SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I7P102

Volume 13 Issue 7, 17-34, July 2026

Original Article

An Attention-Guided Framework for Feature-Level and Decision-Level Fusion in Multimodal Emotion Recognition

Chintan Chatterjee[1], Brijesh Bhatt[2]

[1,2]*Department of Computer Engineering, Dharmsinh Desai University, Gujarat, India.*

[1]*Corresponding Author : chintanchatterjee.ce@ddu.ac.in*

Received: 11 April 2026 Revised: 30 May 2026 Accepted: 17 June 2026 Published: 29 July 2026

Abstract - *The Multimodal Emotion Recognition (MER) is a critical aspect in the development of human-computer interaction since it is a synthesis of non-redundant information based on the use of text, audio, and visual modalities. However, the comparative effectiveness of disparate fusion strategies and mechanisms of attention in MER has not been thoroughly studied on a single experimental paradigm. In line with this, the present study engages in a comparative analysis of four multimodal configurations, namely: early Fusion without attention, early Fusion with attention, late Fusion without attention, and late Fusion complemented by attention. Each of these configurations is evaluated on the Multimodal Emotion Lines Dataset (MELD) using harmonized training protocols to ensure a fair comparison. The methodologies use pre-trained architectures of BERT to support textual representation, Wav2Vec to support acoustic encoding, and TimesFormer to support visual streams to extract features that are modality-specific. These characteristics are then fused through the above fusion tactics. Attention modules that constitute cross attention, hierarchical attention, and self-attention modules are incorporated to enable cross-modal as well as intra-modal features interaction. The results of empirical studies have shown that there are recognizable differences between the models, where attention-enhanced schemes have better measures of performance, especially improved in terms of accuracy and weighted F1-score. However, they still have residual problems, including the imbalance of the classes that influence minority emotion categories. Such findings explain the impact created by specific fusion paradigms and attention structure on the multimodal emotion recognition performance. In turn, this research provides a methodological framework that can be used to develop more effective and understandable MER systems using a systematic and structured comparative framework.*

Keywords - *Attention mechanisms, Early Fusion, Late Fusion, Multimodal Emotion Recognition, Multimodal Fusion.*

1. Introduction

Multimodal Emotion Recognition (MER) addresses this limitation by integrating textual, acoustic, and visual data to construct richer and more accurate representations of emotional states. By leveraging the complementary strengths of these modalities, textual semantic content, audio prosodic cues, and visual facial dynamics, MER systems offer substantially improved emotion detection compared to any single-modality approach. Despite this progress, a critical and underexplored research gap persists: no controlled, unified experimental study has systematically isolated the independent contributions of fusion strategy and attention mechanism to MER classification performance. Prior works have frequently confounded these variables by varying multiple architectural choices simultaneously, making it difficult to determine whether performance gains arise from the fusion paradigm, the attention design, or differences in experimental configuration.

Feature-level Fusion, commonly referred to as early Fusion, combines modality representations at an intermediate stage, allowing the model to jointly learn cross-modal interactions. Decision-level Fusion, or late Fusion, processes each modality independently and integrates results at the final classification stage, preserving modality-specific information at the cost of direct cross-modal interaction. Each paradigm carries distinct trade-offs, yet a rigorous, side-by-side comparison under identical experimental conditions remains absent from the literature.

A further limitation of existing MER research is the insufficient treatment of class imbalance and model interpretability. Naturalistic conversational data is inherently skewed toward neutral emotional expressions, leaving infrequent categories such as fear severely underrepresented, which standard training objectives fail to address adequately. Additionally, deep fusion models frequently operate as black

boxes, offering limited transparency regarding which modalities drive individual predictions, a practical constraint that restricts deployment in high-stakes domains such as clinical affective assessment.

This study directly addresses these gaps through a controlled comparative analysis of four multimodal fusion architectures evaluated on the Multimodal Emotion Lines Dataset (MELD): (1) early Fusion without attention, (2) early Fusion with cross-attention and hierarchical attention, (3) late Fusion without attention, and (4) late Fusion with self-attention refinement. All four architectures employ identical pretrained encoders, BERT for textual representation, Wav2Vec for acoustic encoding, and TimesFormer for visual streams, and are trained under harmonized protocols. Class imbalance is addressed through Focal Loss in early fusion models and label-smoothed cross-entropy in late fusion models. This systematic design enables clear attribution of performance differences to specific architectural choices, providing a transparent and reproducible methodological framework for advancing interpretable and robust MER systems.

The principal novelty of this work lies not in proposing a single new architecture, but in providing the first controlled, four-way comparative study of fusion strategy and attention mechanism on the MELD dataset under harmonized experimental conditions. Unlike prior works that propose a fixed architecture and compare against external baselines, this study holds all variables constant except the design choice under examination, enabling clean and reproducible conclusions about the independent contribution of fusion paradigm and attention mechanism to multimodal emotion recognition performance.

2. Literature Review

2.1. Emotion Recognition in Conversational Systems

The concept of Emotion Recognition in Conversations (ERC) has become an important field of study in affective computing, especially in enhancing human-computer interaction systems. ERC concentrates on the definition of the emotional state elicited in every utterance in a conversation through the discussion of the contextual, linguistic, and behavioral signs. Emotion detection has been shown to help intelligent systems respond more naturally to applications like conversational agents, healthcare monitoring, recommendation systems, and interactive learning environments when done accurately [1]. The initial attempts at recognition of emotions mostly used unimodal data sources, including text, speech, or facial expressions. Text-based systems applied the methods of natural language processing to recognize words carrying sentiment and contextual linguistic patterns, whereas speech-based systems used the acoustic characteristics of the voice, tone, and prosody to determine the emotional state. Systems that were visual-based and concentrated on facial expressions, eye movements, and micro-expressions, which were observed on video data. Though these unimodal methods showed encouraging outcomes in controlled situations, they were not always able to reproduce the emotional expression complexity of real conversation in a natural environment [2]. Human feelings are multimodal and are normally expressed in the form of a mixture of verbal, vocal, and visual messages. Consequently, unimodal methods are incapable of capturing the entire range of emotional cues and, therefore, they cannot be effective in multifaceted conversations. Because of this constraint, researchers have been moving towards the use of multimodal emotion recognition frameworks, which combine information across different modalities to provide stronger and more context-sensitive emotion classification [3]. The Multi-modal Conversation Emotion Recognition (MCER) wants to identify and monitor the speaker's emotional state based on text, speech, and visual information in the conversation scene [4]. Unlike single-utterance emotion classification, MCER must take complicated interactions between turns and speakers into account. MCER is more challenging than single-utterance recognition because of the complex emotional context and speaker dependencies [3]. Recent reviews classify MCER methods according to contextual modelling: context-free methods, which do not consider the past dialogue, sequential context modelling methods, which model temporal chains of utterances; speaker-differentiated methods, which model individual speaker characteristics, and speaker-relationship models, which explicitly model interactions between speakers. These models, based on graphs or attention, have demonstrated better performance than the naive baselines by explicitly learning the structure of the conversation [5].

Despite promising results in controlled environments, unimodal ERC systems consistently underperform in naturalistic conversational settings. Text-based systems often fail to detect emotions conveyed through irony, sarcasm, or emotionally neutral language, where the true affective state is carried by vocal tone or facial expression rather than word choice. Similarly, audio-only systems struggle with speaker variability, background noise, and prosodic ambiguity, while vision-based systems are sensitive to occlusion and lighting conditions. These inherent limitations of unimodal approaches have been documented extensively across multiple benchmark datasets, reinforcing the need for multimodal frameworks that can capture the full spectrum of emotional expression in conversation [2, 6].

Recent work has further refined the taxonomy of ERC models by distinguishing between context-free models, which classify each utterance independently without reference to preceding dialogue turns, and context-aware models, which model temporal dependencies across utterances [3]. Among context-aware approaches, speaker-differentiated methods model individual speaker characteristics separately, while speaker-relationship models

explicitly learn how one participant's emotional state influences another [5, 7]. Graph-based and transformer-based architectures have demonstrated superior performance in this latter category by explicitly encoding dialogue structure [5, 8], though they often sacrifice interpretability for accuracy. This trade-off between performance and transparency motivates the need for more controlled, architecture-comparative studies of the kind presented in this paper.

2.2. Multimodal Emotion Recognition and Feature Fusion

Multimodal Emotion Recognition (MER) has enjoyed great popularity because it can use complementary information provided by other modalities, e.g., text, audio, video, etc. Through integration, MER systems are able to provide more expressive representations based on emotion and better recognition performances than single-mode systems. It has been determined that using multimodal cues allows models to receive the semantic meaning of text as well as the emotional context of non-verbal speech and facial expressions [6]. Nonetheless, heterogeneous modalities introduce several challenges, especially in feature representation and Fusion across modalities. Various modalities generate features of different structures and dimensions, and it is not easy to properly match and integrate them. Examples of traditional fusion strategies are early Fusion and late Fusion; in early Fusion, features are fused at the representation level, whereas in late Fusion, the prediction of each of the modalities is fused at the decision level. Although early Fusion can achieve more interactions between modalities, it can be noisy when the modalities are not well aligned. Late Fusion, on the other hand, retains modality-specific information, but it might not make full use of cross-modal relations [7]. Recent studies have thus been dedicated to the development of advanced fusion architectures that could better capture cross-modal dependencies without destroying the peculiarities of each modality. The purposes of these architectures are to increase feature compatibility, minimize intermodal redundancy, and enhance the robustness of emotion recognition systems during conversations [9].

Beyond early and late Fusion, hybrid fusion strategies have also been explored, wherein features are fused at multiple stages simultaneously. The Tensor Fusion Network models unimodal, bimodal, and trimodal interactions jointly through an outer product-based fusion layer, demonstrating that capturing higher-order cross-modal interactions improves sentiment and emotion classification [10]. Similarly, the Low-rank Multimodal Fusion method, which reduces the computational cost of tensor-based Fusion while maintaining competitive performance, makes hybrid Fusion more practical for deployment [11]. A further challenge in fusion research is the handling of missing or noisy modalities at inference time. In paper [12], demonstrated that models trained on complete trimodal data suffer significant performance degradation when one modality is absent at test time, and a modality-dropout training strategy was proposed to improve robustness. These findings collectively highlight that the choice of fusion stage, fusion mechanism, and robustness strategy each independently affect MER system performance, a distinction that existing literature has not systematically examined within a single controlled study [7-9].

2.3. Deep Learning Approaches for Multimodal Emotion Recognition

Deep Learning architectures have substantially advanced the state of Multimodal Emotion Recognition by enabling models to learn hierarchical, context-sensitive representations from raw or pre-processed inputs. Early deep MER systems relied on Convolutional Neural Networks (CNNs) for spatial feature extraction from visual and audio spectrograms, and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to capture temporal dependencies across utterances [12, 13]. These architectures established the foundation for sequential emotion modelling in conversational data, though they remained limited in their ability to capture long-range dependencies and cross-modal interactions simultaneously [3, 14].

The introduction of transformer-based models has significantly shifted the design space for MER. BERT (Bidirectional Encoder Representations from Transformers), a pretrained language model that produces deep contextual embeddings for textual inputs, has since become the standard encoder for text modality in MER systems [15]. For the acoustic modality, the author of [16] proposed Wav2Vec 2.0, a self-supervised speech representation model that learns directly from raw waveform data, capturing fine-grained paralinguistic features such as pitch, rhythm, and speech rate without manual feature engineering. For visual streams, TimesFormer was introduced as a transformer architecture that models spatiotemporal relationships across video frames through divided space-time attention, enabling robust extraction of facial expression dynamics over time [17]. These three pretrained encoders, BERT, Wav2Vec 2.0, and TimesFormer, form the feature extraction backbone of the proposed framework in this study and represent the current state of the art in modality-specific representation learning.

Building on these pretrained representations, recent MER systems stack learnable fusion layers including attention mechanisms, gating networks, and graph convolutions to encode cross-modal interactions [14, 15]. In paper [13], a comprehensive survey of bimodal speech-text emotion recognition and demonstrated that models combining CNNs, RNNs, and multi-head self-attention consistently outperform pure CNN or RNN baselines, with attention integration yielding the most significant gains. Paper [14] proposed AFR-BERT, which fuses BERT-encoded text with BiLSTM-encoded audio through a cross-

attention fusion network, achieving state-of-the-art performance on standard sentiment benchmarks. These results confirm that the combination of pretrained encoders with learnable attention-based Fusion is among the most effective paradigms for MER, and provide the architectural motivation for the models evaluated in this study.

Recent MER models make a lot of use of deep architectures to learn rich features and interactions. Common types of networks include convolutional neural networks and recurrent neural networks: a survey mentioned in paper [13], bimodal (speech + text) emotion recognition, and highlight models that combine convolutional neural networks, recurrent neural networks, and multi-head self-attention. They propose a multi-learning model based on deep learning, which utilizes the integration of CNNs and RNNs with attention layers; experiments are conducted to demonstrate that the integration of attention is extremely helpful in enhancing performance compared with pure CNN/RNN baselines. In the multimodal setup, AFR-BERT that utilizes BERT to process text and a BiLSTM for audio, then combines them in an attention-based module. This Fusion of cross attention is shown to be significantly better than naive baselines on standard sentiment/emotion benchmarks [14]. In all, deep MER systems benefit from the embeddings of pretrained models (BERT, Wav2Vec, ResNet, etc.) and stack them with learnable fusion layers (attention, gating, or graph layers) to encode complex inter-modal patterns [14, 15].

2.4. Graph-based Learning Approaches for Emotion Recognition in Conversations

Graph-based learning techniques have been proposed as a promising tool for capturing the complicated relationships that exist in conversational data. In conversation, the utterances, speakers, and modalities usually have non-linear interaction that cannot easily be represented in the conventional sequential models. Graph Neural Networks alleviate this weakness by modelling conversation in the form of a graph, whereby nodes may represent utterances or modalities and edges may reflect contextual or speaker relations [5]. Several studies have suggested graph-based architectures of multimodal emotion recognition. Other models like DialogueGCN build graphs that reflect the contextual dependencies and the speaker relations so that the system is able to capture long-distance relationships between utterances [18]. The other methods combine a variety of graphs to be able to represent various modalities and contextual structures to have a model that can make a better understanding of the interactions between textual, acoustic, and visual cues [5]. Despite the benefits they have, graph-based approaches continue to have issues associated with modal heterogeneity and feature alignment. Simple graph structures in most of the instances do not represent the relationship between modalities in their entirety, resulting in suboptimal Fusion of multimodal features. In turn, the recent studies have investigated heterogeneous graph networks, which are more effective in the reflection of the intra-modal and inter-modal relationships [18].

In conversational ERC, a node may represent every utterance, and temporal or speaker relations are represented as an edge. An example is a graph of the conversation as the DialogueGCN constructs the nodes, which are utterances, and the edges, which are self- and inter-speaker dependencies. This GCN-based model overcomes the context-propagation problems of RNNs by explicitly transmitting information between speaker nodes and achieves substantial improvements on recurrent baselines [9]. Other more recent papers also utilize graph convolutions to connect semantically similar or temporally similar utterances. Naturally, these graph-based models can deal with long-range dependencies and conversation context: they learn to spread affective information across speakers, the effect of one participant's emotion on another. Generally, graph learning is an extension of MER, which includes the dialogue structure and interaction between speakers in the learning of features [19].

More recent work has explored heterogeneous graph networks that simultaneously model multiple types of relationships, such as speaker-speaker, utterance-utterance, and modality-modality edges within a single unified graph structure. HGLER [1], a hierarchical heterogeneous graph network that constructs multi-relational graphs over conversational utterances, demonstrated consistent improvements over homogeneous graph baselines on the MELD and IEMOCAP benchmarks. further demonstrated that incorporating modality-aware and identity-aware edges into graph structures significantly improves cross-speaker emotion classification in multi-party conversations [9]. While these heterogeneous graph approaches offer richer relational modelling, they introduce substantial computational complexity and reduced transparency regarding which relational structure contributes most to classification performance [18]. This interpretability limitation, combined with the high computational overhead of graph-based architectures, positions simpler attention-based fusion frameworks such as those examined in this study as a more practical and transparent alternative for MER system design.

2.5. Cross-Modal Attention and Gated Fusion Mechanisms

Attention mechanisms have been a popular method of multimodal feature fusion enhancement. These processes permit models to dynamically place weight on the various modalities or features, enabling the system to highlight the most important emotional cues when making predictions. The cross-modal attention procedures allow the modalities to interact with each other, i.e., the features of a certain modality can be used to direct the learning of the representation in another modality [6, 14]. The recent research suggested the use of cross-modal gated attention to improve the multimodal

feature representation. These techniques bring in gating networks, which control the contribution of each modality in the fusion process. The model is able to use the complementary features across modalities and reduce redundant or irrelevant information by dynamically modifying the weight of modalities. The method can enhance the accuracy of emotion recognition as well as the stability of the model in dialogue settings [18]. Also, contrastive learning methods have been integrated into multimodal emotion recognition systems to improve the cross-modal feature alignment. In contrastive learning, the model is made to reduce the space between correlated multimodal representations and maximize the distance between irrelevant samples. The approach enhances the discriminative learning of emotional characteristics of the model using heterogeneous multimodal inputs [18].

Hierarchical attention mechanisms represent a structured extension of standard attention, wherein multiple levels of attention are applied sequentially to progressively refine the multimodal representation. At the lower level, intra-modal attention captures dependencies within each modality independently, for example, identifying which words in a sentence or which frames in a video are most affectively salient. At the higher level, inter-modal attention integrates these refined unimodal representations, allowing the model to weight the relative contribution of each modality dynamically [20]. Hierarchical cross-attention applied to dense graph Convolutional Networks yields stronger multimodal emotion recognition performance than flat attention, particularly by enabling the model to capture fine-grained intra-modal patterns before cross-modal integration [18]. This hierarchical design principle directly motivates the cross-attention and hierarchical attention modules employed in the early fusion architecture evaluated in this study.

Self-attention, applied within a single modality rather than across modalities, has also been shown to be a powerful tool for refining modality-specific representations before Fusion. The theoretical foundation of self-attention in the original Transformer architecture demonstrates that attending to all positions within a sequence simultaneously enables richer contextual representations than recurrent models [21]. In the context of MER, self-attention enables the model to identify the most emotionally informative segments within a single modality stream, such as the most expressive utterance in a text sequence or the most dynamic facial frames in a video, before the modalities are combined [13]. This principle underlies the self-attention refinement step incorporated in the late Fusion with the attention architecture proposed in this study.

2.6. Large Language Models for Multimodal Emotion Recognition

The emergence of Large Language Models (LLMs) in the recent past has had a profound impact on the creation of multimodal emotion recognition systems, especially in conversational settings. Traditional multimodal models can use hand-crafted fusion mechanisms and modality-specific feature extractors, which can be restrictive in modelling the complex contextual relationship between dialogues. In order to overcome this weakness, [4] introduced DialogueMLLM, a Multimodal Large Language Model with instruction tuning aimed at increasing emotion recognition in a conversational context. The suggested model combines textual, visual, and acoustic modalities into one instruction-following framework, allowing the model to utilize the contextual reasoning of large language models. DialogueMLLM proves to have better contextual and emotive reasoning when compared to dialogue-based datasets, by matching multimodal representations to instruction-based learning goals. Experimental analyses have shown that the instruction-tuned multimodal architecture relative to the traditional multimodal fusion systems results in much stronger emotion classification, which points to the promise of utilizing the LLM-based frameworks to develop emotion recognition in conversations [4].

Despite the impressive capabilities of LLM-based MER systems, several practical limitations constrain their applicability in real-world deployment. The computational cost of fine-tuning or inference with large-scale multimodal LLMs is substantially higher than that of smaller task-specific architectures, making them less suitable for resource-constrained environments [19, 20]. Furthermore, LLMs operate as largely opaque systems, offering limited interpretability regarding which modality or contextual feature drives a specific emotion prediction [21]. In clinical affective computing, educational technology, and human-robot interaction domains, where understanding the basis of an emotion classification is as important as the classification itself, this opacity represents a significant limitation [16, 22]. These considerations motivate the use of lighter, controlled fusion architectures in this study, which prioritize interpretability, reproducibility, and systematic comparison over raw performance maximization.

2.7. Multimodal Emotion Recognition using the MELD Dataset

Several papers have investigated emotion recognition in multimodal mode with benchmark conversational data to measure the performance of the models in a natural conversation. In paper [23], the author introduced a multiprojective sentiment acknowledgement framework that is uniquely created to handle conversational information based on the REmote COLlaborative and Affective (RECOLA) database. The analysis combines the text, audio, and visual representations in order to portray the supplementary emotional cues that are witnessed in the conversational interactions. The framework uses deep learning approaches to extract the semantic meaning of texts, acoustic patterns of speech signals, and visual facial expressions that can allow conducting a rich analysis of

emotional states within dialogical contexts [23]. The experimental findings on the RECOLA dataset indicate that multimodal Fusion can attain a high recognition accuracy on emotions in comparison to unimodal methods. The paper also emphasizes the significance of unifying the heterogeneous sources of data to acquire both verbal and non-verbal expressions of emotion, as well as the capabilities of multimodal data sets in future research of emotion recognition in conversational AI systems [24].

Several additional studies have reported results on the MELD dataset using varying multimodal fusion strategies, providing a useful landscape for contextualizing the results of this study. Proposed multimodal sentiment analysis framework that combined textual, acoustic, and visual features through a memory-based fusion network and reported competitive weighted F1-scores on MELD, highlighting the effectiveness of memory-augmented architectures for conversational emotion modelling [25]. ICON, a conversational emotion recognition model that employs a global-local memory mechanism to capture both individual speaker states and global inter-speaker context, demonstrating improved performance on MELD relative to simpler sequential models [26]. More recently, the generalisation and confidence calibration of multimodal conversational ERC models across multiple datasets, including MELD, identified that most models exhibit significant overconfidence on majority emotion classes while underperforming on minority categories, a finding that directly aligns with the class imbalance observations reported in this study [27].

The class imbalance inherent in the MELD dataset poses a well-documented challenge for MER systems. The neutral class accounts for approximately 47% of all utterances, while low-frequency categories such as disgust and fear represent less than 5% of the data combined [28]. Standard cross-entropy loss optimization under such conditions leads models to disproportionately favour majority classes, resulting in inflated overall accuracy that masks poor performance on minority emotion categories. Prior work has addressed this through class-weighted loss functions, oversampling strategies, and data augmentation techniques [11]. The choice of loss function is therefore not merely an implementation detail but a fundamental design decision with direct consequences for model fairness and applicability. This study explicitly addresses this challenge by employing Focal Loss in early fusion models and label-smoothed cross-entropy in late fusion models, both of which have been shown to improve minority class recall in imbalanced classification settings.

Other commonly used MER benchmarks include IEMOCAP (Interactive Emotional Dyadic Motion Capture), which contains dyadic acted conversations across ten emotion categories [29]; CMU-MOSI and CMU-MOSEI, which focus on sentiment intensity in monologue video recordings [29]; and the SEMAINE corpus, which captures naturalistic human-agent interactions [29]. While each of these datasets offers distinct advantages, MELD is uniquely suited to the goals of this study because it contains multi-party conversational data drawn from a naturalistic television source, includes all three modalities in an aligned format, and presents the class imbalance challenge that is representative of real-world emotion distributions in conversation [29].

The MELD dataset (Multimodal EmotionLines) is a widely-used benchmark for conversational MER. It contains 13,708 utterances from *Friends* TV show scenes, each labelled with one of seven emotions (anger, disgust, fear, joy, neutral, sadness, surprise) [25]. The data include aligned text (subtitles), audio, and video for each utterance, enabling researchers to develop and test multimodal models. Many studies report results on MELD; for instance, a multi-level fusion model significantly improves accuracy on MELD's emotion recognition task [25]. The dataset's conversational context (dialogue threads) also encourages the use of context-aware models (like DialogueGCN or transformers with context). In summary, MELD has become a standard testbed for multimodal emotion systems in dialogue, facilitating direct comparison of architectures (fusion strategies, attention, context modelling) under consistent conditions.

2.8. Identity and Modality-Aware Fusion Networks for Emotion Recognition

The other research direction of multimodal emotion recognition is the design of fusion architecture, which is capable of modelling interactions across modalities and speaker-specific detail in conversations [3]. An Identity and Modality Attributes Driven Multimodal Fusion Network, which integrates modality-specific attributes and ego identity data to learn how to better recognize emotions in conversational data. The suggested model builds on identity-sensitive representations to promote speaker-specific emotional patterns, whereas modality-sensitive attention mechanisms allow effective textual, acoustic, and visual information integration. The framework contributes to the improved modelling of contextual emotional dynamics in conversations by modelling the relationships between speaker identity and multimodal emotional cues. The experimental findings indicate that adding identity attributes to multimodal Fusion has a significant positive effect on emotion classification performance, thereby indicating that the context-related information of the speaker is very important in emotion recognition tasks during conversations [30].

Measuring speaker identity and biases of modality can also enhance MER. Cross-subject generalization is addressed in one of the lines of work. The model reduces the problem of subject discrepancy through adversarial training: a domain discriminator is trained to be incapable of distinguishing

between different speakers, and therefore, the shared feature space must be speaker-invariant. This adaptation significantly enhances cross-subject performance [31]. In the same manner, the fusion system that directly considers the variations of identities: the authors use a combination of facial features with skin conductance (GSR) and EEG in a single model and test one-subject-out performance. Their hybrid fusion approach is highly accurate with a maximum accuracy of 81.2% in a leave-one-subject-out test, which is much higher than traditional baselines [22]. These identity-conscious models demonstrate that it is possible to decrease overfitting to specific speakers by including the subject labels or modality-sensitive sub-networks. The other option is modality-aware Fusion: e.g., gated networks explicitly use information about which modality is most reliable at a specific time. Overall, the most recent fusion networks typically have elements to normalize or gate by identity and modality, such that the model learns emotion signals that are speaker-agnostic and that are capable of dealing with missing or noisy modalities [32, 33].

Collectively, the body of literature reviewed in Sections 2.1 through 2.8 reveals several persistent gaps that the present study is designed to address. First, while a wide range of fusion strategies, early, late, hybrid, graph-based, and LLM-based have been individually proposed and evaluated, no study has systematically compared early and late Fusion with and without attention mechanisms under identical experimental conditions on the same dataset [7-9, 31-33]. Second, the independent contribution of attention mechanisms to MER performance, separate from the choice of fusion stage, has not been cleanly isolated in prior work, making it difficult to draw principled architectural conclusions [14, 15]. Third, the challenge of class imbalance in MELD has been acknowledged in several studies but is rarely addressed through a deliberate, fusion-strategy-specific loss function design [24]. Fourth, model interpretability understanding which modality and which attention pattern drives a given prediction remains underexplored relative to the emphasis placed on raw performance metrics [19, 21, 25]. This study addresses all four of these gaps through a controlled, four-architecture comparative framework using state-of-the-art pretrained encoders, tailored loss functions, and a consistent evaluation protocol.

3. Research Gap

Substantial progress has been made in Multimodal Emotion Recognition (MER) over the past decade, yet several critical gaps remain unaddressed in the existing literature. Existing MER studies predominantly propose a single architectural design and evaluate it against prior baselines without systematically isolating the contribution of individual design choices, such as the fusion stage or attention mechanism. As a result, it remains unclear whether performance improvements in state-of-the-art systems arise from the fusion strategy, the attention design, the choice of pretrained encoder, or simply from differences in experimental configuration. This confounding of variables constitutes the primary methodological gap that motivates the present study.

Specifically, four gaps are identified from the literature: First, no prior study has conducted a controlled comparison of early fusion and late fusion strategies with and without attention mechanisms under completely identical experimental conditions on the MELD dataset. Prior works such as DialogueGCN [31], MMGCN [33], and DialogueCRN [32] each adopt a fixed fusion paradigm without comparing alternative strategies, making it impossible to attribute their performance to fusion design specifically.

Second, the individual contribution of attention mechanisms to MER performance independent of the fusion stage has not been cleanly isolated. Studies such as AFR-BERT [14] and cross-modal gated attention networks [10] incorporate attention as part of a broader architectural change, preventing clean attribution of performance gains to the attention component alone.

Third, while class imbalance in MELD is widely acknowledged, prior work rarely addresses it through fusion-strategy-specific loss function design. Most existing models apply a single loss function regardless of fusion strategy, overlooking the fact that feature-level and decision-level Fusion create structurally different optimization landscapes that may require different training objectives.

Fourth, the interpretability of multimodal fusion models, understanding which modality and which attention interaction drives a specific prediction, remains largely unaddressed relative to the emphasis placed on benchmark accuracy scores [17, 19, 21].

The present study addresses all four of these gaps and makes the following specific novel contributions, which distinguish it from all prior work on MER: A unified four-architecture comparative framework is proposed, comprising early Fusion without attention, early Fusion with attention, late Fusion without attention, and late Fusion with attention, all evaluated under identical training protocols, encoders, and datasets. To the best of the authors' knowledge, no prior study has performed this specific four-way controlled comparison on MELD. The independent effect of attention mechanisms on MER performance is isolated by holding all other variables constant across paired architectures, early Fusion with vs. without attention, and late Fusion with vs. without attention, enabling clean, attributable conclusions about the value of attention at each fusion stage. A fusion-strategy-specific loss function design is introduced: Focal Loss with class-specific weighting is applied in early fusion models to

address the feature-level class imbalance, while label-smoothed cross-entropy is applied in late fusion models to reduce overconfidence in decision-level integration. This tailored approach to loss function design based on the fusion paradigm is a novel methodological contribution not present in prior MER literature. State-of-the-art pretrained encoders BERT [15] for text, Wav2Vec 2.0 [16] for audio, and TimesFormer [17] for video are employed as a shared, consistent feature extraction backbone across all four architectures, ensuring that performance differences between architectures reflect fusion and attention design rather than differences in representation quality.

4. Experimental Setup

This study experiment was structured in such a way that it had a fair and consistent evaluation of the proposed multimodal emotion recognition frameworks. The experiments were carried out with the help of the Multimodal Emotion Lines Dataset (MELD), which is commonly utilized in research on conversational emotion recognition. The data will include around 1,400 conversations and more than 13,000 utterances, and they are labelled with one of seven emotions: angry, disgusted, sad, joyous, neutral, surprised, and fearful. MELD offers multimodal data such as textual transcripts, audio files, and visual information that is gleaned from the Friends television series. Such a multimodal structure allows for the analysis of emotional information through linguistic content, vocal features, and facial expressions in conversations.

To conduct the experiments, the dataset was split into a training and testing set (70/30). It is important to note that the models were trained on a large portion of the data set, and the independent test set was used to evaluate the models. Among the main issues of the MELD dataset is the imbalance in classes, in which some classes of emotions, like joy and neutral, are much more prevalent than others, like fear or disgust. To overcome this problem, dataset balancing policies were included in the training process to eliminate the bias in most classes and enhance the learning of minor classes of emotions. These measures contribute to the need to make sure the evaluation measures are based on the capacity of the model to identify a wide range of emotional states and not focus on prevailing classes.

Preprocessing was done before experimentation to make the multimodal inputs trainable. Embeddings of each modality were extracted and transformed into a normalized NumPy array, and unwanted or erroneous values were verified. A Standard Scaler was used to normalize the features to stabilize the training process since it makes sure that the distributions of features are similar across the modalities. Also, feature dimensions of various modalities were matched by projection operations in order to be merged in the fusion process.

5. Methodology

The paper explores multimodal emotion recognition through a systematic comparison of four fusion systems, namely Early Fusion without Attention, Early Fusion with Attention, Late Fusion without Attention, and Late Fusion with Attention. The purpose is to analyse the impact of various fusion strategies on the process of multimodal information integration to classify emotions in conversational data. The three modalities of video, text, and audio are then converted into representations of embeddings in pretrained models, in all architectures. The TimesFormer is used to generate visual spatiotemporal embeddings of video frames, BERT is used to generate textual contextual embeddings, and Wav2Vec is used to produce acoustic embeddings that reflect speech features. These embeddings give feature representations in the form of standard, and they are the inputs to the proposed frameworks of Fusion.

In the initial fusion methods, each of the modalities is embedded together and at the feature level to create a single multimodal representation before classification. The early Fusion with the attention model does the same thing, except it fuses the features directly and then adds fully connected layers to make the prediction. Conversely, the early Fusion with the attention model proposes modules that are based on attention to facilitate the model to learn cross-modal relationships before the period of classification. In the late fusion strategies, the different modalities are processed separately to acquire modality-specific representations and then integrated. The late Fusion without attention architecture merges the modality-specific outputs later on with feature aggregation, and the late Fusion with attention architecture that integrates an attention mechanism during the extraction of modality-specific features and improves the representation of cross-modal dependencies before the last classification layer. Each of the four architectures applies the identical multimodal embeddings and is tested in the same experimental conditions, which allows achieving a clear comparison between feature-level and decision-level multimodal emotion recognition fusion strategies.

5.1. Early Fusion Approach

Early fusion techniques combine text, audio, and video embeddings at the initial stage of the model to form a unified multimodal representation for emotion classification. By integrating modalities early in the processing pipeline, this approach enables the model to capture cross-modal interactions and temporal relationships important for emotion recognition in conversational data. In this study, two early fusion variants are examined: early Fusion without attention and early Fusion with attention mechanisms, evaluated using the Multimodal Emotion Lines Dataset (MELD).

Video, text, and audio, three modalities, are then transformed into embeddings using pretrained models to create feature representations. TimesFormer is able to extract

spatiotemporal visual embeddings in video, BERT can extract contextual embeddings in text transcripts, and Wav2Vec can generate acoustic embeddings that reflect paralinguistic information (tone, pitch, speech patterns, etc.). Such embeddings have uniform multimodal representations that are employed in all fusion architectures.

By contrast, late fusion methods work with each modality separately and then use their high-level characteristics combined to produce the final emotion. This enables preservation of modality-specific characteristics before integration. Just like in early Fusion, two variants are tested, namely, late fusion (no attention) and late fusion (attention mechanisms), with the same embedding representation between BERT, Wav2Vec, and TimesFormer, to have consistencies in experimental assessment on the MELD dataset.

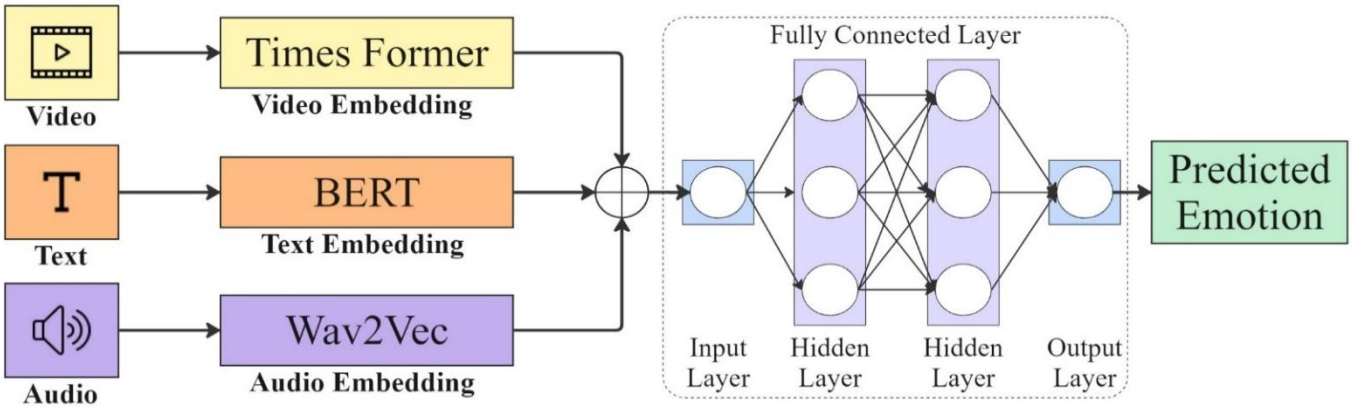

Fig. 1 Early Fusion without attention architecture

5.1.1. Early Fusion without Attention

Figure 1 shows the design of the early fusion method that does not include attention mechanisms. Following the extraction of features, module-specific projection layers are used to match the dimensions of the embeddings. The audio and text embeddings, which were initially 768-dimensional and the 400-dimensional video embeddings, are projected to a shared dimensional space so that they can be compatible during multimodal integration. Embeddings on these aligned points are then concatenated to create a single multimodal feature show. The representation is a combination of all modalities, which allows the model to process all modalities simultaneously and then classify.

Following this, the combined feature vector is fed into a sequence of fully connected layers. These layers are trained to develop more sophisticated representations for emotion classification. The neural network's design includes an input layer, multiple hidden layers, and an output layer. This output layer generates raw scores (logits) corresponding to seven distinct emotion categories: anger, disgust, sadness, joy, neutral, surprise, and fear. RMSNorm is also employed. to provide training stability and generalization. Non-linearity is introduced with the help of ReLU activation, and the dropout rate is set to 0.4 to minimize overfitting in training.

To further manage the issue of class imbalance in the MELD dataset, the model applies Focal Loss, which pays more attention to the hard-to-classify samples and decreases the influence of the majority classes in the optimization process. The Adam optimizer is used for the optimization of model parameters. It offers efficient gradient-based learning as well as stable convergence in the training process.

In general, this architecture builds on the idea of early-stage multimodal integration by concatenating features, allowing the model to be trained on the combination of textual, audio, and video representations, and then performing the classification step.

5.1.2. Early Fusion with Attention

Figure 2 shows the structure of the early fusion architecture that integrates attention mechanisms, as an extension of the fundamental early fusion architecture to include attention-based modules that help the architecture to capture more relationships between modalities. Just like in the case of the previous architecture, the three modalities: video, text, and audio, are initially processed individually through pretrained feature extractor models.

The TimesFormer model produces video embeddings, which are spatiotemporal visual representations including facial expressions and gestures. BERT model generates contextual embeddings that extract the semantic content of a textual conversation. Besides this, Wav2Vec extracts audio embeddings that encode such acoustic and paralinguistic features as pitch, tone, and speech intensity.

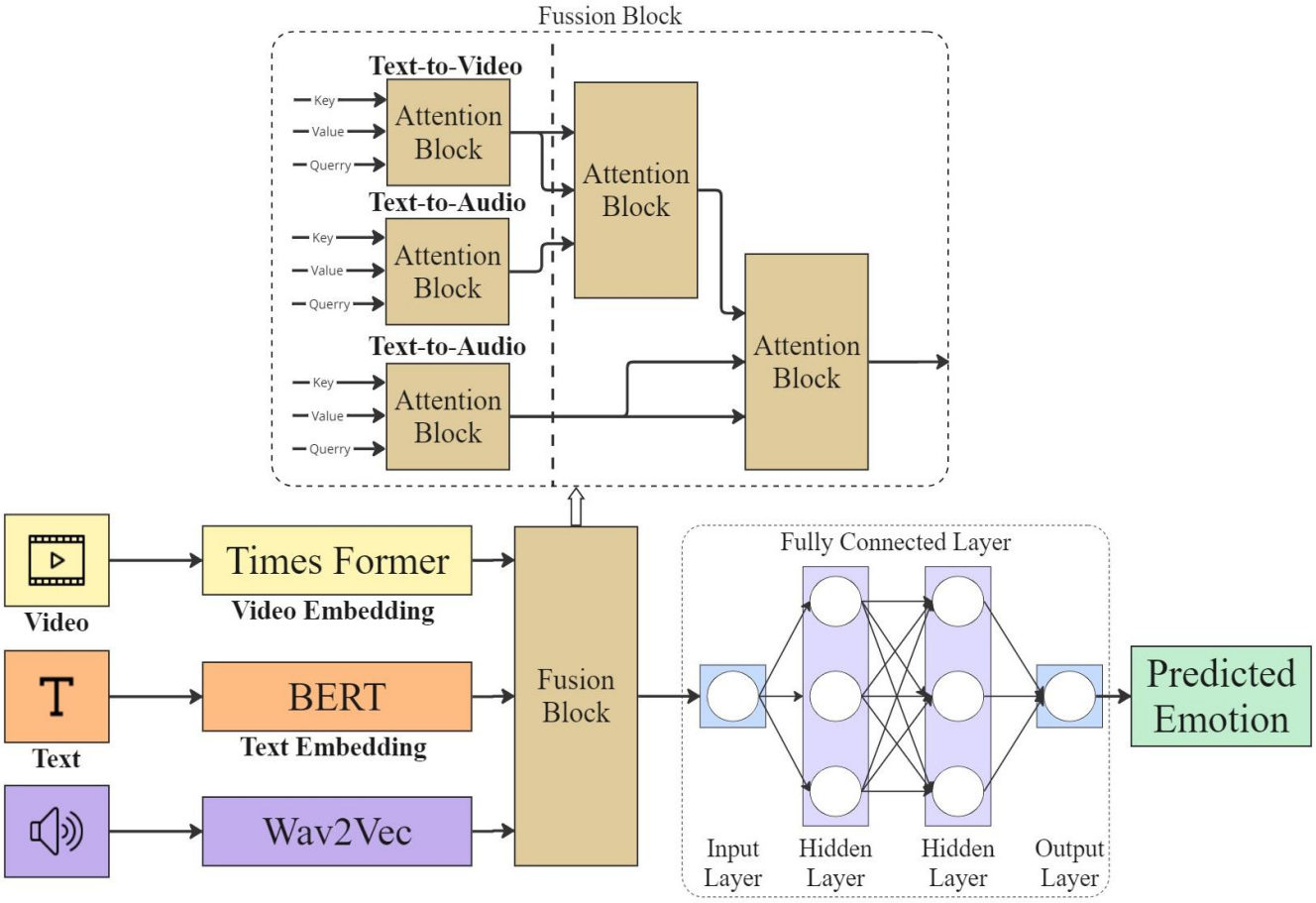

Fig. 2 Early Fusion with attention architecture

The dimensionality between the modality embeddings is brought to match after feature extraction with the help of projection layers. The 768-dimensional text and audio embeddings, as well as the 400-dimensional video embeddings, are mapped into a common representation space so that the two can be compatible with each other in the process of multimodal integration. These aligned embeddings, as opposed to the simplistic early fusion strategy, are initially trained with a fusion block of attention mechanisms that can cause the model to acquire more meaningful cross-modal relationships to be used in the classification.

The architecture has several cross-attention blocks in the fusion block to learn the interaction between the various modalities. More precisely, it is done in three settings, namely text-to-video, text-to-audio, and video-to-audio attention. During cross-attention blocks, one modality is used as a query and another modality is used to provide key and value representation. This design enables the model to selectively pay attention to the information of complementary modalities as it pertains to the learning of emotional cues. The model has the ability to weight features across modalities dynamically to encourage fine dependencies that arise in the expression of emotion in conversations.

After the cross-attention step, the intermediary representations are further represented with hierarchical attention blocks. These blocks gradually increase the multimodal representation through the application of measurement weights to the various modalities and contextual features. This hierarchical model assists in incorporating the multimodal signals in a more effective way and retains the time and contextual ties among features.

Residual connections are also added after attention blocks to enhance the stability of training and efficient gradient flow. The resulting merged representation undergoes a series of fully connected layers, which comprises an input layer, hidden layers, and an output layer that classifies the emotions. Just like the above architecture, RMSNorm is utilized to normalize, ReLU activation is utilized to add non-linearity, and a dropout rate of 0.4 is utilized to curb overfitting.

The last layer generates logits that reflect the seven emotion categories, which include anger, disgust, sadness, joy, neutral, surprise, and fear. Focal Loss is applied in training to correct the problem of the class imbalance in the MELD dataset, so that more emphasis is given to challenging or poorly represented samples. Parameters of the model are optimized with the Adam optimizer, which not only guarantees effective gradient correction and consistent convergence during training. In general, this attention-enhanced early fusion architecture allows modelling cross-modal interactions and temporal dependencies more effectively and allows the system to provide more detailed emotional signals to text, audio, and video modalities.

5.2. Late Fusion Approach

Late fusion methods treat the text, audio, and video modalities independently in the early phases of learning features and then combine the high-level features of the modalities to give the final classification of emotion. According to this strategy, each modality is able to maintain and develop its modality-specific features, so that valuable semantic, acoustic, and visual information is not lost to other modalities at an early stage of processing. Late Fusion allows the model to acquire specialized representations of each modality, and still to receive a combined decision-making process that it can use to recognize emotion.

This paper will examine two types of late Fusion, namely the late Fusion without attention and the late Fusion with attention mechanisms. In the late Fusion without attention, each of the modalities is first processed, and features are extracted, then aggregated together, and finally processed by the last few layers of classification.

Conversely, the late Fusion with the attention model includes attention processes to narrow down the modality-specific representations so that the model can bestow various degrees of significance to each modality at the integration phase.

This cross-modal relationship refinement based on attention assists in capturing additional informative cross-modal relationships and the independence of individual modalities in feature extraction.

Both late fusion architectures are tested on the Multimodal Emotion Lines Dataset (MELD) and assessed in terms of their ability to solve the major problems of Multimodal Emotion Recognition (MER), namely, the ability to integrate features effectively, aligning with the modalities, and reflecting the indicators of emotions in a heterogeneous data stream.

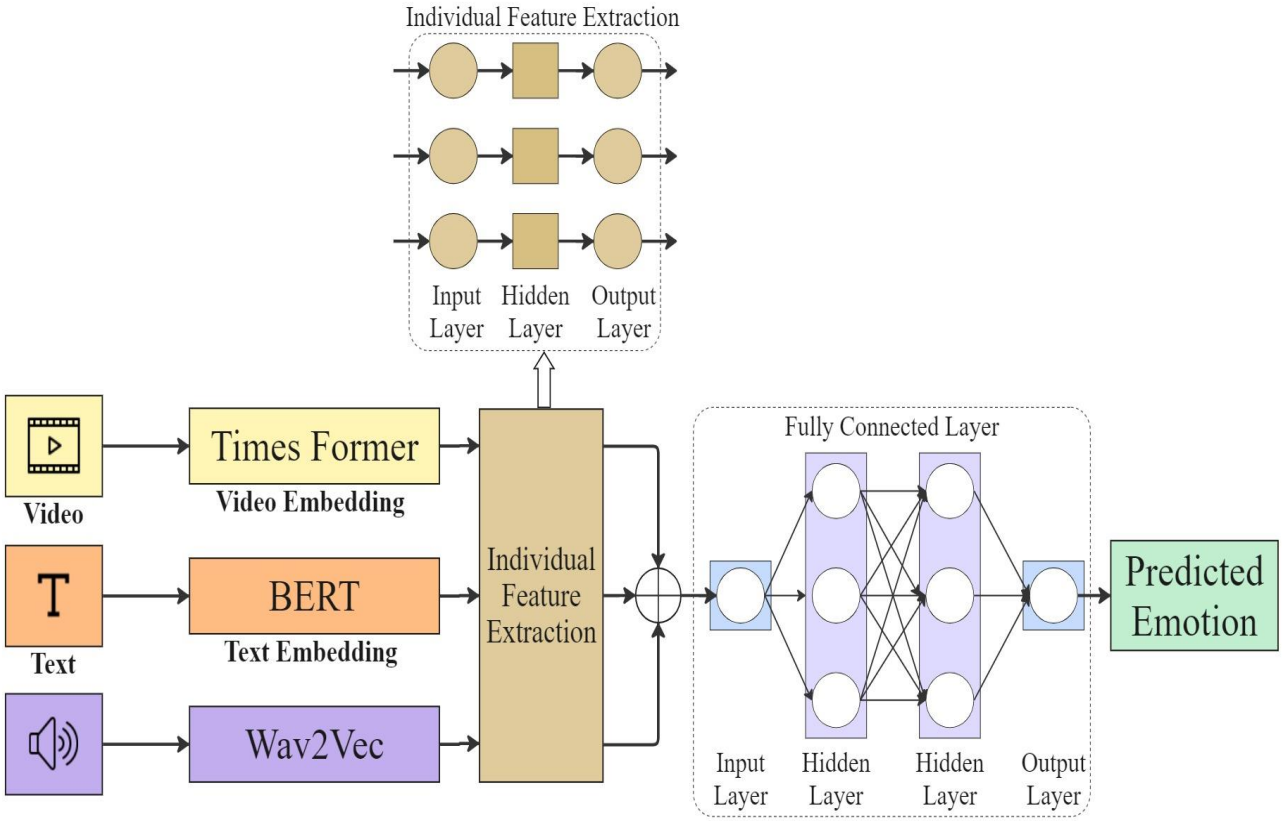

Fig. 3 Late Fusion without attention architecture

5.2.1. Late Fusion without Attention

Figure 3 portrays the architecture of the approach of late Fusion without attention mechanisms, as it involves learning the specific representations of the modality and then integrating with this information in a multimodal manner. Here, embeddings obtained on each modality text (768 dimensions), audio (768 dimensions), and video (400 dimensions) are then processed separately to avoid the blurring of the individual data source features. By lowering the dimensional complexity of high-level representations without compromising important emotional features of each modality, these modality-specific layers learn high-level representations. As an example, semantic and contextual data of dialogue scripts are stored in textual features, paralinguistic data of tone and speech dynamics are stored in audio features, and facial expressions and motion patterns of video frames are encoded in visual features. This processing step, which is independent, enables each modality to perfect its features without other modalities intruding on them. The process continues with intermodal representation that occurs after the individual feature extraction phase, whereby the outputs of that step undergo concatenation to create a composite multimodal feature. This combined representation is subsequently fed through another fully connected classification layer, which learns correlations among the modalities and makes the prediction of emotions. The SoftMax activation is used on the output layer to produce probability distributions of the seven emotion categories: anger, disgust, sadness, joy, neutral, surprise, and fear.

In order to optimize the model, the Cross-Entropy Loss is used as the training goal, which was appropriate in the case of multi-class classification. Adam optimizer is used to update the model parameters, which offer good convergence in the training process. In general, in this late fusion architecture, the focus is placed on saving modality-specific information before the Fusion, and each of the modalities has to make its unique contribution to the final classification. Nevertheless, the lack of attention mechanisms in the model means that it might be weak in the ability to capture complex cross-modal dependencies that occur due to the interaction between textual, acoustic, and visual emotional stimuli.

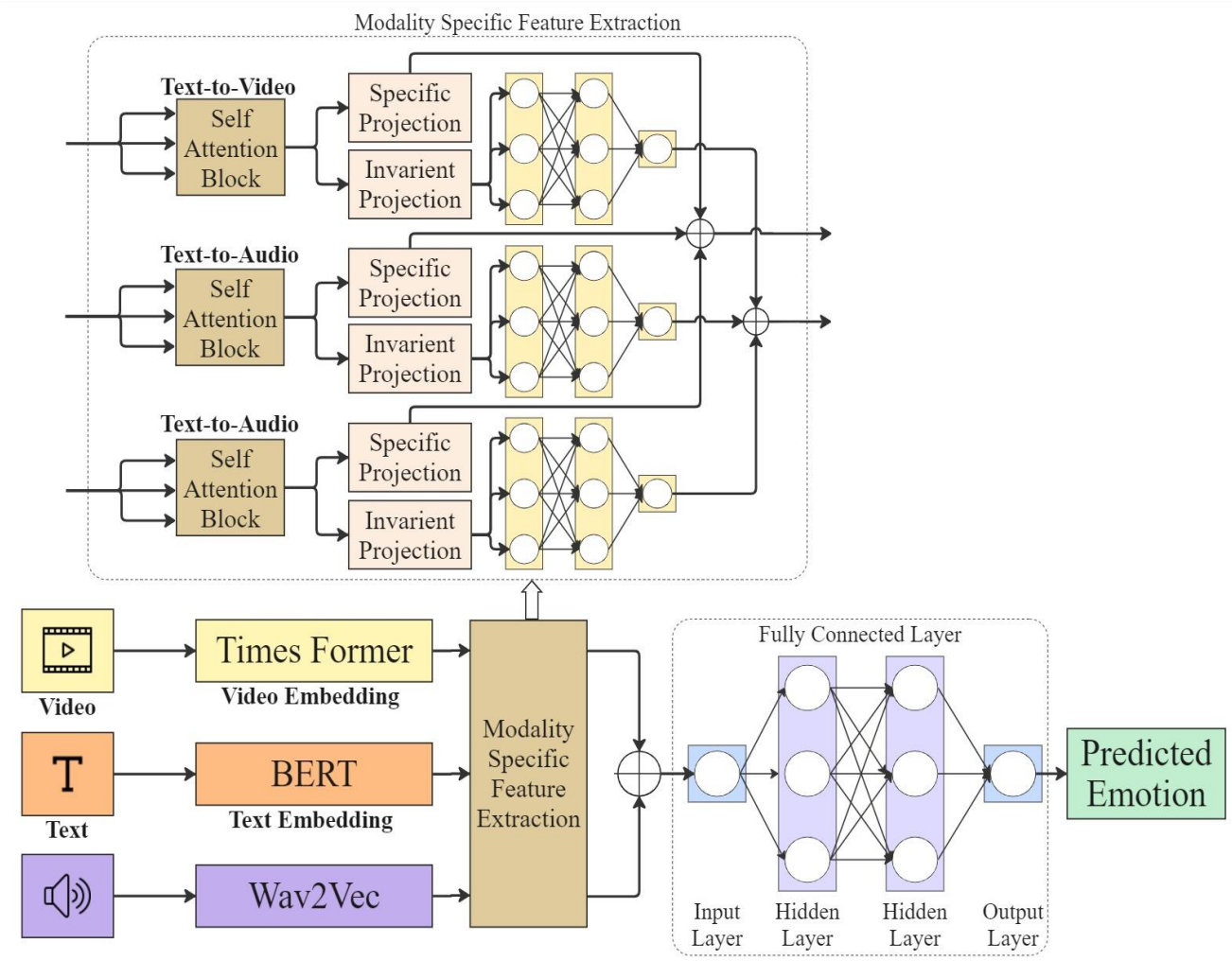

Fig. 4 Late Fusion with attention architecture

5.2.2. Late Fusion with Attention

Figure 4 represents the architecture of late Fusion with attention mechanisms, which adds to the basic late fusion architecture the attention-based modules to optimize the modality-specific feature learning preceding multimodal combination. Just like the former method, three-modality (text, 768 dimensions, audio, 768 dimensions, video, 400 dimensions) embedding representations are initially produced by the pretrained models. Each modality in this architecture goes through self-attention blocks, which fine-tune the generated embeddings by learning contextual and temporal interactions within the modality itself. As an example, self-attention allows the model to define the connection between various segments of a text, speech pattern variations in audio, or facial expression change order in video. The mechanism makes the model place more informative intra-modal features and then multimodal Fusion.

After the self-attention step, the refined features are input into the two projection pathways: a modality-specific projection and a modality-invariant projection. The modality-specific path works on maintaining unique features of each modality, whereas the modality-invariant path obtains features that are common to all modalities and allows heterogeneous data sources to be more aligned. The two predictions are further enhanced using fully connected layers, providing a greater amount of representation learning in each of the modalities, after which, feature representations are then combined to form an overall multimodal representation. This combined feature representation is then run through another fully connected classification layer that learns intermodal relationships and generates the final prediction of emotion. The output is a probability distribution of the seven emotion classes, anger, disgust, sadness, joy, neutral, surprise, and fear, by using a SoftMax activation function.

To maximize performance, the model employs Cross-Entropy Loss, which is adequate in the context of multi-class classification of emotions. The optimization process is done with the Adam optimizer and allows efficient parameter updating and a steady convergence. This architecture, by including attention mechanisms in the extraction of the multimodal specific feature, enhances the correspondence between multimodal representations and the capability of the model to generate salient emotional patterns within and across modalities, facilitating the existence of more robust emotion recognition in conversation.

5.3. Training and Evaluation

The training was done in such a way that it minimized the optimization and fair comparison of all four proposed fusion architectures. The encoders of text (BERT), audio (Wav2Vec), and video (TimesFormer) were pretrained and end-to-end fine-tuned in the process of training. The Adam optimizer was used with a learning rate of 2 x 10-5 and 32 batch size to optimize it. To stabilize the learning process, a linear warm-up schedule was used in the first 1, 000 training steps. All the models were trained up to 50 epochs, and early stopping (patience = 5) was used to prevent overfitting by watching the validation loss. Also, a gradient clipping value of 1.0 was used to ensure the stability of numbers in backpropagation. The fusion strategy relied on different loss functions. In the case of the early fusion architectures, Focal Loss was used to tackle the issue of the imbalance in classes in the MELD dataset. Class weighting a = [0.2, 0.35, 0.1, 0.4, 0.3, 0.15, 0.05] was set to the seven emotion classes of anger to fear, and the focal loss parameters were set to g = 2.0. Conversely, the late fusion systems employed Cross-Entropy Loss using label smoothing (e = 0.1) to give superior generalization and decrease overconfidence throughout classification.

The pipeline training was introduced as a Trainer module, which dealt with the training and validation process. The steps to be made in each batch are gradient initialisation, forward propagation, loss computation, backpropagation, and updating of parameters. The model with the lowest validation loss is automatically stored during training, such that the best model parameters are stored. All the experiments were carried out on a machine that had 2 NVIDIA A100 GPUs, where the multimodal architectures could be trained efficiently. The main methodological elements employed in each architecture, along with the type of Fusion, attention mechanism, loss function, and optimizer, are summarized in Table 1, and they offer a brief overview of the configuration employed in each of the four experimental methods.

Table 1. Summary of key components and configurations for each approach

Fusion	Approach	Attention Mechanism	Loss Function	Optimizer
Early	Without Attention	None	Cross-Entropy Loss	Adam
Early	With Attention	Cross Attention, Hierarchical Attention	Focal Loss	Adam
Late	Without Attention	None	Cross-Entropy Loss	Adam
Late	With Attention	Self-Attention	Focal Loss	Adam

The proposed multimodal emotion recognition frameworks were tested on the basis of the Multimodal Emotion Lines Dataset (MELD), and the evaluation metric was the accuracy. The models were developed to categorize seven emotion types, which are anger, disgust, sadness, joy, neutral, surprise, and fear, through the combination of text, audio, and video modalities. This task was carried out using four methodological settings: early Fusion, no attention, late

Fusion, no attention, early Fusion, no attention, and late Fusion, no attention. The methods allow a comparative analysis of various fusing plans and attention systems of multimodal emotion classification.

The attention-based architectures have their theoretical basis in the cross-attention mechanism that makes the model dynamically learn relationships between modalities. The cross-attention operation can be defined as:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Q is the query embedding derived using the primary modality, and K and V are the key-value pairs derived using complementary modalities. According to Equation (1), such a formulation allows the model to balance dynamically multimodal features that facilitate the interaction of linguistic signs obtained in text and paralinguistic signs obtained in audio and visual cues.

To counter the imbalance in the classes of the MELD dataset, especially in emotions that are not common, like fear, Focal Loss was utilized in the initial fusion models. The Focal Loss is formulated as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (2)$$

Where p_t represents the predicted probability of the true class t, α_t denotes the class-balancing factor, and $\gamma = 2$ acts as a focusing parameter that reduces the contribution of well-classified samples. As described in Equation (2), this mechanism emphasizes difficult or underrepresented emotion categories, enabling the model to learn more effectively from minority classes. The following sections present the experimental results that give the comparative analysis of the four proposed architectures. In the early methods of Fusion, it is the text, audio, and video modalities that are combined at the feature level with projection layers and then fully connected layers. The attention-free version is used as a baseline version, and the attention-based early fusion model applies hierarchical and cross-attention models to learn cross-modal contextual dependencies. However, the late fusion approaches handle the modality-specific features separately and then combine them later on to carry out joint classification of emotions. An attention-based late fusion model is another model that improves the modality-specific representations by using self-attention mechanisms before Fusion. The totality of these architectural variations enables the thorough exploration of the question of how fusion strategies and attention mechanisms affect multimodal emotion recognition performance.

The models are assessed with a Model Evaluator module, which computes the key performance indicators, such as loss, accuracy, weighted F1-score, and a report of detailed classification of every approach. In the evaluation stage, the models run a forward pass on the test data, but gradient calculation is turned down to enhance the computation speed. The predicted emotion labels will then be compared to the ground-truth labels to compute the performance metrics. Specifically, the F1-score is weighted in such a way that the class imbalance of the MELD dataset would be considered, so that minority emotion types would be represented in the evaluation. The obtained measures are aggregated into an overall assessment report, which allows conducting a systematic comparison between early fusion and late fusion structures with and without attention mechanisms. This assessment model creates a stable ground upon which the performance of every approach can be analysed, and the results of the experiments can be interpreted clearly.

6. Result and Discussion

The performance of the proposed models was tested on unknown test data after training them to test their ability to generalize. This evaluation has been carried out with the results presented in the given section.

Table 2. Early Fusion without attention result

	Precision	Recall	F1-Score	Support
0	0.51	0.45	0.48	192
1	0.67	0.05	0.10	38
2	0.23	0.10	0.14	31
3	0.56	0.50	0.53	247
4	0.70	0.85	0.77	707
5	0.47	0.20	0.28	116
6	0.51	0.55	0.53	164
Accuracy			0.62	1495
Macro Avg	0.52	0.39	0.40	1495
Weighted Avg	0.60	0.62	0.60	1495

Table 3. Early Fusion with attention result

	Precision	Recall	F1-Score	Support
0	0.57	0.40	0.47	192
1	0.71	0.13	0.22	38
2	0.38	0.10	0.15	31
3	0.60	0.55	0.58	247
4	0.71	0.88	0.78	707
5	0.46	0.16	0.24	116
6	0.52	0.61	0.56	164
Accuracy			0.64	1495
Macro Avg	0.56	0.41	0.43	1495
Weighted Avg	0.62	0.64	0.61	1495

6.1. Result of Early Fusion Model

The comparative analysis of the performance of Early Fusion without Attention (Table 2) and Early Fusion with Attention (Table 3) indicates that the differences can be viewed in all the evaluation measures under the same set of experimental conditions. The precision is also increased as compared to the non-attention model, which is 0.62, to the attention-based model, which is 0.64, showing a significant difference in the overall classification accuracy. Likewise, the weighted F1-score gets improved by 0.60 to 0.61, indicating a small enhancement in dealing with the class imbalance among categories of emotion. Similar differences can be found in terms of precision, recall, and F1-scores among emotion classes at the level of the classes. As an example, class 4 (the most dominant class) experiences an increase in F1-score of 0.77 to 0.78, and an increase in the recall of 0.85 to 0.88, denoting improved identification of the most frequent samples. Similarly, class 3 shows that the F1-score is raised, and now equals 0.53 to equal 0.58, which indicates the enhanced classification consistency. Nevertheless, some minority classes like class 1 and class 2 still exhibit relatively low recall values in both methods, which indicates the ongoing issue of class imbalance. A class-level breakdown of the attention effect in early Fusion reveals a consistent pattern: attention mechanisms provide the greatest benefit to mid-frequency emotion classes such as class 3 (anger, F1: 0.53→0.58) and class 6 (sadness, F1: 0.53→0.56), while contributing minimally to the majority class 4 (neutral, F1: 0.77→0.78). This suggests that cross-modal and hierarchical attention mechanisms are particularly effective at resolving ambiguity in moderately represented emotion categories, where textual, acoustic, and visual cues provide complementary rather than redundant information. In contrast, minority classes 1 (disgust, F1: 0.10→0.22) and 2 (fear, F1: 0.14→0.15) show marginal or inconsistent improvement even with attention, indicating that attention alone is insufficient to compensate for severe class underrepresentation. The Focal Loss objective applied in both early fusion models partially mitigates this imbalance by up-weighting hard examples during training, yet the inherent scarcity of disgust and fear samples in MELD (38 and 31 test instances, respectively) limits the upper bound of achievable recall for these classes. These findings collectively indicate that while attention enhances cross-modal interaction in early Fusion, targeted data augmentation or synthetic oversampling strategies for minority emotion classes remain an important direction for future work.

Table 4. Late Fusion without attention result

	Precision	Recall	F1-Score	Support
0	0.37	0.57	0.45	345
1	0.00	0.00	0.00	68
2	0.00	0.00	0.00	50
3	0.64	0.43	0.51	402
4	0.71	0.87	0.78	1256
5	0.53	0.10	0.16	208
6	0.57	0.51	0.54	281
Accuracy			0.62	2610
Macro Avg	0.40	0.35	0.35	2610
Weighted Avg	0.59	0.62	0.59	2610

Table 5. Late Fusion with attention result

	Precision	Recall	F1-Score	Support
0	0.40	0.64	0.49	345
1	0.00	0.00	0.00	68
2	0.25	0.02	0.04	50
3	0.62	0.50	0.55	402
4	0.75	0.82	0.78	1256
5	0.50	0.22	0.31	208
6	0.57	0.54	0.55	281
Accuracy			0.63	2610
Macro Avg	0.44	0.39	0.39	2610
Weighted Avg	0.61	0.63	0.61	2610

6.2. Result of Late Fusion Model

Tables 4 and 5 give the comparative data of the late fusion strategies with and without attention mechanisms, respectively. The general accuracy has a slight improvement between 0.62 (when there is no attention) and 0.63 (when attention mechanisms are involved), which demonstrates that there is a measurable difference in the performance in the case when attention mechanisms are involved. On the same note, the weighted F1-score increases by 0.59 to 0.61, which indicates a slight increase in the ability to manage the class imbalance in each of the categories of emotions. There are significant differences at the level of classes. In the case of class 0, the F1-score is also improved, with the F1-score going up by 0.49 to 0.64, and Recall also going up by 0.57 to 0.64, meaning that this class is better with attention. Class 3 also demonstrates that the F1-score improved from 0.51 to 0.55, which indicates a steadier classification. In class 5, the F1-score goes up to 0.31, as compared to 0.16, implying an enhanced ability in detecting the comparatively less prominent class. Likewise, the score of the F1-score of class 6 slightly increases (0.54 to 0.55). Nevertheless, some classes, like class 1, do not change with an F1-score of 0.00 in both methods, which means that it is still hard to detect this type of class. Class 2 demonstrates that there is a slight increase between 0.00 and 0.04 in F1-score, though the recall is still low, which indicates that there is still a problem in the recognition of minority classes.

The macro average measures have also increased, with the F1-score rising from 0.35 to 0.39, which suggests a more balanced performance in all classes when the attention mechanisms are used. The same holds for precision and recall, which indicate a better representation of various classes. In general, the results in Tables 4 and 5 will provide a systematic comparison of late fusion strategies with and without attention mechanisms. The statistical variations between the measures of evaluation permit the analysis of the effects of attention mechanisms on the refinement of features of the modality and multimodal integration in the emotional recognition tasks.

6.3. Comparative Analysis

T, A, and V denote text, audio, and video modalities, respectively. Tables 6 and 7 represent the proposed methods introduced in this study and prior multimodal baseline approaches. The Weighted F1-score is reported as a percentage (%). Early Fusion is denoted as (EF), whereas Late Fusion is represented as (LF). All values for prior works are directly taken from their respective published sources.

Table 6. Other models trained on the MELD dataset

Method	Modalities	Features / Encoders	Weighted F1 (%)	Notes
DialogueGCN [31]	T+A+V	GloVe, OpenSmile, DenseNet	58.10	GCN-based
DialogueCRN [32]	T	RoBERTa	58.39	Contextual RNN
MMGCN [33]	T+A+V	TextCNN, OpenSmile, DenseNet	58.65	Multimodal GCN

Table 7. Our models trained on the MELD dataset

Method	Modalities	Features / Encoders	Weighted F1 (%)	Notes
EF without Attention	T+A+V	BERT, Wav2Vec, TimesFormer	60.00	Early Fusion, Focal Loss
EF with Attention	T+A+V	BERT, Wav2Vec, TimesFormer	61.00	Cross + Hierarchical Attention
LF without Attention	T+A+V	BERT, Wav2Vec, TimesFormer	59.00	Late Fusion, Label Smooth
LF with Attention	T+A+V	BERT, Wav2Vec, TimesFormer	61.00	Self-Attention Refinement

The results presented in Tables 6 and 7 provide a focused comparison between selected prior multimodal approaches and the proposed fusion-based architectures on the MELD dataset. Baseline models such as DialogueGCN (58.10%), DialogueCRN (58.39%), and MMGCN (58.65%) demonstrate moderate performance, relying on graph-based reasoning and conventional multimodal feature extraction techniques.

In contrast, the proposed models consistently achieve improved performance, with early Fusion with attention and late Fusion with attention both reaching a weighted F1-score of 61.00%, outperforming the selected baselines. Even the non-attention variants (60.00% and 59.00%) remain competitive, indicating the effectiveness of the underlying multimodal representation.

A key strength of the proposed framework lies in the integration of state-of-the-art pretrained encoders, namely BERT, Wav2Vec, and TimesFormer, enabling robust and consistent trimodal feature extraction. Furthermore, unlike prior works, this study provides a controlled and systematic evaluation of fusion strategies (early vs late) and attention mechanisms, allowing clear insight into their individual contributions to performance.

Overall, the results highlight that the proposed architectures achieve a strong balance between performance and design simplicity, while offering a structured analysis of multimodal fusion strategies for emotion recognition.

7. Conclusion

This paper made a comparative analysis of four multimodal emotion recognition techniques that used early fusion and late fusion strategies, and each of them was considered with the presence and absence of attention mechanisms. The findings illustrate that the combination of attention mechanisms results in differences in the evaluation measures, such as accuracy, weighted F1-score, and macro F1-score. In all the methods, the results indicate the significance of multimodal feature fusion in enhancing emotion classification judgments. Although both early and late fusion strategies are effective in integrating the information of the text, audio, and video modalities, they differ in their performance, which suggests that the interaction between multimedia is modelled differently. The findings also demonstrate the same pattern, with major classes of emotions constantly performing better, whereas minority ones are difficult to categorize, despite the presence of complex fusion and attention processes. In general, the comparative analysis informs about the influence of fusion strategies and mechanisms of attention on multimodal emotion recognition. The presented results provide a systematic perception of the impact of various architectural decisions on performance, which enables the readers to explain the efficiency of each solution on the basis of the numerical results presented.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Qingping Zhou, "HGLER: A Hierarchical Heterogeneous Graph Networks for Enhanced Multimodal Emotion Recognition in Conversations," *PLoS One*, vol. 20, no. 9, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Yuntao Shou et al., "A Comprehensive Survey on Multi-Modal Conversational Emotion Recognition with Deep Learning," *ACM Transactions on Information Systems*, vol. 44, no. 2, pp. 1-48, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[3] Wei Dai et al., "A Novel Approach for Multimodal Emotion Recognition: Multimodal Semantic Information Fusion," *arXiv preprint*, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[4] Yuanyuan Sun, and Ting Zhou, "DialogueMLLM: Transforming Multimodal Emotion Recognition in Conversation through Instruction-Tuned MLLM," *IEEE Access*, vol. 13, pp. 121048-121060, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Chhavi Dixit, and Shashank Mouli Satapathy, "Deep CNN with Late Fusion for Real Time Multimodal Emotion Recognition," *Expert Systems with Applications*, vol. 240, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Yong Zhang, Cheng Cheng, and YiDie Zhang, "Multimodal Emotion Recognition based on Manifold Learning and Convolution Neural Network," *Multimedia Tools and Applications*, vol. 81, pp. 33253-33268, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jiarong He, "A Multimodal Approach for Emotion Recognition in Conversations Using the MELD Dataset," *2025 Asia-Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC)*, Rimini, Italy, pp. 54-58, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Yun Liu et al., "Toward Multimodal Sentiment Analysis with a Self-Supervised Knowledge-Augmented Network," *Expert Systems with Applications*, vol. 314, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[9] Wuzhen Shi et al., "Identity and Modality Attributes Driven Multimodal Fusion Networks for Emotion Recognition in Conversations," *IEEE Transactions on Multimedia*, vol. 27, pp. 4361-4371, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Shiyun Zhao, Jinchang Ren, and Xiaojuan Zhou, "Cross-modal Gated Feature Enhancement for Multimodal Emotion Recognition in Conversations," *Scientific Reports*, vol. 15, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Gyanendra K. Verma, and Uma Shanker Tiwary, "Multimodal Fusion Framework: A Multiresolution Approach for Emotion Classification and Recognition from Physiological Signals," *NeuroImage*, vol. 102, no. 1, pp. 162-172, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[12] Komal Anadkat et al., "Enhancing Emotion Recognition with Multimodel Approach using Deep Neural Networks," *Reliability: Theory & Applications*, vol. 20, no. 1, pp. 1-13, 2025. [Google Scholar] [Publisher Link]

[13] Samuel Kakuba, Alwin Poulose, and Dong Seog Han, "Deep Learning Approaches for Bimodal Speech Emotion Recognition: Advancements, Challenges, and a Multi-Learning Model," *IEEE Access*, vol. 11, pp. 113769-113789, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Mingyu, Zhou Jiawei, and Wei Ning, "AFR-BERT: Attention-based Mechanism Feature Relevance Fusion Multimodal Sentiment Analysis Model," *PLoS One*, vol. 17, no. 9, pp. 1-20, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Licai Sun et al., "MAE-DFER: Efficient Masked Autoencoder for Self-supervised Dynamic Facial Expression Recognition," *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6110-6121, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Alexei Baevski et al., "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *arXiv preprint*, pp. 1-19, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is Space-Time Attention All you Need for Video Understanding?," *arXiv Preprint*, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Cheng Cheng et al., "Dense Graph Convolutional with Joint Cross-Attention Network for Multimodal Emotion Recognition," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 5, pp. 6672-6683, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Yunhong Liao et al., "Attention-Driven Adaptive Deep Canonical Correlation Analysis for Multimodal Sentiment Analysis with EEG and Eye Movement Data," *2025 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, pp. 409-415, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Mohammad Soleymani, Maja Pantic, and Thierry Pun, "Multimodal Emotion Recognition in Response to Videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211-223, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[21] Wei-Bang Jiang et al., "SEED-VII: A Multimodal Dataset of Six Basic Emotions with Continuous Labels for Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 16, no. 2, pp. 969-985, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[22] Nicu Sebe, Ira Cohen, and Thomas S. Huang, *Multimodal Emotion Recognition*, Handbook of Pattern Recognition and Computer Vision, 4th ed., pp. 387-409, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[23] Panagiotis Tzirakis et al., "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301-1309, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[24] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu, "Emotion Recognition using Multimodal Deep Learning," *23rd International Conference Neural Information Processing*, Kyoto, Japan, vol. 9948, pp. 521-529, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[25] You Wu, Qingwei Mi, and Tianhan Gao, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions," *Biomimetics*, vol. 10, no. 7, pp. 1-24, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[26] Huan Liu et al., "EEG-Based Multimodal Emotion Recognition: A Machine Learning Perspective," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1-29, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Wei-Long Zheng et al., "EmotionMeter: A Multimodal Framework for Recognizing Human Emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110-1122, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[28] Zebang Cheng et al., "Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning," *arXiv preprint*, pp. 110805-110853, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Samira Hazmoune, and Fateh Bougamouza, "Using Transformers for Multimodal Emotion Recognition: Taxonomies and State of the Art Review," *Engineering Applications of Artificial Intelligence*, vol. 133, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Cristina Luna-Jiménez et al., "Multimodal Emotion Recognition on Ravdess Dataset using Transfer Learning," *Sensors*, vol. 21, no. 22, pp. 1-29, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[31] Deepanway Ghosal et al., "DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, pp. 154-164, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[32] Dou Hu et al., "DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1, pp. 7042-7052, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[33] Jingwen Hu et al., "MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1, pp. 5666-5675, 2021. [CrossRef] [Google Scholar] [Publisher Link]