Enriching Multimodal Sentiment Analysis Through Textual Emotional Descriptions of Visual-Audio Content

Sheng Wu[1,2], **Dongxiao He**[3], **Xiaobao Wang**[3,2,*], **Longbiao Wang**[3,*], **Jianwu Dang**[4]

[1]School of New Media and Communication, Tianjin University, Tianjin, China
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[3]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
[4]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
{2022245014, hedongxiao, wangxiaobao, longbiao_wang}@tju.edu.cn, jdang@jaist.ac.jp

Abstract

Multimodal Sentiment Analysis (MSA) stands as a critical research frontier, seeking to comprehensively unravel human emotions by amalgamating text, audio, and visual data. Yet, discerning subtle emotional nuances within audio and video expressions poses a formidable challenge, particularly when emotional polarities across various segments appear similar. In this paper, our objective is to spotlight emotion-relevant attributes of audio and visual modalities to facilitate multimodal fusion in the context of nuanced emotional shifts in visual-audio scenarios. To this end, we introduce DEVA, a progressive fusion framework founded on textual sentiment descriptions aimed at accentuating emotional features of visual-audio content. DEVA employs an Emotional Description Generator (EDG) to transmute raw audio and visual data into textualized sentiment descriptions, thereby amplifying their emotional characteristics. These descriptions are then integrated with the source data to yield richer, enhanced features. Furthermore, DEVA incorporates the Text-guided Progressive Fusion Module (TPF), leveraging varying levels of text as a core modality guide. This module progressively fuses visual-audio minor modalities to alleviate disparities between text and visual-audio modalities. Experimental results on widely used sentiment analysis benchmark datasets, including MOSI, MOSEI, and CH-SIMS, underscore significant enhancements compared to state-of-the-art models. Moreover, fine-grained emotion experiments corroborate the robust sensitivity of DEVA to subtle emotional variations.

Introduction

Sentiment analysis is a classic language understanding task, where traditionally, the analysis of user sentiment is conducted through text (Lei, Qian, and Zhao 2016). With the evolution of social media, there has been a significant increase in user-generated videos, making multimodal sentiment analysis (MSA) gradually emerge as a hotspot in research. Its objective is to comprehensively analyze people's emotional states through text, audio, and visual data (Poria et al. 2020; Dong et al. 2024). MSA plays a crucial role in various fields such as healthcare (Doctor et al. 2016; Jiang et al. 2020), social media analytics (Melville, Gryc, and Lawrence 2009), and human-computer interaction (Peter and Urban 2012). Compared to unimodal approaches, multimodal analysis offers greater robustness and comprehensiveness in understanding human emotions.

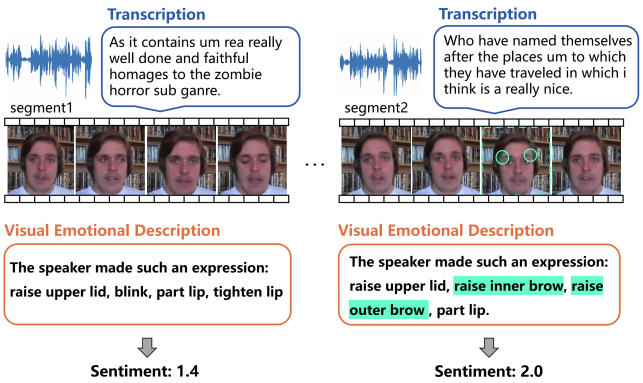

Figure 1: The illustration of our motivation is as follows: The text transcription of the audio is indicated within the blue box, while the visual emotional description is within the orange box. The teal highlighting indicates highly emotionally relevant descriptions, and teal circles are used to mark the corresponding microexpressions in the source data.

Leveraging the advancements in deep learning techniques (Chen et al. 2022; Huang et al. 2022; Fu et al. 2021; Wang et al. 2023), recent approaches in multimodal sentiment analysis (MSA) primarily concentrate on representation learning and fusion strategies across modalities. In terms of representation learning, various methods have emerged, including feature decoupling techniques aimed at mapping features into shared and private spaces (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022a,b). Moreover, contrastive learning (Yang et al. 2023) and multitask learning (Yu et al. 2021) have been explored to enhance representation learning. Regarding multimodal fusion, the initial approach is often early fusion, where features from text, audio, and visual modalities are concatenated for downstream tasks. Subsequently, more sophisticated techniques such as outer product (Zadeh et al. 2017), Convolutional

*Corresponding author.

Neural Networks (Huang et al. 2020), and Recurrent Neural Networks (Sun et al. 2020) have been adopted. Furthermore, attention mechanisms (Chen et al. 2017; Tsai et al. 2019) have been explored for multimodal data fusion in recent studies.

However, we find that discerning subtle differences in emotional intensity becomes challenging when the emotional polarities of different segments are closely aligned, particularly when analyzing raw data such as audio and visual inputs. As depicted in Figure 1, segments 1 and 2 demonstrate highly similar audio temporal distributions, and their overall facial expressions exhibit considerable resemblance. Relying solely on transcription, audio, and visuals poses difficulties in accurately determining the sentimental polarity of segment 2. To the best of our knowledge, previous studies have not explicitly tackled the scenario of fine-grained emotional changes in visual-audio content.

Fortunately, several studies suggest that different modalities contribute disparately to the Multimodal Sentiment Analysis (MSA) task, with text often serving as the core modality, audio and visual as auxiliary modalities (Zhang et al. 2023; Wu et al. 2021; Li et al. 2022). We find that articulating emotions in audio and depicting expressions using text can inherently accentuate disparities in sentimental polarity, leading to more precise sentimental assessments. This observation holds true across multiple datasets, as illustrated in Figure 1. For instance, when we describe facial expressions in emotional terms, the micro-expression of "raising eyebrows" translates into "raise inner brow, raise outer brow" in textual form, thereby suggesting that segment 2 elicits a more positive sentiment compared to segment 1.

Based on the above observations, we propose DEVA, a novel approach designed to accentuate emotional expressions in visual-audio content through textual descriptions. DEVA constructs these descriptions by narrating minor modalities with text and progressively integrating them with the source data under textual guidance. Initially, DEVA utilizes a pre-trained BERT model (Devlin et al. 2019) alongside separate feature extractors for audio and visual modalities to process text, audio, and visual inputs. Following this, three Transformers encode each modality into a unified format. Subsequently, Emotional Description Generator (EDG) leverages OpenFace (Baltrušaitis et al. 2018) and OpenSMILE (Eyben, Wöllmer, and Schuller 2010) tools to extract emotionally relevant visual and acoustic features from the video, generating natural language descriptions to highlight emotionally significant features, particularly subtle emotional shifts in audio and visual cues. These descriptions are then fused with the source data features to enhance modality features. Furthermore, we introduce the Text-Guided Progressive Fusion Module (TPF), utilizing text as the core modality to guide the fusion of audio and visual modalities into minor modality fusion features. Finally, the core and minor modality features are employed in cross-modal Transformers for fusion, effectively bridging distribution differences between the core and minor modalities.

The main contributions can be summarized as follows:

- We propose DEVA, a progressive fusion framework based on emotional description to highlight the emotional characteristics of visual-audio content. This method transforms audio and visual source data into textual emotional descriptions.
- We introduce a novel Emotional Description Generator (EDG), which textualizes minor modalities into emotional descriptions to highlight the emotional representation in audio and visual, addressing the issue of fine-grained variations in visual-audio emotion features. Meanwhile, we design a Text-guided Progressive Fusion method (TPF), which progressively fuses audio and visual data into minor modality features guided by text to bridge the gap between core and minor modalities.
- DEVA achieves state-of-the-art performance on three popular MSA datasets, thoroughly analyzing and validating the method's effectiveness and advancements.

Related Work

Multimodal Sentiment Analysis (MSA) is a widely studied research topic. Unlike previous approaches that solely utilized unimodal data (Aman and Szpakowicz 2007; Shirian and Guha 2020), MSA integrates text, audio, and visual non-verbal information to obtain more rich and robust features. Previous research has mainly focused on representation learning and multimodal fusion. For methods centered around representation learning, approaches like (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022a,b) decompose each modality into modality-invariant and -specific representations, utilizing squared norm loss as a constraint. In terms of methods focused on multimodal fusion, Zedeh (Zadeh et al. 2016b) first proposed a multimodal dictionary, learning dynamic interactions between facial gestures and spoken vocabulary to model sentiment. Later, the Tensor Fusion Network was introduced (Zadeh et al. 2017), which encodes each modality with its corresponding subnetwork and models interactions at the unimodal, bimodal, and trimodal levels through the triple Cartesian product. Approaches like (Zadeh et al. 2018; Yu et al. 2021; Yang et al. 2023) adopt late fusion at the decision level.

In recent years, Transformers (Vaswani et al. 2017) have dominated the field of deep learning. In the field of Multimodal Sentiment Analysis (MSA), Transformer is widely used for feature extraction, representation learning, and multimodal fusion. The Tsai et al. (Tsai et al. 2019) introduced a multimodal transformer to align multimodal sequences, but the pairwise fusion approach is less efficient. (Lv et al. 2021; Zhang et al. 2023) combine textual and non-textual features for multimodal interaction and fusion. Unlike the aforementioned work, we partition each modality into traditional data and text data based on emotional descriptions, which complement each other, thereby obtaining a unimodal representation with enhanced expressive power.

Method

Problem Definition

The original inputs of multimodal sentiment analysis includes text (t), audio (a), and visual (v). The goal of this task is to fuse the data from different modalities and output the pretictive sentimental polarity $\hat{y}$.

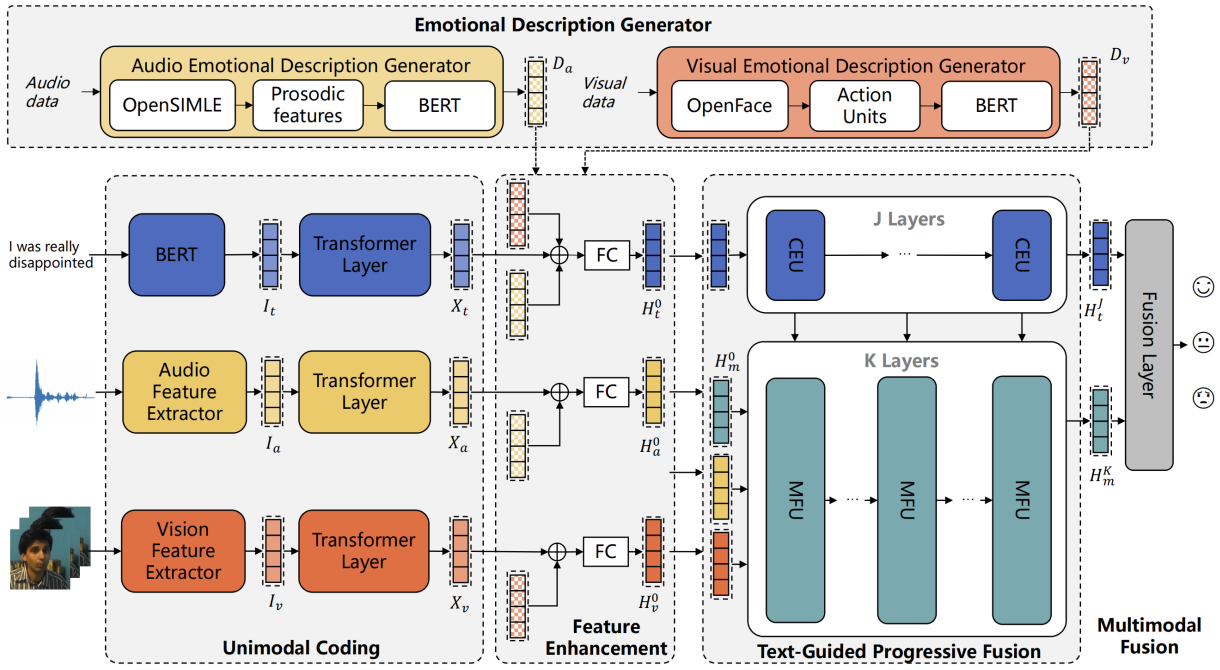

Figure 2: The overall architecture of DEVA consists of unimodal coding, an emotional description generator, feature enhancement, text-guided progressive fusion, and multimodal fusion.

Model Overview

The overall processing pipeline of the proposed Emotional Description Progressive framework (DEVA) is shown in Figure 2. We begin by preprocessing the source data into a unified format. Simultaneously, the Emotional Description Generator (EDG) is employed to extract emotion-relevant features from audio and visual modality source data, transforming them into textual Emotional Description. These textual features are then concatenated with the source data and input into a fully connected layer for fusion enhancement. Subsequently, the Text-Guided Progressive Fusion module (TPF) is applied, using the core modality (text) from different Transformer layers as guidance to progressively fuse the minor modalities (audio and visual) and generate minor modality features. Finally, a cross-modal Transformer is utilized to merge core and minor modalities, yielding the ultimate multimodal representation.

Unimodal Coding

For unimodal encoding, we adopt a two-stage approach. In the first stage, following prior work, we utilize BERT (Devlin et al. 2019) , Librosa (McFee et al. 2015), and OpenFace (Baltrušaitis et al. 2018) to individually encode text, audio, and visual source data, resulting in the representation $I_m \in \mathbb{R}^{T_m \times d_m}$, where $T_{m\in\{t,a,v\}}$ represents the length of each modal data, and $d_{m\in\{t,a,v\}}$ represents the dimension of each modal data.

In the second stage, we initialize a token E_m for each modality and concatenate it after I_m, inputting the combined representation into the Transformer Layer (Vaswani et al. 2017) to obtain traditional features for each modality:

$$X_m = Trans([I_m; E_m], \theta_{Trans}) \in \mathbb{R}^{T \times d}, \qquad (1)$$

where X_m is the unified feature of each modality $m \in \{t, a, v\}$ with a size of $T \times d$, E_{Trans} and θ_{Trans} respectively represent the Transformer feature extractor and corresponding parameters, $[\cdot;\cdot]$ represent the concatenation.

It is noteworthy that we do not use the output of the Transformer; instead, we select the first T tokens (where $T < T_m$) as the traditional feature information. This choice is made because most of the information in the Transformer tends to aggregate in the embeddings of the initial tokens. Aggregating unimodal information in the initial tokens helps to eliminate redundant information.

Emotional Description Generator

We use a third-party tool to extract emotionally strongly correlated features from audio and visual (facial expressions) and convert them into text descriptions. Specifically, we describe the audio modality as text that contains loudness, pitch, jitter, and shimmer, and the visual modality with text that contains multiple facial expression action units.

Audio Emotional Description. OpenSMILE (Eyben, Wöllmer, and Schuller 2010) is a feature extractor used for audio signal processing, commonly utilized in fields such as speech recognition and affective computing. Through OpenSMILE, we extract four prosodic features closely related to emotion: pitch, loudness, jitter, and shimmer. Among them, jitter represents the variability in pitch, while shimmer represents the variability in loudness. We extract the prosodic features of the entire dataset and calculate the

Prosodic features	Description
loudness	low loudness, normal loudness, high loudness
pitch	low pitch, normal pitch, high pitch
jitter	low jitter, normal jitter, high jitter
shimmer	low shimmer, normal shimmer, high shimmer

Table 1: Descriptions of Prosodic features.

numerical tertiles for each of the four features, dividing them into low, normal, and high levels based on the tertiles. This results in 4×3 AED Units, as shown in the Table 1. Ultimately, we obtain AEDs in the following format: *"The Speaker made such an tone: pitch, loudness, jitter, and shimmer at different levels."*

Visual Emotional Description. Since facial expression is the most important basis for visual judgment of sentiment, we utilize OpenFace (Baltrušaitis et al. 2018) to extract facial Action Units (AUs) as Visual Emotional Description. OpenFace provides 16 common AUs, each AU corresponding to a textual description. For instance, AU04 represents "lower brow," and AU20 represents "stretch lip." The specific AU descriptions are detailed in the Table 2 below. Additionally, we establish a criterion for selecting AUs for description, which involves the following steps: a) If an AU appears continuously for three frames, it is considered a candidate. b)The candidate AUs are sorted based on their duration, from highest to lowest, and the top-k AUs are selected for description.

The reason for selecting k AUs instead of all candidate AUs is to reduce the redundant impact of minor facial emotional features, allowing the model to capture the most significant emotional characteristics. Ultimately, we generate VEDs in the following format: *"The speaker made such an expression: k AUs."*

Finally, we encode the obtained AED and VED using the same encoding method as the text modality to acquire emotional description $D_a \in \mathbb{R}^{T\times d}$ and $D_v \in \mathbb{R}^{T\times d}$:

$$F_a^{'} = AEDG(I_a) \in \mathbb{R}^{T\times d}, \tag{2}$$

$$F_v^{'} = VEDG(I_v) \in \mathbb{R}^{T\times d}, \tag{3}$$

$$D_m = Trans([F_m^{'}; E_m]), \theta_{Trans}) \in \mathbb{R}^{T\times d}, \tag{4}$$

where D_m is the unified feature of audio and visual emotional description, $m \in \{a, v\}$, and $[\cdot;\cdot]$ represents the concatenation operation.

Feature Enhancement

We concatenate the traditional and description feature and merge them through fully connected layers, enhancing the traditional representation to include more emotional information. Specifically, we fuse X_t with D_a and D_v to obtain the text-enhanced modality H_t^0,

$$H_t^0 = FC([X_t; D_a; D_v]) \in \mathbb{R}^{T\times d}, \tag{5}$$

then merge X_a with D_a to obtain the audio-enhanced modality H_a^0,

$$H_a^0 = FC([X_a; D_a]) \in \mathbb{R}^{T\times d}, \tag{6}$$

AU	Description	AU	Description
AU01	raise inner brow	AU12	pull lip corner
AU02	raise outer brow	AU15	depress lip corner
AU04	lower brow	AU20	stretch lip
AU05	raise upper lid	AU23	tighten lip
AU06	raise cheek	AU25	part lip
AU07	tighten lid	AU26	drop jaw
AU09	wrinkle nose	AU28	suck lip
AU10	raise upper lip	AU45	blink

Table 2: Descriptions of Action Units.

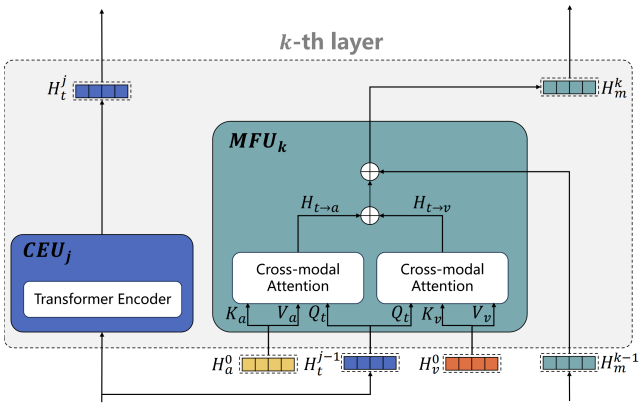

Figure 3: The architecture of TPF.

and we fuse X_v with D_v to obtain the visual-enhanced modality H_v^0:

$$H_v^0 = FC([X_v; D_v]) \in \mathbb{R}^{T\times d}, \tag{7}$$

where FC is fully connected network, $[\cdot;\cdot]$ represents the concatenation operation.

Text-Guided Progressive Fusion

After obtaining enhanced text, audio, and visual features, we apply Text-Guided Progressive Fusion (TPF) to merge the minor modality. TPF consists of K layers of Minor-modality Fusion Units (MFU) and J layers of Core-modality Enhancement Units (CEU), where K and J satisfy the following relationship: $K = J + 1$. The purpose of MFU is to capture text features from different layers of the Transformer as the core modality. TPF uses the output of each layer of MFU as a guide and progressively fuses audio and visual features layer by layer, merging them into minor modality features. Finally, the core modality and minor modality are further combined to obtain the ultimate fused features. For ease of discussion, we consider the k-th layer of MFU and the j-th layer of CEU as one layer of TPF, as illustrated in Figure 3.

Core-modality Enhancement Unit To obtain text features at different levels, we utilize the Transformer Encoder as the j-th layer of CEU to extract deeper-level text features:

$$H_t^j = Trans_j(h_t^{j-1}, \theta_{Trans_j}) \in \mathbb{R}^{T\times d}, \tag{8}$$

where $Trans_j$ and θ_{Trans_j} represent the j-th Transformer encoder and corresponding parameters.

Minor-modality Fusion Unit We initialize a minor modality feature H_m^0, then use text as a guide to fuse audio and visual features, as shown in Figure 3. The input to MFU includes H_a^0, H_v^0, the output Ht from the previous CEU layer, and the output H_m from the previous MFU layer. First, we guide the fusion with text for audio, using H_t as Q in Cross-modal Attention, and H_a as K and V, performing the fusion with the following formula:

$$\begin{aligned} H_{t\to a} &= CMA(H_t^{j-1}, H_a^0) \\ &= softmax(\frac{H_t^{j-1}W_{Q_t}W_{K_a}^T H_a^{0^T}}{\sqrt{d_k}})W_{V_a}^T H_a^{0^T}. \end{aligned} \tag{9}$$

Similarly, the fusion guided by text for visual features is expressed by the following formula:

$$\begin{aligned} H_{t\to v} &= CMA(H_t^{j-1}, H_v^0) \\ &= softmax(\frac{H_t^{j-1}W_{Q_t}W_{K_v}^T H_v^{0^T}}{\sqrt{d_k}})W_{V_v}^T H_v^{0^T}. \end{aligned} \tag{10}$$

Then we perform a weighted sum of the two fusion vectors, and the minor modality feature is updated by adding it to the sum, serving as the output of this layer of MFU.

$$H_m^k = H_m^{j-1} + \alpha H_{t\to a} + \beta H_{t\to v} \in \mathbb{R}^{T\times d}, \tag{11}$$

where H_m^k represnets the output minor-modality features of k-th MFU, α and β are learnable parameters.

Ultimate Multimodal Fusion Finally, the core modality feature H_t^J and the minor modality feature H_m^K are used as the input for the Crossmodal Transformer, resulting in the fused modality vector. Specifically, H_t^J is used as Q, and H_m^K is used as K and V:

$$\begin{aligned} H &= CMT(H_t^J, H_m^K) \\ &= softmax(\frac{H_t^J W_{Q_t}W_{K_m}^T H_m^{K^T}}{\sqrt{d_k}})W_{V_m}^T H_m^{K^T} \in \mathbb{R}^{T\times d}. \end{aligned} \tag{12}$$

Learning Objectives

Finally, we add a classifier after the Crossmodal Transformer to obtain the final prediction results $\hat{y}$. For classification tasks, we use the standard cross-entropy loss whereas for regression tasks, we use the mean squared error (MSE) loss as the MSA basic optimization objective, which is:

$$\begin{aligned} \mathcal{L} &= -\frac{1}{N_b}\sum_{i=0}^{N_b} y_i \cdot log\hat{y}_i \\ &= \frac{1}{N_b}\sum_{i=0}^{N_b} |\hat{y}_i - y_i|, \end{aligned} \tag{13}$$

where N_b is the number of training samples, $\hat{y}$ is the presiction of DEVA, and y is the ground truth.

Experiments

Datasets

We conduct extensive experiments on three standard multimodal sentiment analysis benchmarks: MOSI (Zadeh et al. 2016a), MOSEI (Bagher Zadeh et al. 2018), and CH-SIMS (Yu et al. 2020).

MOSI. The MOSI dataset is a popular benchmark dataset in MSA research. This dataset is a collection of YouTube monologues. MOSI contains 2199 subjective words-video clips. These utterances are artificially labeled as consecutive opinion scores between -3 to 3, where -3/+3 represents strong negative/positive sentiment.

MOSEI. The MOSEI dataset is an improvement on MOSI. It contains 23,454 YouTube monologues video segments covering 250 distinct topics from 1,000 distinct speakers. Each utterance also has sentiment consecutive opinion scores between -3 to 3.

CH-SIMS. The CH-SIMS dataset is a Chinese MSA dataset with fine-grained annotations of modality. The dataset comprises 2,281 video clips collected from various sources, such as different movies and TV serials with spontaneous expressions, various head poses, etc. Human annotators label each sample with a sentiment score from -1 (strongly negative) to 1 (strongly positive).

Evaluation Metrics

Following the previous works (Yu et al. 2020, 2021; Zhang et al. 2023), we present our experimental findings in two distinct formats: classification and regression. In terms of classification, we provide metrics such as Weighted F1 score (F1-Score) and 2-class accuracy (Acc-2). Specifically, for the MOSI and MOSEI datasets, we compute Acc-2 and F1-Score in two configurations: negative / non-negative (including zero) and negative / positive (excluding zero). Moreover, we include additional metrics such as 5-class accuracy (Acc-5) and 7-class accuracy (Acc-7). For the CH-SIMS dataset, we calculate Acc-2, F1, 3-class accuracy (Acc-3), and Acc-5. Regarding regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). In all metrics except MAE, higher values indicate better performance.

Baselines

In order to verify the superiority of our proposed DEVA, we conduct an experimental comparison with the following state-of-the-art baseline models:

- Utterance-vector fusion approaches that use tensor-based fusion and low-rank variants: **TFN** (Zadeh et al. 2017), **LMF** (Liu et al. 2018).
- Models which Learn invariant and specific representations through feature decomposition: **MISA** (Hazarika, Zimmermann, and Poria 2020), **FDMER** (Yang et al. 2022a), **MFSA** (Yang et al. 2022b), **ConFEDE** (Yang et al. 2023).
- Models which utilize attention and transformer modules to improve token representations using non-verbal signals: **MulT** (Tsai et al. 2019), **PMR** (Lv et al. 2021), **ALMT** (Zhang et al. 2023).

Methods	MOSI						MOSEI					
	Acc-2	Acc-5	Acc-7	F1	MAE	Corr	Acc-2	Acc-5	Acc-7	F1	MAE	Corr
TFN*	77.99/79.08	-	34.46	77.95/79.11	0.947	0.673	78.50/81.89	-	51.6	78.96/81.74	0.572	0.714
LMF*	77.90/79.18	-	33.82	77.80/79.15	0.950	0.651	80.54/83.48	-	51.59	80.94/83.36	0.575	0.716
MulT*	79.71/80.98	42.68	36.91	79.63/80.95/	0.879	0.702	81.15/84.63	54.18	52.84	81.56/84.52	0.559	0.733
ICCN	-/83.07	-	39.01	-/83.02/	0.862	0.714	-/84.18	-	51.58	-/84.15	0.565	0.713
MISA*	81.84/83.54	47.08	41.37	81.82/83.58	0.776	0.778	80.67/84.67	53.63	52.05	81.12/84.66	0.557	0.751
MAG-BERT	82.37/84.43	-	43.62	82.50/84.61	0.727	0.781	82.51/84.82	-	52.67	82.77/84.71	0.543	0.755
PMR	-/82.40	-	40.60	-/82.10	-	-	-/83.10	-	51.80	-/82.80	-	-
MFSA	-/83.3	-	41.1	-/83.7	0.856	0.722	-/83.8	-	53.2	-/83.6	0.574	0.724
FDMER	-/84.6	-	44.1	-/84.7	0.724	0.788	-/86.1	-	54.1	-/85.8	0.536	0.773
Self-MM†	82.54/84.45	**52.22**	45.56	82.46/84.44	**0.719**	**0.794**	82.09/84.76	53.54	53.65	82.43/84.67	0.535	0.761
ALMT†	83.24/85.37	50.29	44.75	83.41/85.46	0.738	0.776	82.34/85.94	55.05	53.32	81.85/85.93	0.534	0.771
ConFEDE	84.17/85.52	-	42.27	84.13/85.52	0.742	0.784	81.65/85.82	-	**54.86**	82.17/85.83	**0.522**	**0.780**
DEVA	**84.40/86.29**	51.78	**46.32**	**84.48/86.30**	0.730	0.787	**83.26/86.13**	**55.32**	52.26	**82.93/86.21**	0.541	0.769

Table 3: Comparison on MOSI and MOSEI Datasets. * represents results obtained from (Mao et al. 2022) and its corresponding GitHub page[1]. Models with † are reproduced under the same conditions. Best results are marked in bold.

Methods	Acc-2	Acc-3	Acc-5	F1	MAE	Corr
TFN*	78.38	65.12	39.30	78.62	0.432	**0.591**
LMF*	77.77	64.68	40.53	77.88	0.441	0.575
MulT*	78.56	64.77	37.94	79.66	0.453	0.564
MISA*	76.54	-	-	76.59	0.447	0.563
MAG-BERT	74.44	-	-	71.75	0.492	0.399
Self-MM†	77.64	64.68	41.76	77.85	0.428	0.590
ALMT†	78.59	64.98	40.70	78.94	0.450	0.535
DEVA	**79.64**	**65.42**	**43.07**	**80.32**	**0.424**	0.583

Table 4: Comparison on CH-SIMS. *represents the result is from (Mao et al. 2022) and its corresponding GitHub page [1] . Models with † are reproduced under the same conditions.

- Learning the multimodal and unimodal representations based on the multimodal label and generated unimodal labels: **Self-MM** (Yu et al. 2021).
- Learning textbased audio and text-based video features by optimizing canonical loss: **ICCN** (Sun et al. 2019).
- Model which allows audio and video information to leak into the BERT model for multimodal fusion: **MAG-BERT** (Rahman et al. 2020).

Comparison of Results

Tables 3 and 4 summarize the comparative results of our proposed method and all baseline models on the MOSI, MOSEI, and CH-SIMS datasets.

As shown in Table 3, our proposed DEVA outperforms all baseline models in Acc-2 and F1 (non-negative, negative). On the MOSI dataset, our model exhibits a 0.65% improvement in Acc-7 over the second-best result. Similarly, on the MOSEI dataset, our model shows a 0.27% improvement in Acc-5 over the second-best result. In other metrics, our model also approaches state-of-the-art results.

Scenarios in CH-SIMS are more complex than those in MOSI and MOSEI, presenting a greater challenge for multimodal sentiment recognition tasks. However, on the CH-SIMS dataset, our DEVA achieves state-of-the-art performance across all metrics except the Corr indicator. Notably, in binary classification tasks, our method gains a 1.05% improvement over the best baseline, and it surpasses the best baseline by 1.34% in the multi-classification metric Acc-5 and outperforming the highest baseline by 0.66% in F1, indicating the outstanding performance of our model on the challenging CH-SIMS dataset, which features a more complex environment.

Several baseline models also treat text as the core modality, but their results are inferior to those of our model. This demonstrates the effectiveness and advancements of the DEVA model in capturing emotional description.

Ablation Study and Analysis

Impacts of Different Modality Combinations In order to investigate the contributions of different modalities, particularly the Audio Emotional Description and Visual Emotional Description, which are carried by the text modality, to the overall performance of the model, we conduct various combinations of modalities on the MOSI and CH-SIMS datasets. We represent Text, Audio, visual, Audio Emotional Description, and Visual Emotional Description as T, A, V, AED, and VED, respectively. The results of the ablation experiments are presented in Table 5.

Initially, ablations are performed on Emotional Description (ED), including AED and VED. It can be observed that, compared to the complete model, removing either AED or VED leads to varying degrees of performance degradation across all metrics on the MOSI and CH-SIMS datasets, indicating a positive role of ED in the overall performance of our model. Furthermore, the addition of VED shows a more noticeable improvement in average performance compared

[1]https://github.com/thuiar/MMSA/blob/master/results/result-stat.md

Methods	Modality					MOSI					CH-SIMS				
	A	V	T	AED	VED	Acc-2	Acc-7	F1	MAE	Corr	Acc-2	Acc-3	F1	MAE	Corr
w/o ED	✓	✓	✓			82.36/84.15	43.00	82.46/84.19	0.744	0.784	77.89	63.23	78.69	0.442	0.545
	✓	✓	✓	✓		82.36/84.45	40.96	82.45/84.48	0.749	0.781	78.11	66.30	78.03	0.425	0.577
	✓	✓	✓		✓	83.38/85.21	45.77	83.43/85.20	0.738	0.777	78.33	65.42	78.19	0.424	0.583
w/o AV			✓	✓	✓	82.07/84.15	43.29	82.21/84.21	0.751	0.773	**79.86**	64.98	79.79	**0.416**	**0.604**
DEVA	✓	✓	✓	✓	✓	**84.40/86.28**	**46.21**	**84.48/86.31**	**0.737**	**0.786**	79.64	**65.42**	**80.32**	0.424	0.583

Table 5: The modality ablation studies on MOSI and CH-SIMS. AED means Audio Emotional Desciption and VED means visual Emotional Description.

Methods	MOSI			CH-SIMS		
	Acc-5	MAE	Corr	Acc-2	Acc-3	Corr
w/o CEU	48.98	0.751	0.780	76.36	62.36	0.551
w/o MFU	23.29	1.133	0.458	76.14	58.42	0.455
w/o EDG	48.98	0.744	0.784	77.89	63.23	0.545
w/o Fusion Layer	49.71	0.782	0.770	77.46	63.01	0.552
DEVA	**50.44**	**0.737**	**0.786**	**79.64**	**65.42**	**0.583**

Table 6: The ablation results by subtracting each component individually on the MOSI and CH-SIMS.

to the addition of AED. We infer that the emotional information contained in descriptions of facial expressions and actions might be richer than the emotional information conveyed by broad descriptions of audio signal characteristics such as pitch and volume.

Interestingly, we attempt to replace the original audio and visual source data with modal emotional descriptions for audio and visual modalities, ensuring that the entire model includes only the text modality. The results indicate that, for the MOSI dataset, the model maintains results comparable to traditional MSA inputs. On the CH-SIMS dataset, DEVA (w/o AV) demonstrates competitiveness with our complete model, which simultaneously includes traditional inputs and emotional descriptions. This suggests a novel direction for future MSA research.

Impacts of Different Components To validate the effectiveness of each component in DEVA, we provide the ablation results in Table 6. We observe a decrease in all metrics when CEU and MFU are removed (replaced with a single multi-layer Transformer Encoder and addition), with MFU having the greatest impact. This demonstrates the significance of the guidance from the text core modality and the progressive interaction between the core modality and the minor modality. Additionally, when removing EDG, i.e., discarding additional emotional description, we find that there is some degradation in model performance, supporting the idea that emotional domain information can enhance the emotion information capture capability in MSA. Finally, when removing the Fusion Layer from the last layer (replaced with a simple concatenation), there is a noticeable decrease in overall scores, highlighting the importance of crossmodal Transformer for effective fusion across different modalities.

Methods	Polarity	Acc-2	Acc-4	Acc-5	MAE	Corr
DEVA	[-3,-2)	**56.25**	**30.20**	**8.33**	**1.006**	0.065
	[-2, -1)	**55.68**	27.54	**20.76**	**0.677**	**0.128**
	[-1, 0)	**65.51**	**39.65**	**33.62**	0.652	**0.196**
	(0, +1]	**59.63**	**37.61**	**25.85**	0.794	**0.140**
	(+1, +2]	**56.66**	**33.33**	15.00	**0.674**	0.215
	(+2, +3]	**66.66**	**43.75**	**10.41**	**0.760**	**0.316**
ALMT	[-3,-2)	53.75	28.12	4.16	1.099	**0.102**
	[-2, -1)	54.49	**28.74**	20.35	0.687	0.047
	[-1, 0)	64.65	38.79	27.58	**0.584**	0.180
	(0, +1]	58.71	34.86	24.77	**0.766**	0.130
	(+1, +2]	57.50	29.16	**17.50**	0.693	**0.250**
	(+2, +3]	62.50	35.41	6.25	1.033	0.312

Table 7: Fine-grained study on MOSI. The green number indicates the predominance of DEVA, and the red number indicates the predominance of ALMT.

Fine-grained Study

To explore the performance of DEVA within finer ranges of sentimental polarity, we conduct a fine-grained MSA comparison experiment on the MOSI dataset between DEVA and ALMT. We subdivide the sentimental polarity of MOSI into seven sub-intervals, each with a range of 1. Then, we train models on the entire MOSI dataset and perform regression and classification tasks within each sub-interval separately. The experimental results are shown in Table 7. It can be observed that our proposed model performs better than the baseline under the conditions of fine-grained sentimental polarity, especially in binary, ternary, and quinary classification metrics. This demonstrates that our model is more capable of distinguishing subtle changes in sentiment.

Conclusion

This paper introduces DEVA, a novel method for MSA, which utilizes generated emotional description for progressive fusion. DEVA seamlessly integrates the traditional features obtained from pre-trained models with emotional description features, treating text as the core modality and progressively fusing it with audio and visual modalities. This approach bridges the gap between different modalities. Rigorous experiments on several popular MSA datasets demonstrate the superiority of our proposed method. In future work, we aim to enhance the fluency of emotional descriptions to make them more closely daily-life expressions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62302333, 62422210, 62276187) and the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (No. GML-KF-24-16).

References

Aman, S.; and Szpakowicz, S. 2007. Identifying expressions of emotion in text. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, TSD'07, 196–205. Springer-Verlag. ISBN 3540746277.

Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.

Baltrušaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66.

Chen, M.; Wang, S.; Liang, P. P.; Baltrušaitis, T.; Zadeh, A.; and Morency, L.-P. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, 163–171. ISBN 9781450355438.

Chen, Z.; Li, B.; Xu, J.; Wu, S.; Ding, S.; and Zhang, W. 2022. Towards Practical Certifiable Patch Defense with Vision Transformer. *In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15127–15137.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Doctor, F.; Karyotis, C.; Iqbal, R.; and James, A. E. 2016. An intelligent framework for emotion aware e-healthcare support systems. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8.

Dong, Y.; He, D.; Wang, X.; Jin, Y.; Ge, M.; Yang, C.; and Jin, D. 2024. Unveiling Implicit Deceptive Patterns in Multi-Modal Fake News via Neuro-Symbolic Reasoning. In *AAAI Conference on Artificial Intelligence*.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, 1459–1462. Association for Computing Machinery. ISBN 9781605589336.

Fu, Y.; Okada, S.; Wang, L.; Guo, L.; Song, Y.; Liu, J.; and Dang, J. 2021. CONSK-GCN: Conversational Semantic- and Knowledge-Oriented Graph Convolutional Network for Multimodal Emotion Recognition. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *the Proceedings of the 28th ACM International Conference on Multimedia*, 1122–1131.

Huang, H.; Hu, Z.; Wang, W.; and Wu, M. 2020. Multimodal Emotion Recognition Based on Ensemble Convolutional Neural Network. *IEEE Access*, 8: 3265–3271.

Huang, H.; Wang, Y.; Chen, Z.; Li, Y.; Tang, Z.; Chu, W.; Chen, J.; Lin, W.; and Ma, K.-K. 2022. CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 989–997.

Jiang, Y.; Li, W.; Hossain, M. S.; Chen, M.; Alelaiwi, A.; and Al-hammadi, M. 2020. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Inf. Fusion*, 53: 209–221.

Lei, X.; Qian, X.; and Zhao, G. 2016. Rating Prediction Based on Social Sentiment From Textual Reviews. *IEEE Transactions on Multimedia*, 18: 1910–1921.

Li, Z.; Zhou, Y.; Zhang, W.; Liu, Y.; Yang, C.; Lian, Z.; and Hu, S. 2022. AMOA: Global Acoustic Feature Enhanced Modal-Order-Aware Network for Multimodal Sentiment Analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, 7136–7146.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Bagher Zadeh, A.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256.

Lv, F.; Chen, X.; Huang, Y.; Duan, L.; and Lin, G. 2021. Progressive Modality Reinforcement for Human Multimodal Emotion Recognition from Unaligned Multimodal Sequences. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2554–2562.

Mao, H.; Yuan, Z.; Xu, H.; Yu, W.; Liu, Y.; and Gao, K. 2022. M-SENA: An Integrated Platform for Multimodal Sentiment Analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 204–213.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P. W.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference 2015 (SciPy 2015)*, 18–24.

Melville, P.; Gryc, W.; and Lawrence, R. D. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. KDD '09, 1275–1284. ISBN 9781605584959.

Peter, C.; and Urban, B. 2012. Emotion in Human-Computer Interaction. In *Expanding the Frontiers of Visual Analytics and Visualization*, 239–262.

Poria, S.; Hazarika, D.; Majumder, N.; and Mihalcea, R. 2020. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*, 14: 108–132.

Rahman, W.; Hasan, M. K.; Lee, S.; Bagher Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2359–2369. Association for Computational Linguistics.

Shirian, A.; and Guha, T. 2020. Compact Graph Architecture for Speech Emotion Recognition. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6284–6288.

Sun, L.; Lian, Z.; Tao, J.; Liu, B.; and Niu, M. 2020. Multimodal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*, 27–34.

Sun, Z.; Sarma, P. K.; Sethares, W. A.; and Liang, Y. 2019. Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis. In *AAAI Conference on Artificial Intelligence*, 8992–8999.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Curran Associates Inc. ISBN 9781510860964.

Wang, X.; Dong, Y.; Jin, D.; Li, Y.; Wang, L.; and Dang, J. 2023. Augmenting Affective Dependency Graph via Iterative Incongruity Graph Learning for Sarcasm Detection. In *AAAI Conference on Artificial Intelligence*.

Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; and Zhu, L.-N. 2021. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4730–4738.

Yang, D.; Huang, S.; Kuang, H.; Du, Y.; and Zhang, L. 2022a. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1642–1651.

Yang, D.; Kuang, H.; Huang, S.; and Zhang, L. 2022b. Learning Modality-Specific and -Agnostic Representations for Asynchronous Multimodal Language Sequences. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 1708–1717.

Yang, J.; Yu, Y.; Niu, D.; Guo, W.; and Xu, Y. 2023. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7617–7630.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3718–3727.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *AAAI Conference on Artificial Intelligence*, 10790–10797.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5634–5641.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016a. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *IEEE Intelligent Systems*, abs/1606.06259: 82–88.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016b. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31: 82–88.

Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; and Yu, T. 2023. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 756–767.