

# Joint Modal Circular Complementary Attention for Multimodal Aspect-Based Sentiment Analysis

1<sup>st</sup> Hao Liu  
*School of Information and  
 Communication Engineering  
 Xi'an Jiaotong University  
 Shaanxi, P.R. China  
 haoliu88@stu.xjtu.edu.cn*

2<sup>nd</sup> Lijun He\*  
*School of Information and  
 Communication Engineering  
 Xi'an Jiaotong University  
 Shaanxi, P.R. China  
 lijunhe@mail.xjtu.edu.cn*

3<sup>rd</sup> Jiaxi Liang  
*School of Information and  
 Communication Engineering  
 Xi'an Jiaotong University  
 Shaanxi, P.R. China  
 liangjiaxi2002@163.com*

**Abstract**—Existing approaches to Multimodal Aspect-Based Sentiment Analysis have drawbacks: (i) Aspect extraction and sentiment classification always exhibit loose connections, overlooking aspect correlations which leads to inaccurate analysis of indirectly described aspects. (ii) Image pixels are coarsely treated equally in most methods, introducing visual noise that compromise sentiment analysis accuracy. (iii) Additionally, most rely on extra pre-training image-text relation detection networks, limiting their generality. To address these issues, we propose the Joint modal Circular Complementary attention framework (JCC) which optimizes aspect extraction and sentiment classification jointly by incorporating global text to enhance the model’s awareness of aspect correlations. JCC utilizes text for visual highlighting to mitigate the impact of visual noise. Furthermore, we design the Circular Attention module (CIRA) for general feature-focused aspect extraction and the Modal Complementary Attention module (MCA) for detailed information-focused sentiment classification. Experimental results across three MABSA subtasks demonstrate the superiority of JCC over existing methods.

**Index Terms**—Multimodal, Sentiment Analysis, Aspect Correlation, Visual Noise, Attention

## I. INTRODUCTION

Multimodal Aspect-Based Sentiment Analysis (MABSA) is a highly fine-grained task in sentiment analysis that has garnered significant attention recently [1]–[4]. Previous studies primarily focused on individual subtasks, including Multimodal Aspect Term Extraction (MATE) [5]–[8] and Multimodal Aspect Sentiment Classification (MASC) [9]–[14]. MATE predicts aspects presented in the text based on text-image pairs, while MASC predicts the sentiment polarity corresponding to known aspects. However, the accuracy of aspect term extraction in MATE directly impacts the results of aspect sentiment classification in MASC, thus integrating the two subtasks is imperative. To address this, Ju et al. [1] introduced the Joint Multimodal Aspect Sentiment Analysis (JMASA) task, aiming to predict aspects and their corresponding sentiment polarity simultaneously.

In Fig. 1, JMASA aims to identify aspect-sentiment pairs like (Taylor Swift, Positive) from text-image pairs. In instances where direct descriptions are unavailable for certain aspects (e.g., Taylor Swift), it becomes necessary to integrate sentiment semantics from related aspects (e.g., voice, lyrics) for

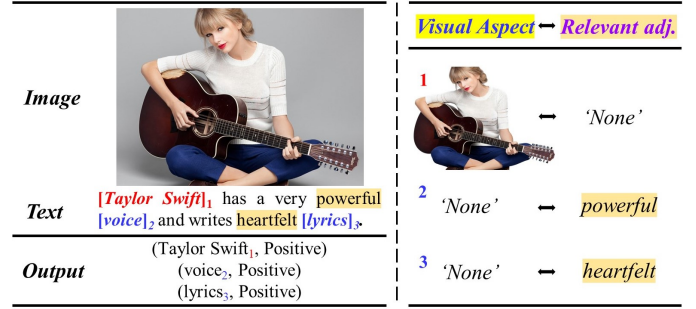


Fig. 1. The left side presents an example of JMASA task. The right side illustrates the corresponding scenarios with missing visual aspects and indirect descriptions. ‘None’ indicates non-existence.

discrimination. Moreover, while images offer more detailed features compared to text, some aspects may be absent. As depicted in Fig. 1, visual information effectively aids sentiment analysis for “Taylor Swift” but falls short in extracting details about “voice” and “lyrics”. These challenges contribute to the difficulty of MABSA.

As mentioned above, we emphasized the importance of integrating sentiment semantics from pertinent aspects for sentiment analysis on indirectly described aspects. However, for textual information in MASC, Ju et al. [1] solely considered aspectual representations, limiting the model’s awareness of aspect interrelations and making it challenging to analyze aspects without direct descriptions. During multimodal fusion, Ju et al., Ling et al. [2] and Yang et al. [3] directly correlated entire visual representations with textual content. Yet, coarse entire image visualization intuitively introduces aspect-independent visual noise, potentially hindering sentiment analysis. Furthermore, the incongruity in quantity of semantic information and detail levels between images and text poses a challenge as previously noted. In the MABSA subtask, the asymmetry in the amount of image and text information makes modal fusion susceptible to modal overlay. Additionally, MASC demands more detailed information compared to MATE. Addressing this challenge, Ju et al. and Zhou et al. [4] introduced a module for image-text relation detection to determine the appropriate level of image information for a

\*Lijun He is the corresponding author.

given task and fused the features directly. However, establishing such image-text relations in the real world is both difficult and resource-intensive and direct modal fusion may lead to information loss.

In this paper, we propose the **Joint modal Circular Complementary attention framework (JCC)** to handle the aforementioned challenges. The main contributions are summarized as follows:

- 1) *A joint multimodal aspect-based sentiment analysis framework introducing global text in MASC*: The framework concatenates MATE and MASC tasks to align with real-world needs. Unlike other approaches that overlook the correlations between aspects, JCC tackles the challenge of indirect aspect sentiment analysis by incorporating global text (GText) in the MASC task, enabling the model to learn aspect correlations.
- 2) *Reducing the impact of visual noise by utilizing text-highlighted visual features*: JCC utilizes textual representations to highlight visual features, reducing visual noise from coarse entire images. This allows the model to concentrate on aspect-dependent image features.
- 3) *CIRA and MCA modules for modal fusion focusing on different levels of detail features*: Due to the difficulty in obtaining image-text relations in reality, JCC introduces a CIRA module for the MATE task and a MCA module for the MASC task to focus on different levels of detail features. These modules circumvent the strong constraint of image-text relation detection and enhance the generalizability of method.

## II. METHODOLOGY

### A. Task Definition

Formally, we define three subtasks within MABSA: JMASA, MATE, and MASC. For a dataset comprising multimodal samples  $\mathcal{D} = \{(I_i, S_i)\}_{i=1}^N$ , each sample containing an image  $I$  and a sentence  $S$  consisting of  $n$  words  $\{W_0, \dots, W_n\}$ , we annotated each word  $W_j$  using the BIO sequence tagging method with entity labels. These labels indicate whether the word is an aspect term and specify its sentiment polarity  $p \in \{\text{positive, neutral, negative}\}$ . For these three subtasks, we formulate their outputs as follows:

- JMASA: Output =  $[(a_1^b, a_1^e, p_1), \dots, (a_i^b, a_i^e, p_i), \dots]$ ,
- MATE: Output =  $[(a_1^b, a_1^e), \dots, (a_i^b, a_i^e), \dots]$ ,
- MASC: Output =  $[(\underline{a_1^b}, \underline{a_1^e}, p_1), \dots, (\underline{a_i^b}, \underline{a_i^e}, p_i), \dots]$ ,

where  $a_i^b$ ,  $a_i^e$  and  $p_i$  inform the start index, end index, and sentiment polarity of an aspect term in the sentence. In other words, we want the model to be able to predict the corresponding aspects, sentiment polarity or both simultaneously based on text and its associated image. The underlined token denotes that it was given during inference.

### B. Feature Extractor

**Text Representation.** We tokenize the input text and convert it into textual representation using BERT<sup>1</sup>. Let

$Tok = \{tok_1, \dots, tok_N\}$  represent the input tokens, and  $T = \{t_1, \dots, t_N\}$  denote the resulting textual representation.

**Visual Representation.** We employ the pre-trained ResNet152<sup>2</sup> to extract image features. The resulting output feature tensor is represented as  $\bar{V} \in \mathbb{R}^{2048 \times 49}$ , where 49 corresponds to image region and 2048 is the feature dimension. To ensure consistency in dimension between image and text representations, we project the image features to match the text dimension, denoted as  $V \in \mathbb{R}^{d \times 49}$ .

### C. Circular Attention for MATE

The top-left part of Fig. 2 illustrates the MATE architecture. Initially, the cross-modal attention network (CA) employs text as a guide to generate highlighted visual features (HLV) denoted as  $H \in \mathbb{R}^{d \times 49}$ , thereby mitigating the influence of irrelevant visual noise on the model. The CA is shown on the left side of the Fig. 2.

We devise the CIRA for closed-loop inter-attentional fusion among three unimodal features:  $T$ ,  $V$ , and  $H$ . The CIRA module is multi-stage and treats each modality feature equally to align the requirements of the MATE task. The fusion order is altered across stages to facilitate the integration of information from diverse perspectives. The fusion algorithm has the flexibility to modify the type and number of fused modalities, showcasing excellent scalability adaptable to various tasks.

As depicted in Fig. 2, the CIRA module is divided into two stages: the multimodal inter-attention fusion stage and the reverse order mutual attention stage. The entire module is cascaded with CAs, and each stage consists of three CAs:  $CA_H()$ ,  $CA_V()$ , and  $CA_T()$ , corresponding to the CAs guided by  $H$ ,  $V$ , and  $T$ . The CA achieves the fusion of two modal information, for example, the vision-text cross-modal attention fusion network  $CA_T()$  takes textual features and the output features from the previous network as input and produces dominant textual fusion features. Taking the first  $CA_H()$  as an example, its  $H$  as query and  $T$  as key and value, the output is

$$CA_H(H, T, T) = \text{softmax}\left(\frac{[W_Q H][W_K T]^T}{\sqrt{d}}\right)[W_V T], \quad (1)$$

where  $W_Q$ ,  $W_K$  and  $W_V \in \mathbb{R}^{d \times d_{head}}$ . In the first stage, the fusion order is arranged as  $CA_H()$ ,  $CA_V()$ ,  $CA_T()$ , while in the subsequent stage, it is reversed. The fusion order differs at each stage and the fusion between any two modalities is constrained by another modality. Subsequently, by concatenating the input feature  $T$  with  $\mathcal{F}_c$  which is the output of CIRA, the outcome is processed through the feedforward layer, producing  $\mathcal{F}'$  as the input for the aspect extraction algorithm.  $\mathcal{F}' = FFN(\mathcal{F}_c \oplus T)$ , where  $\oplus$  denotes concatenation in dimensions. Subsequently,  $\mathcal{F}'$  is mapped through a linear layer to form an unstandardized vector  $f$  representing the aspect's start-end positions, which is then used to generate the corresponding probability distribution  $p$ :

$$f^s = \text{Linear}^s(\mathcal{F}'), f^e = \text{Linear}^e(\mathcal{F}'), \quad (2)$$

<sup>1</sup><https://huggingface.co/bert-base-uncased>

<sup>2</sup><https://download.pytorch.org/models/resnet152-b121ed2d.pth>

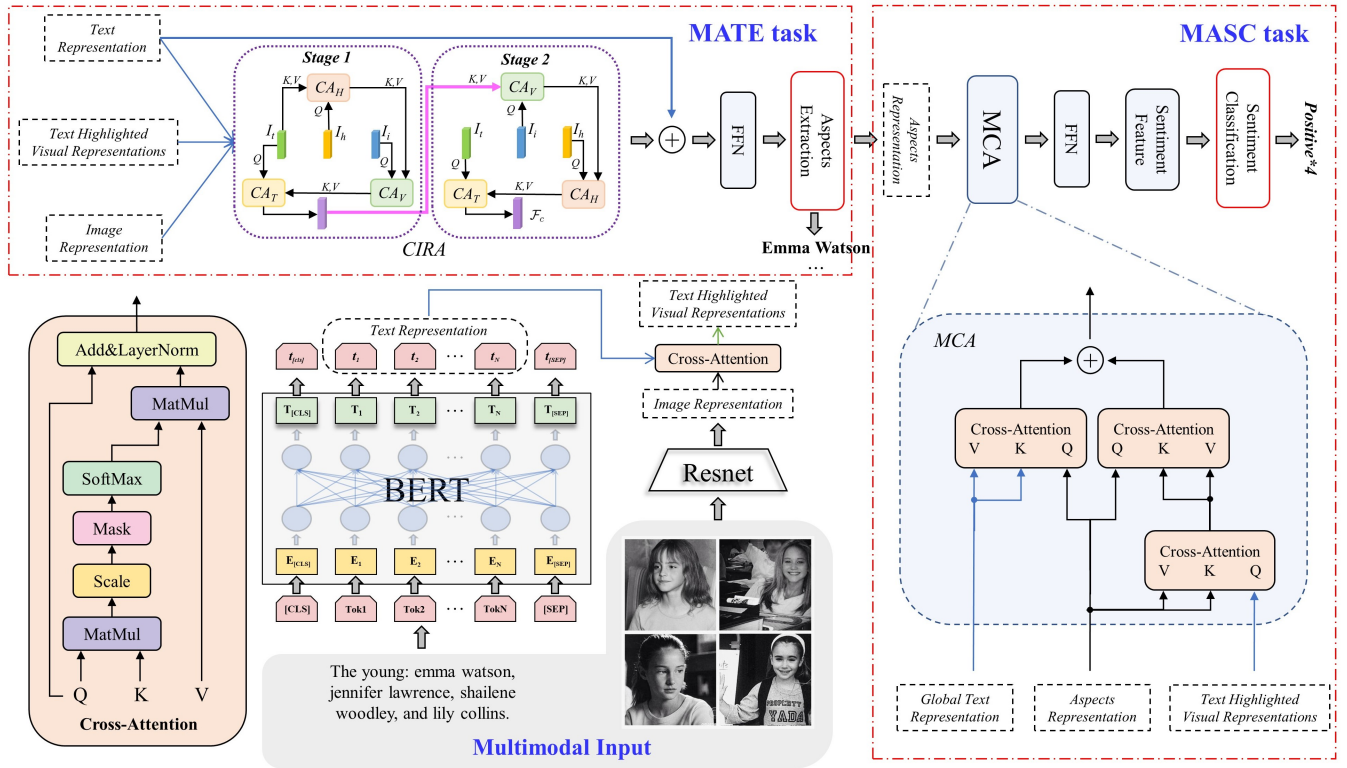


Fig. 2. The pipeline of our proposed JCC.

$$p^s = \text{softmax}(f^s), p^e = \text{softmax}(f^e). \quad (3)$$

During training, considering the possibility of multiple aspects within each input sentence, start position label vector  $y^s$  and end position label vector  $y^e$  are defined based on aspect term sequences  $A = \{a_1, a_2, \dots, a_k\}$  from the dataset. The  $n$ -th dimension  $y_n^s$  of the start position label vector indicates whether the  $n$ -th position represents the beginning of an aspect term, while  $y_n^e$  of the  $n$ -th dimension of the end position label vector indicates whether it denotes the end of an aspect term. The optimization objective is formulated as the sum of the predicted probability distribution vectors and the corresponding loss values of the label vectors:

$$L_{mate} = - \sum_{n=1}^{l-2} y_n^s \log(p_n^s) - \sum_{n=1}^{l-2} y_n^e \log(p_n^e), \quad (4)$$

where  $l$  is the length of the text term sequence,  $p_n^s$  and  $p_n^e$  denote the  $n$ -th dimension of  $p^s$  and  $p^e$ , respectively.

We analyze the aspect terms based on the obtained  $p^s$ ,  $p^e$ ,  $f^s$  and  $f^e$ . Intuitively, the positions of aspect terms are determined by selecting the top  $K$  largest values from the vector of unstandardized scores of aspect term start and end positions:  $f_k^s + f_l^e$  ( $k \leq l$ ), where  $k$  is the start position and  $l$  is the end position. However, simply taking the first  $K$  aspect term positions based on the sum of the two scores may lead to multiple predicted aspect terms referring to the same text.

Therefore, we adopt the inference algorithm outlined in Algorithm 1. Initially, the Top- $M$  score value is selected from

$f^s$  and  $f^e$  of each sample, and the eligible start positions become candidate start positions of the aspect terms. The corresponding candidate aspect term scores  $u_l$  and candidate aspect term positions  $r_l$  are consolidated into their respective sets.  $\gamma$  serves as the score threshold.

Subsequently,  $r_l$  corresponding to the highest score value is iteratively removed from the set and incorporated into the aspect term position set  $O$ . Simultaneously, the corresponding data is eliminated from the set of to-be-selected aspect term positions  $R$  and the set of scores  $U$  until the size of the aspect term location set exceeds a predefined value of  $K$  or the set of to-be-selected aspect term locations  $R$  is empty. The definition of  $u_l = f_{s_i}^s + f_{e_j}^e - (e_j - s_i + 1)$  signifies the objective of selecting aspect term start and end positions with higher unstandardized scores and shorter aspect term lengths. Each time a new position is merged in, the remaining to-be-selected set and the score set are checked for any overlap, and if any exists, it is removed to minimize redundancy in the predicted aspect terms.

#### D. Modal Complementary Attention for MASC

On the right side of Fig. 2, we truncate the text features to obtain aspect textual features  $T_a \in \mathbb{R}^{d \times n}$  utilizing the aspect positions obtained from the prediction of MATE, where  $n$  denotes the number of aspects. Subsequently,  $T_a$ ,  $T$ , and  $H$  are inputted into MASC subtask.

We introduced the MCA to address the need for a closer focus on more detailed visual features in MASC compared

---

**Algorithm 1** Aspect term extraction algorithms

---

```
1: Input:  $f^s, f^e, \gamma, K$ 
2: Output:  $O$ 
3: procedure
4:   Initialize the set of to-be-selected aspect term locations
    $R = \{\}$ , scores  $U = \{\}$ , and aspect term locations  $O = \{\}$ .
5:   The position indexes corresponding to the Top- $M$  values
   from  $f^s$  and  $f^e$  are denoted as  $S$  and  $E$ , respectively.
6:   for  $s_i$  in  $S$  do
7:     for  $e_j$  in  $E$  do
8:       if  $s_i \leq e_j$  and  $f_{s_i}^s + f_{e_j}^e \geq \gamma$  then
9:          $u_l = f_{s_i}^s + f_{e_j}^e - (e_j - s_i + 1)$ 
10:         $r_l = (s_i, e_j)$ 
11:         $R = R \cup \{r_l\}; U = U \cup \{u_l\}$ 
12:      end if
13:    end for
14:  end for
15:  while  $R \neq \{\}$  and  $\text{size}(O) < K$  do
16:     $R = R \cup \{r_l\}; U = U \cup \{u_l\}$ 
17:     $O = O \cup \{r_l\}; R = R - \{r_l\}; U = U - \{u_l\}$ 
18:    for  $w_r$  in  $R$  do
19:      if  $\text{fl}(r_l, w_r) \neq 0$  then
20:         $R = R - \{w_r\}; U = U - \{u_w\}$ 
21:      end if
22:    end for
23:  end while
24: end procedure
```

---

to MATE. However, the  $H$  remain sparse and coarse in comparison to  $T_a$  and  $T$ . To improve information density, we initially perform cross-modal fusion between  $T_a$  and  $H$ . Next,  $T_a$  serve as queries, while  $T$  and the fused highlighted visual representations  $H_a$  act as key and value for cross-modal complementary fusion. The results are concatenated and then fed to a feedforward layer to obtain the sentiment feature  $\mathcal{F}_s$ .

From  $\mathcal{F}_s$ , the emotional polarity score is initially obtained using the linear network as defined in Eq.5. Subsequently, the obtained score is normalized through the softmax function, as outlined in Eq.6, to produce the polarity probability:

$$f^p = \text{Linear}(\tanh(\text{Linear}(\mathcal{F}_s))), \quad (5)$$

$$p^p = \text{softmax}(f^p). \quad (6)$$

Thus, we can formulate the optimization objective for the MASC subtask as Eq.7:

$$L_{\text{masc}} = - \sum_{j=1}^m \sum_{n=1}^{\epsilon} y_{jn}^p \log(p_{jn}^p), \quad (7)$$

where  $m$  is the number of aspect terms,  $\epsilon$  represents the count of lexical elements in an aspect term. The sentiment label  $y_{jn}^p$  corresponds to the  $n$ -th lexical element of the  $j$ -th aspect term, while  $p_{jn}^p$  is the  $n$ -th dimension of the vector representing the probability distribution of sentiment polarities for aspect terms.

During inference, we calculate the probability of sentiment polarity within the target span for each aspect term in the set  $O$ . The sentiment category with the highest value in  $p^p$  is then selected as the sentiment category for the current aspect term.

As our proposed method involves the joint training of the MATE subtask and the MASC subtask, the overall optimization objective for the method is:

$$L = L_{\text{mate}} + L_{\text{masc}}. \quad (8)$$

### III. EXPERIMENTS

#### A. Experimental Settings

**Downstream datasets.** To assess our model’s performance, we utilize two multimodal datasets, namely Twitter-15 and Twitter-17. These datasets consist of sentences containing multiple aspects along with associated pictures.

**Implementation Details.** We implement our method using PyTorch on an NVIDIA GTX 3090, setting the hidden dimension  $d$  of BERT to 768, and using 8 heads for the  $CA$  module. Training is performed on the training dataset, validation on the dev dataset, and the model with the highest F1 value is selected for testing on the test dataset. All models underwent 50 training epochs with a fixed batch size of 128, employing the Adam optimizer with learning-rate set to  $2e-5$ .

**Evaluation Metrics.** We validate our model across the three MABSA subtasks. For JMASA and MATE, we employ Micro-F1 score (Mic-F1), Precision (P), and Recall (R) as evaluation metrics. In the case of MASC, we adhere to Accuracy (ACC) and Macro-F1 (Mac-F1) to ensure a fair comparison with prior methods.

#### B. Main Results

**Results of JMASA.** Table I summarizes the performance of various methods for JMASA. Among them, JCC demonstrates the highest overall performance, except for the R-value metric on the Twitter-15 dataset. Clearly, JCC enhances the network’s ability to perceive aspect relations and reduces visual noise by incorporating global text and highlighting image representations. Consequently, it reduces the number of false positives. Moreover, the integration of CIRA and MCA efficiently fuses multimodal features. At the same time, we observed that OSCGA-collapse has a slightly higher R-value than JCC on Twitter-15. This may be attributed to its use of Mask RCNN for object-level visual representation extraction which comes at the cost of increased complexity.

**Results of MATE.** The performance of each method on the MATE task is presented in Table II. JCC-MATE achieves the highest P and Mic-F1 values on both datasets. Additionally, JCC without CIRA achieves the highest R-value on Twitter-15 and Twitter-17. This highlights the architectural advantages of our joint method, where the highlighted visual representation exhibits a higher signal-to-noise ratio. The CIRA module effectively integrates multimodal information, significantly reducing the number of false positive samples.

**Results of MASC.** Table III showcases the performance of different methods for MASC. JCC-MASC stands out with

TABLE I  
RESULTS OF DIFFERENT APPROACHES FOR JMASA. \* DENOTES THE RESULTS ARE FROM [1].

Modality	Approaches	Twitter-15			Twitter-17		
		P	R	Mic-F1	P	R	Mic-F1
Text-based	SPAN* [15]	53.9	53.9	53.8	59.6	61.7	60.6
	D-GCN* [16]	58.3	58.8	59.7	64.2	64.1	64.1
Multi-modal Joint Task	UMT+TomBERT* [6] [11]	58.4	61.3	59.8	62.3	62.4	62.4
	OSCGA+TomBERT* [7] [11]	61.7	63.4	62.5	63.4	64.0	63.7
	UMT-collapse* [6]	60.4	61.6	61.0	60.0	61.7	60.8
	OSCGA-collapse* [7]	63.1	<b>63.7</b>	63.2	63.5	63.5	63.5
	RpBERT* [17]	49.3	46.9	48.0	57.0	55.4	56.2
	JML [1]	62.1	62.1	62.1	66.5	65.5	66.0
	<b>JCC</b>	<b>63.3</b>	63.4	<b>63.3</b>	<b>67.3</b>	65.2	<b>66.2</b>
	<b>JCC</b> w/o GText	59.8	60.5	60.1	65.7	62.8	64.2
	<b>JCC</b> w/o HLV	63.2	63.1	63.1	66.3	65.0	65.7
	<b>JCC</b> w/o CIRA	57.1	63.3	60.0	64.3	<b>66.9</b>	65.6
<b>JCC</b> w/o MCA	62.3	58.0	60.0	63.7	61.4	62.5	

TABLE II  
RESULTS OF DIFFERENT APPROACHES FOR MATE. \* DENOTES THE RESULTS ARE FROM [1].

Approaches	Twitter-15			Twitter-17		
	P	R	Mic-F1	P	R	Mic-F1
RAN* [5]	80.5	81.5	81.0	90.7	90.0	90.3
UMT* [11]	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA* [7]	81.7	82.1	81.9	90.2	90.7	90.4
JML-MATE [1]	81.0	81.1	81.1	90.8	89.9	90.3
<b>JCC-MATE</b>	<b>84.0</b>	81.7	<b>82.7</b>	<b>91.8</b>	89.5	<b>90.7</b>
<b>JCC-MATE</b> w/o HLV	83.0	81.7	82.5	90.3	90.1	90.2
<b>JCC-MATE</b> w/o CIRA	80.3	<b>83.8</b>	82.0	90.3	<b>90.8</b>	90.6

TABLE III  
RESULTS OF DIFFERENT APPROACHES FOR MASC. \* DENOTES THE RESULTS ARE FROM [1].

Approaches	Twitter-15		Twitter-17	
	ACC	Mac-F1	ACC	Mac-F1
TomBERT* [11]	74.0	71.8	70.5	68.0
CapTrBERT [10]	78.0	73.2	72.3	70.2
ESAFN* [12]	70.9	-	65.5	-
JML-MASC [1]	76.1	72.3	72.3	69.9
<b>JCC-MASC</b>	<b>78.5</b>	<b>73.9</b>	<b>73.6</b>	<b>72.2</b>
<b>JCC-MASC</b> w/o GText	76.3	71.8	72.0	70.8
<b>JCC-MASC</b> w/o HLV	77.5	72.1	72.4	71.3
<b>JCC-MASC</b> w/o MCA	76.5	72.6	71.4	70.2

the highest performance on both datasets. The introduction of global text allows the model to comprehend the correlations between aspects. Through the MCA module, it focuses on detailed multimodal features, thereby enhancing the accuracy of sentiment classification.

### C. Ablation Study

To validate the effectiveness of each module in JCC, we designed a series of ablation experiments.

**Contribution of GText.** In order to confirm the role of GText in MASC, we removed GText in the JMASA and MASC tasks, as shown in Table I and Table III respectively. The results decreased significantly, which shows that not only  $T_a$  but also  $T$  are necessary for polarity classification.

**Contribution of HLV.** In highlight the role of HLV, we substitute HLV with  $V$ . As demonstrated in Tables I, II, and III, the results decrease, signifying that HLV emphasizes certain image details. This enables the model to concentrate more on key features, thereby enhancing performance.

**Contribution of CIRA.** To investigate the validity of CIRA, we conducted an experiment by removing it and inputting  $T$ ,  $H$ , and  $V$  into the FFN after concatenating them. The results, as shown in Tables I and II, reveal that while the R-value become higher after removing CIRA, the P-value decrease significantly. This is because the model no longer focuses on integrating information but shifts its attention to all information, thereby increasing the predicted quantity, resulting in an increase in the R-value and a decrease in the P-value.

**Contribution of MCA.** As shown in Tables I and III, the removal of the MCA module resulted in a significant

decrease in all indicators for both JMASA and MASC tasks. This suggests that the MCA module plays a crucial role in integrating information from each modality and refining emotional features.

#### D. Case Study


Visual Modality		Textual Modality	
		charlie is decidedly not excited about @ ussoccer_ynt at 4 am.	
<b>GT</b>	(Charlie, negative) (ussoccer_ynt, neutral)		
<b>JML</b>	(Charlie, negative) ✓ (ussoccer_ynt, positive) ✗		
<b>JCC w/o MCA</b>	(Charlie, positive) ✗ (ussoccer_ynt, positive) ✗		
<b>JCC</b>	(Charlie, negative) ✓ (ussoccer_ynt, neutral) ✓		

Fig. 3. Visualization of predictions for JML, JCC w/o MCA, and JCC.

To demonstrate the effectiveness of our joint multimodal approach, Fig. 3 illustrates a case where JCC correctly predicts and compares it with JCC w/o MCA and JML. In this case, all models accurately extract two terms, while JML incorrectly analyzes the sentiment polarity for the term “ussoccer\_ynt”. However, after removing the MCA module from JCC, the polarity of both terms is incorrectly analyzed. This case study underscores that JCC correctly extracts aspects and predicts polarity by joint frameworks and effectively fusing visual and textual information.

#### IV. CONCLUSION

In this paper, we propose the JCC framework for MABSA task. To tackle the challenge of indirect aspect sentiment analysis, we introduce global text in the MASC task. Additionally, to mitigate the impact of visual noise, we utilize text to cross-modally highlight images, guiding the model to focus on aspect-related information. Moreover, to accommodate varying levels of detail in input features for the MATE and MASC tasks, we employ two customized modules for modal semantic fusion with different levels of detail. Experimental results across the three MABSA subtasks demonstrate the effectiveness of our approach.

#### REFERENCES

- [1] Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou, “Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection,” in *EMNLP*, 2021, pp. 4395–4405.
- [2] Yan Ling, Jianfei Yu, and Rui Xia, “Vision-language pre-training for multimodal aspect-based sentiment analysis,” in *ACL*, 2022, pp. 2149–2159.
- [3] Li Yang, Jin-Cheon Na, and Jianfei Yu, “Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis,” *Inf. Process. Manag.*, vol. 59, no. 5, pp. 103038, 2022.

- [4] Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan, “Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis,” in *ACL-Findings*, 2023, pp. 8184–8196.
- [5] Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi, “Multimodal aspect extraction with region-aware alignment network,” in *NLPCC*, 2020, pp. 145–156.
- [6] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia, “Improving multimodal named entity recognition via entity span detection with unified multimodal transformer,” in *ACL*, 2020, pp. 3342–3352.
- [7] Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li, “Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts,” in *MM’20*, 2020, pp. 1038–1046.
- [8] Meysam Asgari-Chenaghlu, Mohammad-Reza Feizi-Derakhshi, Leili Farzinvash, M. A. Balafar, and Cina Motamed, “CWI: A multimodal deep learning approach for named entity recognition from social media using character, word and image features,” *Neural Comput. Appl.*, vol. 34, no. 3, pp. 1905–1922, 2022.
- [9] Zenan Xu, Qinliang Su, and Junxi Xiao, “Multimodal aspect-based sentiment classification with knowledge-injected transformer,” in *ICME*, 2023, pp. 1379–1384.
- [10] Zaid Khan and Yun Fu, “Exploiting BERT for multimodal target sentiment classification through input space translation,” in *MM ’21*, 2021, pp. 3034–3042.
- [11] Jianfei Yu and Jing Jiang, “Adapting BERT for target-oriented multimodal sentiment classification,” in *IJCAI*, 2019, pp. 5408–5414.
- [12] Jianfei Yu, Jing Jiang, and Rui Xia, “Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 429–439, 2020.
- [13] Jianfei Yu, Kai Chen, and Rui Xia, “Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis,” *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1966–1978, 2023.
- [14] Chen Zhang, Qiuchi Li, and Dawei Song, “Aspect-based sentiment classification with aspect-specific graph convolutional networks,” in *EMNLP-IJCNLP*, 2019, pp. 4567–4577.
- [15] Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv, “Open-domain targeted sentiment analysis via span-based extraction and classification,” in *ACL*, 2019, pp. 537–546, Association for Computational Linguistics.
- [16] Guimin Chen, Yuanhe Tian, and Yan Song, “Joint aspect extraction and sentiment analysis with directional graph convolutional networks,” in *COLING*, 2020, pp. 272–279.
- [17] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng, “Rpbert: A text-image relation propagation-based BERT model for multimodal NER,” in *AAAI*, 2021, pp. 13860–13868.