*2025 International Conference on Intelligent Computing*
*July 26-29, Ningbo, China*
https://www.ic-icc.cn/2025/index.php

# MMCFusionNet: A Multimodal Mixture of Experts and Collaborative Attention Fusion Network for Abnormal Emotion Recognition

Haitao Xiong, Xin Zhou, Wei Jiao, and Yuanyuan Cai[*]

School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China
caiyuanyuan@btbu.edu.cn

**Abstract.** The growing popularity of short videos on social media has introduced new challenges for content moderation, particularly in detecting abnormal emotions like hate and sarcasm. These emotions usually exhibit higher concealment and multimodal inconsistency compared to conventional ones. While prior studies have primarily focused on conventional emotion recognition, research on abnormal emotions remains limited. Moreover, existing models often fail to leverage the complementary nature of multimodal data fully and lack robust intermodal interactions. This study proposes MMCFusionNet, a novel multimodal fusion framework designed for abnormal emotion recognition in short videos. The model extracts and aligns features from four modalities (text, visual, audio, and facial) through a dedicated feature encoder and alignment module to improve the ability of hate and sarcasm emotion recognition. At its core, the model integrates two key mechanisms: 1) Mixture of Experts (MoE) modules to enhance intramodal representations across temporal frames for identifying concealed emotional cues; 2) Dual-channel collaborative attention (Co-Attention) modules to facilitate intermodal complementarity for resolving multimodal contradictions. Experimental results on the HateMM and MUStARD datasets show that MMCFusionNet outperforms baseline models across various evaluation metrics, with ablation studies confirming the effectiveness and robustness of each module.

**Keywords:** Emotion recognition; Multi-modal learning; Multi-modal fusion; Mixture of Experts; Collaborative Attention

## 1 Introduction

Short videos have become a pivotal medium for emotional expression and information dissemination [1]. The exponential growth of platforms like YouTube and TikTok has increased the scope and impact of emotional communication. Traditional manual content moderation mechanisms for short videos are criticized for labour concerns, lack of transparency, perpetuating biases, and potential harm to marginalized communities, making them inadequate for handling the sheer volume of content [2]. To address this challenge, the development of automated emotion analysis tools is necessary.

Multimodal technology, through the integration of text, images, audio, and other modalities, can deeply analyze complex emotional signals in videos. It surpasses the limitations of single-modal methods and has become a hotspot in emotion recognition [3].

In emotion recognition tasks, identifying abnormal emotions (e.g., hate, sarcasm, violence) is more urgent than detecting conventional emotions (e.g., happiness, sadness, neutrality). This is due to their potential threats to individual cognition and social order [4]. Abnormal emotions are often conveyed through incendiary language, symbolic images, manipulative audio, and contradictory facial expressions. They subtly exacerbate social fragmentation and group antagonism. For instance, a shooter once exploited the contradictory combination of a gentle narration and violent imagery to disseminate hate, triggering over a dozen copycat crimes worldwide. Similarly, a short video used a "praising" tone coupled with eye-rolling expressions to spread false information about epidemic prevention and mislead the public. This ironic sentiment, characterized by a disconnect between surface semantics and underlying intent, resulted in decreased government credibility and increased resistance to policies. These examples illustrate that, unlike traditional emotions, abnormal emotions are highly concealed and multimodally contradictory, making their recognition particularly challenging [5]. Therefore, the core of the abnormal emotion detection tasks lies in the dual perception mechanism of modality features and cross-modal interactions [6].

Recent research has made some progress. Hu et al. [7] proposed a joint network for speech emotion recognition that combines a pre-trained model and a spectral model. By designing different interaction attention modules to fuse the intermediate features and optimizing the joint network with a multi-branch training strategy, this method accurately recognizes emotions such as anger, happiness, sadness, and neutrality on the IEMOCAP dataset [8]. Liu et al. [9] proposed the TMSON model, which estimates the uncertainty distribution of each modality to address issues such as noise, semantic ambiguity, and missing modalities in multimodal data. This method employs Bayesian rules to fuse single-modality distributions, accurately recognizing positive and negative emotions on datasets such as CMU-MOSI [10]. Rana et al. [11] proposed a multimodal deep learning framework to combine the acoustic features representing emotion and the textual features to detect hateful content. Waligora et al. [12] extracted spatiotemporal features from facial, voice, or biosensors modalities for emotion recognition and pain estimation respectively. It relies on a joint multimodal transformer for fusion with key-based cross-attention to dynamically learn information between modalities.

However, these studies still have limitations: 1) They mostly focus on conventional emotion recognition, with relatively scarce attention to abnormal emotions like hate, sarcasm, and violence. 2) Most existing methods primarily utilize video, audio, and text modalities, seldom incorporating a comprehensive set of modalities, including facial expressions. Facial expressions, such as furrowed brows, glaring eyes, and clenched teeth, can also provide crucial cues for detecting emotions in videos. 3) Previous methods often tend to connect different modalities directly during modal fusion without considering the interaction between modalities. An effective combination of different modalities can boost performance by fully exploiting complementary information.

To address these issues, this study proposes a model for recognizing hate and sarcasm. Our model leverages four modalities: visual, audio, textual, and facial. It preprocesses and extracts multimodal features, aligns features within the spatial domain, enhances intramodal features using multimodal mixture of experts (MoE) modules, and strengthens intermodal interactions through dual-channel collaborative attention (Co-Attention) modules. The main contributions of this work are as follows:

- We propose a model MMCFusionNet, designed for recognizing abnormal emotions, effectively integrating features from text, visual, audio, and facial modalities. It significantly improves the accuracy of abnormal emotion recognition and demonstrates good generalization ability.
- We introduce a multimodal feature fusion method that combines MoE modules and Co-Attention mechanisms, enhancing both intramodal feature representation and intermodal complementarity, improving the model's effectiveness and robustness.
- Experimental evaluations on the HateMM [13] and MUStARD [14] datasets demonstrate that our model outperforms existing baseline methods. A series of ablation experiments further validate the rationality of the module design and confirm the individual contributions of each component.

## 2 Method

### 2.1 Task Definition

Given a video, the abnormal emotion recognition task aims to classify it as either abnormal emotion ($y = 1$) or non-abnormal emotion ($y = 0$). The visual modality of the video can be represented as a sequence of video frames $F = \{frame_1,\ frame_2, \ldots,$ $frame_n\}$, the associated audio $A = \{\alpha udio_1,\ \alpha udio_2, \ldots,\ \alpha udio_n\}$, the transcribed text of the audio $T = \{word_1,\ word_2, \ldots,\ word_m\}$, and the sequence of faces extracted from the video frames $Face = \{face_1,\ face_2, \ldots,\ face_k\}$. The classifier is defined as $(F; A; T; Face) \rightarrow y$, where $y \in \{0, 1\}$ represents the video label.

### 2.2 Overview

The MMCFusionNet model comprises three main components, as shown in **Fig. 1**. The multimodal feature encoder extracts features from four modalities: text (using fine-tuned BERT [15]), visual (using ViT [16]), audio (using MFCC [17]), and facial (using InsightFace [18]). The Alignment Module aligns these features using LSTM and MLP. The Fusion Module interacts and fuses features using MoE and Co-Attention modules.

### 2.3 Multimodal Feature Encoder Module

The video is sampled at a rate of one frame per second to obtain 100 frames for each video segment. For videos with fewer than 100 frames, we add an image with a white background as padding. For videos with more than 100 frames, 100 frames are uniformly sampled from the total available frames. After obtaining video frames, ViT is used to learn the features of each frame, resulting in visual features $X_v \in \mathbb{R}^{n \times 768}$.

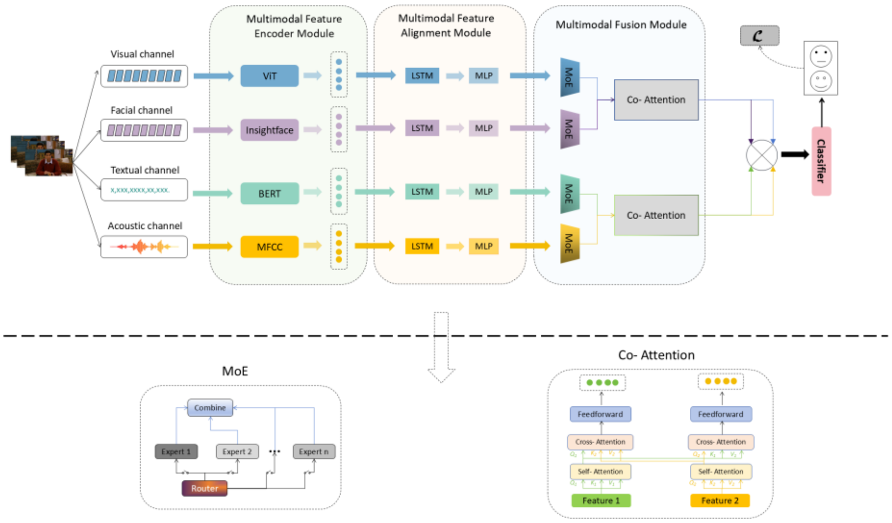


**Fig. 1.** Framework of MMCFusionNet. The video data is divided into video frames, audio, text extraction, and facial recognition. Each modality is encoded using specialized deep-learning models (ViT, MFCC, BERT, and InsightFace). Features are aligned using LSTM and MLP. The fusion module employs MoE and dual-channel Co-Attention, followed by classification.

The audio modality is processed by dividing the audio into 100 segments. Mel-frequency cepstral Coefficients (MFCCs) are extracted from each segment to capture the prosodic and phonetic characteristics. The resulting acoustic feature sequence is denoted as $X_a \in \mathbb{R}^{n \times 128}$.

For the textual modality, the content is transcribed from audio using Automatic Speech Recognition (ASR) technology [19]. We truncate and pad the text information to align sequence lengths. Through systematic analysis of the dialogue text length distributions, we use truncation thresholds: a fixed sequence length of 200 tokens for HateMM and 100 tokens for MUStARD. Features are then extracted using a fine-tuned BERT model. The text feature sequence is denoted as $X_t \in \mathbb{R}^{n \times 768}$.

Facial features are extracted using InsightFace, an open-source tool, to detect and capture facial expressions from each sampled frame. InsightFace utilizes deep learning methods to detect and localize human faces accurately, producing bounding box coordinates, facial key points, and discriminative facial features. The facial feature sequence is represented as $X_f \in \mathbb{R}^{n \times 512}$ .

## 2.4 Multimodal Feature Alignment Module

The dimensions of features extracted by each modality encoder differ. To refine and align these features while considering their interdependencies, we employ Long Short-Term Memory (LSTM) [20]and Multilayer Perceptron (MLP). These methods not only achieve dimensionality reduction and alignment of multimodal features but also significantly enhance their representational power.

LSTM is primarily used to capture and store temporal dependencies within each modality. By applying LSTM to the feature sequences of all four modalities, we enhance the representations while simultaneously reducing the dimensionality of each modality's features. MLP is then used to further process the features output by LSTM. It is capable of learning more complex feature representations. By adjusting the number of layers and neurons per layer, MLP provides flexibility in controlling the model's capacity. Each layer of the MLP uses the ReLU activation function to introduce non-linearity, further strengthening the model's ability to express intricate feature relationships. The specific formulas are shown as follows.

$$h_m = W_m^2\ ( Relu\ ( W_m^1\ LSTM\ ( X_m\ ) + b_m^1) + \ b_m^2 \quad , m \in \{v,a,t,f\} \tag{1}$$

In these equations, $h_m$ denotes the aligned feature for modality m, $v,a,t,f$ represent four modalities respectively. $W_m^1$ and $b_m^1$ represent the weight matrices and biases for the first fully connected layer acting on each modality. Similarly, the second fully connected layer weights $W_m^2$ and biases $b_m^2$ follow the same notation.

### 2.5 Multimodal Fusion Module

The multimodal fusion module is central to our model. Each video frame has unique information, requiring specialized enhancement. Previous methods often ignored the distinct sequential characteristics and complementary natures between modalities. To address this, we employ multimodal MoE modules [21] to strengthen features from different frames within a modality. In addition, we use dual-channel Co-Attention [22] to facilitate effective intermodal interactions and capture complementary information between modalities. This mechanism allows the model to prioritize crucial information within each modality while filtering out irrelevant details.

**Mixture of Experts.** The MoE module accelerates convergence and improves performance without adding significant complexity. It computes the weights of various experts via routing, multiplies them with corresponding features, and sums the top-k features. More specifically, this process begins with the computation of logits for each expert through the gating network, which assesses the input features $h_m$ and outputs a set of logits $\ell_m$. A Top-K selection mechanism is then applied to introduce sparsity by retaining the K highest-scoring experts. These selected logits are normalized using the softmax function to form a probability distribution. Finally, the top-K features, weighted by these probabilities, are summed to produce the final output $M_m$ for each modality. The feature enhancement process is formalized as follows:

$$\ell_m = h_m * \mathrm{W}_m + \mathrm{b}_m \quad , m \in \{v,a,t,f\} \tag{2}$$

$$TopK(\ell_m)_i = \begin{cases} \ell_{m_i}, & if\ \ell_{m_i}\ is\ the\ topK\ elements. \\ -\infty, & otherwise. \end{cases} \tag{3}$$

$$G(h_m) = Softmax\big(TopK(\ell_m)\big) \tag{4}$$

$$M_m = KeepTop\ ( G(h_m) * h_m\ , k) \tag{5}$$

where $G(h_m)$ represents the weights of different expert modules for the modality m, and $KeepTop$ represents the enhanced feature obtained by summing the top-k features.

**Dual-channel Co-Attention.** In the context of abnormal emotion recognition in short videos, the visual modality primarily provides contextual background through capturing scene and action information, while the facial modality offers emotional cues through recognizing facial expressions. The same action paired with different facial expressions can convey different emotions. Likewise, the acoustic modality primarily captures vocal intonation and background music, while the textual modality offers semantic insights through transcribing spoken words. The same text combined with varying intonations can express different emotions. Thus, considering computational complexity and correlation degrees between modalities for the specific task, we use a dual-channel collaborative attention mechanism, focusing on $M_v - M_f$ and $M_t - M_a$ pairs. The processes are shown in the following equations:

$$m_t, m_a = CoAttention(M_t; M_a) \tag{6}$$

$$m_v, m_f = CoAttention\big(M_v; M_f\big) \tag{7}$$

$m_t, m_a$ represent the features obtained through collaborative attention between textual and acoustic modalities. $m_v, m_f$ represent the features obtained through collaborative attention between visual and facial modalities.

Finally, the features from all four modalities are concatenated and passed through a classifier for emotion recognition. The process is described as:

$$Z = [m_t; m_a; m_v; m_f] \tag{8}$$

$$\hat{y} = W_z Z + b_z \tag{9}$$

where $Z$ is the concatenated multimodal fusion feature, $W_z \in \mathbb{R}^{4d*\mathrm{m}}$ represents the weight matrix, and $b_z \in \mathbb{R}^{1*\mathrm{m}}$ is the bias. $[;]$ indicates the vector concatenation.

To obtain the optimized parameters, we minimize cross-entropy loss between a predicted probability and a ground-truth label, which is computed as follows:

$$\mathcal{L} = -\big(ylog(\hat{y}) + (1-y)log(1-\hat{y})\big) \tag{10}$$

# 3 Experiments

## 3.1 Implementation Details

The experiments are conducted on a Mac with an M1 chip, running the Sequoia operating system. We use the PyTorch 1.13.1 deep learning framework with Python 3.9 as the interpreter. We select Adam optimizer for weight updates and adopt five-fold cross-validation to validate performance. The dataset is split into training, validation, and test sets with a ratio of 7:1:2. Detailed experimental parameters are shown in **Table 1**.

**Table 1.** The experimental parameters used for the two datasets.

| **Dataset** | | **HateMM** | **MUStARD** |
|---|---|---|---|
| Batch Size | | 32 | 32 |
| Learning Rate | | 0.0001 | 0.0001 |
| Number of Experts | Total | 8 | 8 |
| | Text | 3 | 3 |
| | Visual | 5 | 3 |
| | Audio | 3 | 3 |
| | Facial | 5 | 3 |

### 3.2 Datasets and Baselines

Two publicly available datasets are used: HateMM and MUStARD. Based on these datasets, we focus on two types of abnormal emotions: hate and sarcasm.

The HateMM dataset, introduced by a research team at the Indian Institute of Technology in 2023, consists of 1083 video clips. Sourced from the real social media platform BitChute, this dataset was curated through manual annotation and filtering. The dataset contains two emotion labels: hate and non-hate. The ratio of hate to non-hate samples is approximately 4:6. The MUStARD dataset is specifically designed for detecting sarcastic emotions. Sarcasm is a complex linguistic phenomenon that involves a discrepancy between literal meaning and underlying intent. It comprises 690 videos sourced from popular TV shows such as The Big Bang Theory and Friends, with a nearly 1:1 ratio of sarcastic to non-sarcastic samples.

To validate the performance of the proposed MMCFusionNet model, we compare it with the following baseline methods:

- FEF-Net [23]: A multimodal humor prediction model designed to improve accuracy by reducing redundancy in auxiliary modalities (e.g., audio and video). It uses cross-modal attention to enrich these modalities with knowledge from the text modality and assigns appropriate weights to redundant features across time slices, minimizing the impact of irrelevant information.
- MCER [24]: A method that introduces a gated joint multimodal fusion mechanism. The model integrates features from text, speech, and video modalities, considering contextual and speaker information to enhance performance. It recognizes emotions in sarcastic statements and is evaluated using the MUStARD dataset.
- PriSA [25]: A model that combines preference fusion with distance-aware contrastive learning. By employing self-attention modules, it efficiently extracts emotional information from the combined modality-relatedness and unique intramodal features of visual and audio modalities.

### 3.3 Results Analysis

To evaluate the effectiveness of the MMCFusionNet model, five commonly used evaluation metrics in multimodal emotion analysis tasks are used: Accuracy (Acc), F1

Score (F1), Macro-average F1 Score (M-F1), Precision (P), and Recall (R) [26]. We assess models using Text (T), Visual (V), Audio (A), and Facial (F) modalities in single, three, and four-modal configurations. Single-modal uses MLP classifiers. The three-modal group (T+V+A) includes FEF-Net, MCER, and PriSA. The four-modal group (T+V+A+F) includes a naive concatenation baseline (ConcatFusion) and our proposed MMCFusionNet. **Table 2** presents experimental results on two datasets.

The results show that MMCFusionNet consistently outperforms baseline models across most evaluation metrics. Single-modal models exhibit relatively poor performance, while three-modal models show significant improvements, highlighting the complementary nature of multimodal information and the benefits of combining them. Notably, our model still surpasses these models. The simple concatenation of features in the ConcatFusion model performs worse than our advanced multimodal fusion approach, emphasizing the importance of mechanisms like MoE and Co-Attention.

**Table 2.** Comparison of experimental results.

| Dataset | HateMM | | | | | MUStARD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Modality&Model | Acc | M-F1 | F1 | P | R | Acc | M-F1 | F1 | P | R |
| Single | | | | | | | | | | |
| Text (T) | 74.8 | 71.7 | 62.4 | 78.8 | 51.8 | 70.6 | 70.3 | 69.8 | 70.1 | 70.6 |
| Visual (V) | 72.5 | 69.8 | 60.8 | 72.5 | 52.7 | 67.4 | 67.1 | 68.0 | 66.9 | 70.0 |
| Audio (A) | 71.7 | 68.2 | 57.8 | 73.2 | 47.7 | 63.2 | 62.7 | 59.8 | 64.8 | 56.5 |
| Facial (F) | 67.0 | 58.5 | 56.8 | 76.5 | 45.3 | 67.0 | 66.7 | 65.2 | 67.8 | 63.5 |
| Three (T+V+A) | | | | | | | | | | |
| FEF-Net | 78.5 | 76.3 | 68.2 | 79.4 | 59.6 | 71.2 | 71.0 | 69.5 | 72.4 | 67.1 |
| MCER | 81.8 | 80.6 | **76.2** | 76.9 | **76.6** | 70.9 | 70.8 | 70.5 | 70.6 | 70.9 |
| PriSA | 80.0 | 78.8 | 73.9 | 74.6 | 74.2 | 71.6 | 71.2 | 68.7 | 73.9 | 65.3 |
| Four (T+V+A+F) | | | | | | | | | | |
| ConcatFusion | 80.9 | 79.5 | 74.2 | 82.2 | 67.8 | 73.3 | 73.1 | 71.5 | 75.6 | 68.2 |
| **MMCFusionNet** | **83.4** | **82.1** | 74.6 | **83.1** | 67.5 | **76.4** | **76.3** | **76.1** | **76.0** | **76.8** |

Additionally, the performance of almost all models on the MUStARD dataset is lower than that on the HateMM dataset. This is largely due to the smaller size of the MUStARD dataset, which makes it more challenging for the model to learn effectively. In conclusion, our model's advanced multimodal alignment and fusion strategies significantly outperform traditional unimodal and simple multimodal models, highlighting the value of advanced fusion techniques in improving emotion recognition tasks.

### 3.4 Ablation Study

To verify the contributions of facial features and different modules to the model, we conduct three ablation experiments on both datasets. Specifically, "w/o Co-Attention" indicates the removal of the dual-channel collaborative attention module, "w/o MoE" represents the removal of the MoE module, and "w/o F" denotes the exclusion of the facial modality. **Table 3** shows the results of ablation experiments.

The results reveal that MMCFusionNet consistently outperforms other models across most evaluation metrics on both datasets. Each module—facial information, the MoE module, and the co-attention mechanism—contributes significantly to the model's performance. The integration of these components collectively enhances the model's ability to recognize abnormal emotions. Specifically, the facial information provides critical visual cues, the MoE module effectively integrates multimodal data, and the co-attention mechanism ensures robust feature fusion. Together, these elements synergistically improve the overall efficacy of the model.

Notably, both the "w/o MoE" and "w/o Co-Attention" models perform better than the "w/o F" model, underscoring the critical role of the facial modality in hate and sarcasm detection from videos. This performance disparity highlights that while the co-attention mechanism and the MoE module are pivotal, integrating facial modality among other modalities is indispensable for achieving optimal detection accuracy.

**Table 3.** Ablation experiment results.

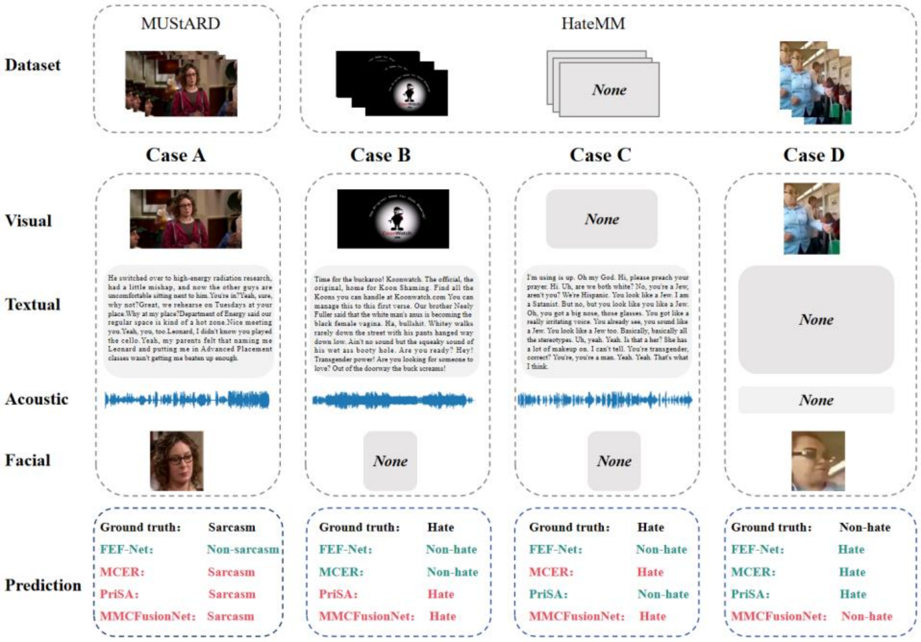
| Dataset | HateMM | | | | | MUStARD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Acc | M-F1 | F1 | P | R | Acc | M-F1 | F1 | P | R |
| w/o Co-Attention | 82.4 | 80.1 | 72.2 | 78.9 | 66.5 | 73.5 | 73.3 | 75.0 | 70.4 | **80.6** |
| w/o MoE | 81.3 | 81.3 | 73.1 | 80.4 | 66.9 | 75.5 | 75.4 | 74.5 | **76.5** | 72.9 |
| w/o F | 77.6 | 75.4 | 68.0 | 80.7 | 58.9 | 72.3 | 71.4 | 68.7 | 69.9 | 67.7 |
| **MMCFusionNet** | **83.4** | **82.1** | **74.6** | **83.1** | **67.5** | **76.4** | **76.3** | **76.1** | 76.0 | 76.8 |

### 3.5 Case-driven Analysis

Four typical cases are shown to analyze the performance in **Fig. 2**. For case A, we select a sarcastic video from MUStARD. This case represents one of the primary types in both datasets, containing all modalities with clear video frames and complete audio. In this video, several people engage in a conversation, using exaggerated and self-deprecating humor to mock the government's overreaction to minor issues and people's excessive attention to trivial matters. Our model accurately identified the emotional category of this video as sarcasm.

For Cases B and C, we select hate videos from HateMM, while Case D is a non-hate video from the same dataset. In Case B, the facial modality is missing, but the other modalities capture the use of the racial slur "coon" to demean and insult Black people. Case C lacks both visual and facial modalities, containing only audio and text that reference negative stereotypes about Jewish people. Case D contains only the video track, thus missing both text and audio modalities. From the perspective of video content moderation processes, recognizing abnormal emotions in videos with missing modalities is challenging. However, our model's predictions in these three cases are accurate and outperform the baseline models.

These cases illustrate that MMCFusionNet's multimodal fusion strategy is effective and stable even when the modalities are incomplete. This approach enables accurate recognition and complementary integration of modality information.



He switched over to high-energy radiation research, had a little mishap, and now the other guys are uncomfortable sitting next to him.You're in?Yeah, sure, why not?Great, we rehearse on Tuesdays at your place.Why at my place?Department of Energy said our regular space is kind of a hot zone.Nice meeting you.Yeah, you, too.Leonard, I didn't know you played the cello.Yeah, my parents felt that naming me Leonard and putting me in Advanced Placement classes wasn't getting me beaten up enough.

Time for the buckaroo! Koonwatch. The official, the original, home for Koon Shaming. Find all the Koons you can handle at Koonwatch.com You can manage this to this first verse. Our brother Neely Fuller said that the white man's anus is becoming the black female vagina. Ha, bullshit. Whitey walks rarely down the street with his pants hanged way down low. Ain't no sound but the squeaky sound of his wet ass booty hole. Are you ready? Hey! Transgender power! Are you looking for someone to love? Out of the doorway the buck screams!

I'm using is up. Oh my God. Hi, please preach your prayer. Hi. Uh, are we both white? No, you're a Jew, aren't you? We're Hispanic. You look like a Jew. I am a Satanist. But no, but you look like you like a Jew. Oh, you got a big nose, those glasses. You got like a really irritating voice. You already see, you sound like a Jew. You look like a Jew too. Basically, basically all the stereotypes. Uh, yeah. Yeah. Is that a her? She has a lot of makeup on. I can't tell. You're transgender, correct? You're, you're a man. Yeah. Yeah. That's what I think.

**Fig. 2.** Comparison of prediction cases between MMCFusionNet and baseline models. The prediction results marked in green font are incorrect, while those in red font are correct.

### 3.6 Discussion: MoE Mechanism for Fast Convergence

This section evaluates the MoE module's impact on training convergence speed. We compare loss trends between MMCFusionNet and the variant without MoE (w/o MoE) over 20 epochs, using five-fold cross-validation.

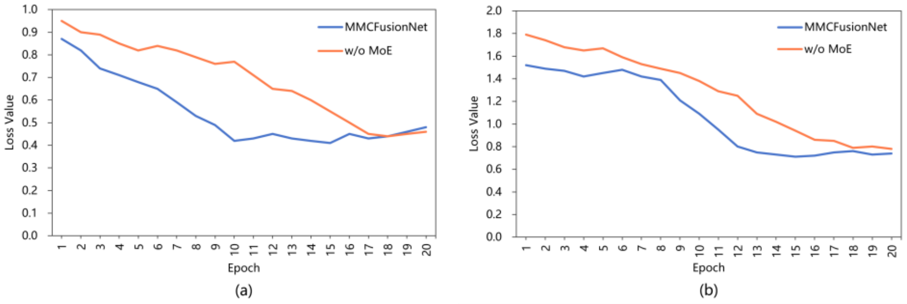


**Fig. 3.** Loss Trend Comparison: (a) results on HateMM, and (b) results on MUStARD.

**Fig. 3** illustrates the loss trajectories of both models on the two datasets. As shown, MMCFusionNet results in significantly lower loss values and a more rapid convergence rate on both datasets. This improvement is attributed to the dynamic weight allocation mechanism of the MoE module, which enhances early optimization efficiency by mitigating gradient conflicts between modalities. Our model's rapid convergence is a crucial advantage for practical applications, as it substantially reduces training time and the need for extensive computational resources. The cross-dataset experiments also indicate that it is effective for both hate and sarcasm recognition tasks.

## 4 Conclusion

This study proposes MMCFusionNet, a novel framework for detecting abnormal emotions, particularly hate and sarcasm. The model consists of three modules: a multimodal feature encoder module, an alignment module, and a fusion module. The encoders extract features from text, visual, audio, and facial modalities. The alignment module utilizes LSTM and MLP to reduce dimensionality and align features. The fusion module leverages MoE to strengthen intramodal representations by capturing concealed cues missed by conventional methods, while the dual-channel Co-Attention mechanism resolves intermodal conflicts by promoting complementary interactions.

Extensive experiments on the HateMM and MUStARD datasets have demonstrated MMCFusionNet's superior performance and reliability, even when some modalities are incomplete. Beyond technical merits, its MoE architecture offers practical advantages by accelerating training and reducing computational costs. This is a critical advantage for content moderation systems operating under dynamic update requirements and resource constraints.

**Acknowledgments.** This research was supported by the National Natural Science Foundation of China (under Grant Nos. 72171004, 72301010), the Project of Cultivation for Young Top-notch Talents of Beijing Municipal Institutions (under Grant No. BPHR202203061).

### References

1. Wei, Q., Zhou, Y., Xiang, S., et al.: MEAS: Multimodal Emotion Analysis System for Short Videos on Social Media Platforms. IEEE Transactions on Computational Social Systems 1–13 (2024)
2. Wang, H., Hee, M.S., Awal, R., Choo, K.T.W., Lee, R.K.W.: Evaluating GPT-3 Generated Explanations for Hateful Content Moderation. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI 2023), pp. 6255–6263. IJCAI Organization, Macao (2023)
3. Qi, A., Liu, Z., Zhou, X., et al.: Multimodal Emotion Recognition with Vision-Language Prompting and Modality Dropout. In: Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing, pp. 49–53. ACM, New York (2024)

4. Wang, S.: The Power of Images: How Multimodal Hate Speech Shapes Prejudice and Prosocial Behavioral Intentions. Social Media + Society 10(4), 1-12 (2024)
5. Jahan, M.S., Oussalah, M.: A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing. Neurocomputing 546, 126232 (2023).
6. Guo, W., Wang, J., Wang, S.: Deep Multimodal Representation Learning: A Survey. IEEE Access 7, 63373–63394 (2019)
7. Hu, Y., Hou, S., Yang, H., Huang, H., He, L.: A Joint Network Based on Interactive Attention for Speech Emotion Recognition. In: 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 1715–1720. IEEE, Brisbane (2023)
8. Busso, C., Bulut, M., Lee, C.C., et al.: IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. Language resources and evaluation 42, 335–359 (2008)
9. Xie, Z., Yang, Y., Wang, J., et al.: Trustworthy Multimodal Fusion for Sentiment Analysis in Ordinal Sentiment Space. IEEE Transactions on Circuits and Systems for Video Technology 34(8), 7657–7670 (2024)
10. Zadeh, A., Zellers, R., Pincus, E., et al.: MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. arXiv preprint arXiv:1606.06259 (2016)
11. Rana, A., Jha, S.: Emotion-Based Hate Speech Detection Using Multimodal Learning. arXiv preprint arXiv: 2202.06218 (2022)
12. Waligora, P., Aslam, M.H., Zeeshan, M.O., et al.: Joint Multimodal Transformer for Emotion Recognition in the Wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4625–4635. IEEE, Seattle (2024)
13. Das, M., Raj, R., Saha, P., et al.: HateMM: A Multi-Modal Dataset for Hate Video Classification. In: Proceedings of the International Conference on Web and Social Media (ICWSM), vol. 17, pp. 1014–1023. AAAI Press, Limassol (2023)
14. Castro, S., Hazarika, D., Pérez-Rosas, V., et al.: Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper). arXiv preprint arXiv:1906.01815 (2019)
15. Devlin, J., Chang, M.W., Lee, K., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019), pp. 4171–4186. ACL, Minneapolis (2019)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020)
17. Chen, H., Wu, S., Wu, X., Lin, Z.: Speech Emotion Recognition Using Multiple Classification Models Based on MFCC Feature Values. Applied and Computational Engineering. 6, 1037–1047 (2023)
18. Deng, J., Guo, J., Yang, J., et al.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4690–4699. IEEE, California (2019)
19. Radford, A., Kim, J.W., Xu, T., et al.: Robust Speech Recognition via Large-Scale Weak Supervision. In: Proceedings of the International Conference on Machine Learning (ICML 2023), pp. 28492–28518. PMLR, Virtual Conference (2023)
20. Khare, S.K., Blanes-Vidal, V., Nadimi, E.S., et al.: Emotion Recognition and Artificial Intelligence: A Systematic Review (2014–2023) and Research Recommendations. Information Fusion 102, 102019 (2024)

21. Fedus, W., Zoph, B., Shazeer, N.: Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Journal of Machine Learning Research. 23(120), 1–39 (2022)
22. Lu, J., Yang, J., Batra, D., et al.: Hierarchical Question-Image Co-Attention for Visual Question Answering. In: Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), vol. 29, pp. 1–12. Neural Information Processing Systems Foundation, Barcelona (2016)
23. Gao, P., Tao, C., Guan, D.: FEF-Net: Feature Enhanced Fusion Network with Crossmodal Attention for Multimodal Humor Prediction. Multimedia Systems 30(4), 195 (2024)
24. Ray, A., Mishra, S., Nunna, A., et al.: A Multimodal Corpus for Emotion Recognition in Sarcasm. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), pp. 6992–7003. ELRA, Marseille (2022)
25. Ma, F., Zhang, Y., Sun, X.: Multimodal Sentiment Analysis with Preferential Fusion and Distance-Aware Contrastive Learning. In: 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 1367–1372. IEEE, Brisbane (2023)
26. Lai, S., Hu, X., Xu, H., et al.: Multimodal Sentiment Analysis: A Survey. Displays 80, 102563 (2023)