Sentiment-enhanced Multi-hop Connected Graph Attention Network for Multimodal Aspect-Based Sentiment Analysis

Linlin Zhu[1] , **Heli Sun**[1,*] , **Xiaoyong Huang**[1] , **Qi Zhang**[2] , **Ruichen Cao**[1] , **Liang He**[1]
[1]College of Computer Science and Technology, Xi'an Jiaotong University
[2]College of Software Engineering, Xi'an Jiaotong University
{zhulinlin, hxy_computer, zq1230, 2127382011}@stu.xjtu.edu.cn, {hlsun, lhe}@xjtu.edu.cn

Abstract

Multimodal aspect-based sentiment analysis aims to extract aspects from different data sources and recognize the corresponding sentiments. While current research has broadly focused on syntax relation-driven semantic comprehension, the impact of the importance of different syntactic relations on semantic understanding has not been adequately investigated. To address this issue, we propose a Sentiment-enhanced **M**ulti-hop **C**onnected **G**raph Attention Network (**MCG**), aiming to enhance the discriminative capability of model for sentiments and to delve into the syntactic relationships within the text. Firstly, we design a contrastive sentiment-enhanced pre-training task that expands the diversity and complexity of training samples to improve the recognition of multiple sentiments. Secondly, we construct a multi-hop connected syntactic dependency graph to deeply explore the rich syntactic dependencies in the text and to reveal the differences among syntactic relations. Moreover, we develop a multi-hop connected graph attention mechanism that enables the model to focus on the key syntactic relations within the syntactic structure, thereby enhancing the comprehension and predictive capabilities of model in multimodal sentiment analysis. Experimental results on two benchmark datasets demonstrate that our method outperforms state-of-the-art methods. The source code is provided in the supplementary materials.

1 Introduction

With the rapid development of internet, multimodal data are playing an increasingly important role in social events and user attitude expression. Multimodal aspect-based sentiment analysis (MABSA) has garnered increasing attention from researchers due to its ability to accurately discern the sentiment inclinations of various aspects within multimodal data. It typically encompasses three subtasks: joint multimodal aspect-based sentiment analysis (JMASA), multimodal aspect term extraction (MATE), and multimodal aspect sentiment classification (MASC).

Figure 1: An example sentence contains multiple aspects with multiple sentiment polarities. ⓐ illustrates the syntactic dependency relationships of the text, the aspect terms, and their corresponding sentiment polarities. ⓑ displays the targets in the image corresponding to the aspects. ⓒ presents the syntactic structure of text in the form of a tree diagram.

Early MABSA studies [Ju *et al.*, 2021; Yu *et al.*, 2022; Li *et al.*, 2024b] enhanced the integration of cross-modal information through modality alignment techniques. For instance, JML [Ju *et al.*, 2021] utilized cross-modal relation detection, while DTCA [Yu *et al.*, 2022] adopted a dual-encoder transformer with cross-modal alignment. However, these methods overlooked the fine-grained associations between aspects and opinions. To address this issue, AoM [Zhou *et al.*, 2023] introduced aspect-oriented attention modules and graph convolutional networks, and AETS [Zhu *et al.*, 2025] developed an aspect-enhanced module to increase sensitivity to aspects, but their performance suffers in scenarios with sparse data. Current research improves sentiment detection by parsing the syntactic structure of text. AESAL [Zhu *et al.*, 2024] proposed an adaptive syntactic learning mechanism relying on word distance to evaluate semantic relationships, overlooking the differences in syntactic structure connectivity, which may affect the understanding of text semantics.

While these studies have demonstrated promising performance, our analysis reveals two challenges in MABSA: **1) In multimodal multi-sentiment scenarios, the sentiment polarity of aspect terms can be nuanced and complex, challenging models to make accurate judgments with lim-**

*Contact Author

ited data. As shown in Fig. 1 ⓐ, there are three aspects "*Barcelona*", "*La Liga*" and "*Suarez*", with corresponding sentiments of "positive", "neutral" and "negative". However, due to the influence of "*Sink*", "*Suarez*" is misjudged as negative. Integrating visual information (Fig. 1 ⓑ) in pre-training tasks for sentiment enhancement can significantly improve the model's ability to accurately distinguish sentiments, especially in data-limited multi-sentiment task. **2) Differences in syntactic relationships revealed by syntactic connectedness have different effects on MABSA.** As shown in the syntactic dependency tree in Fig. 1 ⓒ, syntactic connectivities like "*Crowned-Champions*", "*Champions-Liga*" and "*Champions-Sink*" are involved in multiple syntactic relationships, demonstrating strong syntactic connectivity and encompassing significant connections with aspect and sentiment words. This intricate syntactic structure is crucial for the precise identification of aspects and sentiment analysis. In contrast, the connectivity of "*Liga-La*", "*Sink-Suarez*" is relatively lower, with the syntactic relationships being of lesser importance. Different syntactic relationships have a certain impact on the understanding of the text. Therefore, it is crucial to pay full attention to the importance of multiple sentiments and syntactic relations for MABSA task.

To address these issues, we propose a sentiment-enhanced multi-hop connected graph attention network (MCG) for MABSA. The method enhances the sensitivity of model to sentiments and fully considers the importance differences driven by syntactic structure through a multi-hop connected graph attention mechanism. Specifically, we firstly construct sentiment-enhanced pre-training based on positive and negative samples to increase the diversity and complexity of training samples, thereby improving the ability to distinguish multiple sentiments. Secondly, we constructed a multi-hop connected syntactic dependency graph to capture the importance of different syntactic relationships, and utilized the attention mechanism of the multi-hop connected graph to effectively highlight key relationships in the syntactic structure. Our contributions are as follows:

- We are the first to consider the impact of syntactic relation differences based on connectivity for MABSA task and propose a sentiment-enhanced multi-hop connected graph attention network that analyzes the importance of key syntactic relations.
- We design a sentiment-enhanced pre-training task that expands the diversity of training samples and enhances the discriminative ability for multiple sentiments.
- We introduce a multi-hop graph attention mechanism to capture sentence structure and semantics by constructing syntactic dependency graphs and focusing on important syntactic relations for deeper text understanding.
- Experimental results on three tasks (JMASA,MATE and MASC) across two benchmark datasets indicate that MCG achieves state-of-the-art performance.

2 Related Work

Multimodal Aspect-Based Sentiment Analysis is a task that is both challenging and requires multiple processing, necessitating models that can deeply understand multimodal data and effectively extract sentiment cues [Ye *et al.*, 2022; Zhang *et al.*, 2023]. Early research primarily relied on attention mechanisms to achieve alignment between different modalities. For instance, JML [Ju *et al.*, 2021] optimizes the utilization of visual information by constructing an auxiliary text-image relation detection module and employed a hierarchical framework to facilitate multimodal interaction between MATE and MASC. DTCA [Yu *et al.*, 2022] introduces two auxiliary tasks to enhance cross-attention performance and aligned text and image modalities by minimizing the Wasserstein distance between them. VLP-MABSA [Ling *et al.*, 2022] innovatively transformes MABSA into a text generation task, incorporating, in addition to conventional masked language modeling and masked region modeling tasks, pre-training tasks such as text aspect-opinion extraction, visual aspect-opinion generation, and multimodal sentiment prediction, to more finely identify aspects, opinions, and their cross-modal consistency.

Existing research focuses on aspect enhancement and syntax mining. CMMT [Yang *et al.*, 2022] uses two auxiliary tasks to learn intra-modal representations for aspects or sentiment perception and introduces a text-guided cross-modal interaction module to dynamically control the contribution of visual information to each word's representation. AoM [Zhou *et al.*, 2023] designs an aspect-oriented attention module to select text labels and image patches relevant to aspect semantics simultaneously, incorporating sentiment embeddings into aspect semantics. Atlantis [Xiao *et al.*, 2024] enhances multimodal data with visual aesthetic attributes. AESAL [Zhu *et al.*, 2024] develops an aspect enhancement module to learn aspect correlations in multimodal input data and a syntax-adaptive learning mechanism for syntactic relationships.

Despite the significant progress made, previous studies have overlooked the handling of diverse sentiments and different syntactic relationships. To address these issues, this paper introduces the MCG model, which enhances the sentiment discriminative capability of model through contrastive sentiment pre-training and employs the multi-hop connected graph attention mechanism to capture the significance of different syntactic relationships within syntactic structures.

3 Method

In this section, we first introduce the task formulation, followed by a detailed description of the proposed sentiment-enhanced multi-hop connected graph attention network (MCG). The overall framework is shown in Fig. 2, and the network consists of the following components: feature extraction, sentiment-enhanced pre-training, multi-hop connected syntactic dependency graph, multi-hop connected graph attention, and the prediction.

3.1 Task Formulation

We assume the dataset $D = \{(T_i, V_i, A_i, S_i)_{i=1}^{K}\}$, which contains K samples. Each sample $x \in D$ includes the text $T = \{t_1, t_2, \cdots, t_n\}$, the image $V \in \mathbb{R}^{3\times H\times W}$, aspect terms $A = \{a_1, a_2, \cdots, a_m\}$ and their corresponding sentiments $S = \{s_1, s_2, \cdots, s_m\}$, where n represents the number of

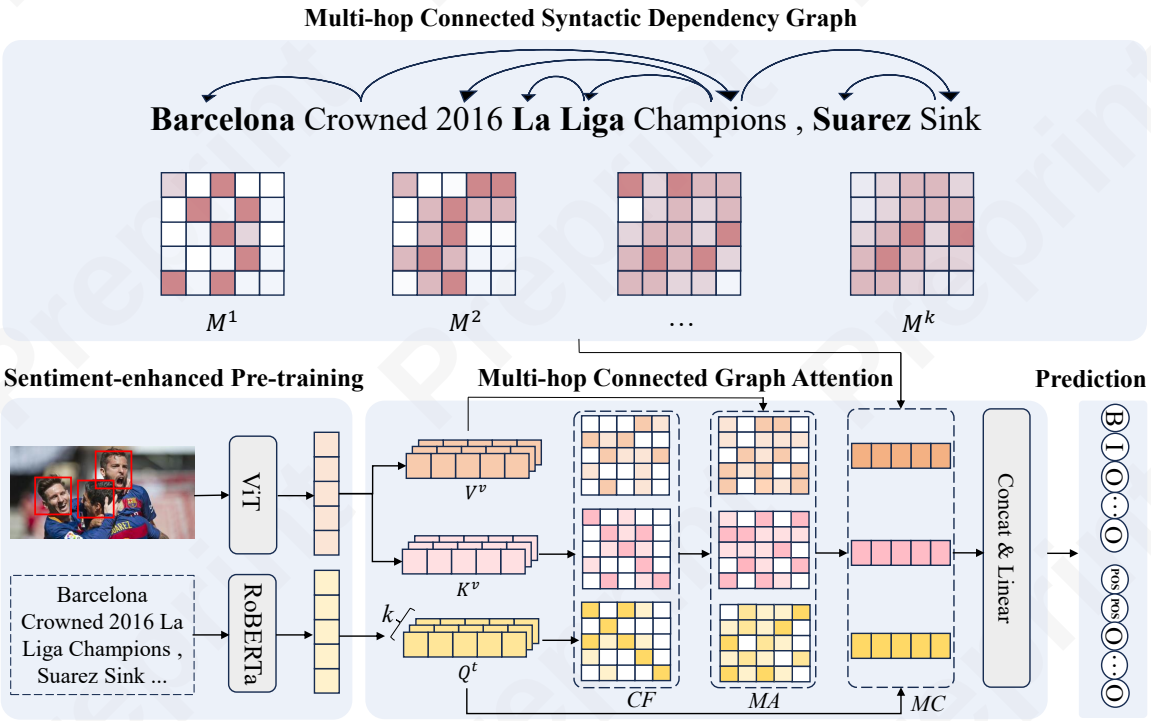

Figure 2: The overall framework of MCG.

words, 3, H, W denote the number of channels, height, width of the image, and m represents the number of aspects and sentiments. $s_i \in \{POS, NEU, NEG\}$, POS, NEU, NEG denote positive, neutral, and negative, respectively. We define the input and output for the three sub-tasks MATE, MASC, and JMASA as follows:

MATE: $input = [\{t_1, t_2, \cdots, t_n\}, V]$; $output = [a_1, a_2, \cdots, a_m]$.

MASC: $input = [\{t_1, t_2, \cdots, t_n\}, V, a_1, a_2, \cdots, a_m]$; $output = [s_1, s_2, \cdots, s_m]$.

JMASA: $input = [\{t_1, t_2, \cdots, t_n\}, V]$; $output = [\{a_1, s_1\}, \{a_2, s_2\}, \cdots, \{a_m, s_m\}]$.

3.2 Feature Extraction

Given the superior performance of RoBERTa [Liu *et al.*, 2019] for text representation and ViT [Dosovitskiy *et al.*, 2020] for visual representation, we utilize RoBERTa and ViT to encode text and images, respectively. Specifically, we append the special tokens $[cls]$ and $[sep]$ to the beginning and end of the text to indicate the start and end of a sentence, with $[cls]$ also used to mark the beginning of the image. In Equations (1) and (2), we feed the text and image into RoBERTa and ViT to derive the hidden layer states and for text and image. During this process, we use an MLP to adjust the dimensions of image to align with the text representation.

$$H^t = RoBERTa(T) \tag{1}$$

$$H^v = MLP(ViT(V)) \tag{2}$$

$H^t, H^v \in \mathbb{R}^{n \times d}$, n denotes the number of words and d denotes the hidden state dimension.

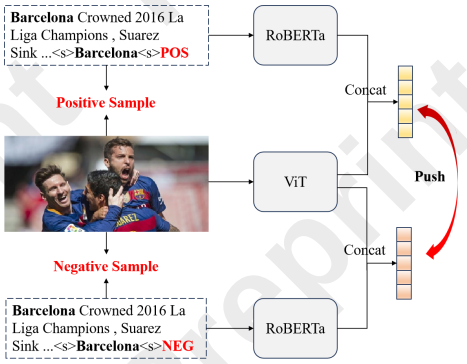

Figure 3: The framework of sentiment-enhanced pre-training.

3.3 Sentiment-enhanced Pre-training

In order to improve the discriminative capability of model in the domain of multiple sentiments recognition, we devise a contrastive sentiment-enhanced pre-training task, with the specific architecture depicted in Fig. 3.

For each training sample x, we employ the following strategy for sample processing. First, we append the aspect term a_i and its corresponding sentiment label s_i to the end of the text, forming a positive training sample P by inserting a special delimiter "<s>". Next, we randomly select a non-real sentiment label $\overline{s_i}$ from a collection of labels, which differs from the real sentiment label s_i (where $\overline{s_i} \in \{POS, NEU, NEG\}$ and $\overline{s_i} \neq s_i$), and combine it with the aspect term a_i to generate a negative training sample

N. During this process, we treat (P, V) as a positive training pair and (N, V) as a negative training pair. Subsequently, the positive and negative training pairs are input into the feature extraction module to obtain text embeddings and image embeddings. These two embedding vectors are then concatenated, and a cross-entropy loss function $\mathcal{L}_f$ is employed to increase the distance between the positive and negative training pairs as in Equation (3).

$$\mathcal{L}_f = -B \log(\hat{B}) - (1 - B) \log(1 - \hat{B}) \tag{3}$$

B represents the true label of the sample pair, and $\hat{B}$ is the probability that the model predicts the sample pair as positive sample.

The method of randomly selecting a non-real sentiment label as a negative example helps to balance positive sentiment information and noise (including irrelevant or misleading information), thereby encouraging the model to focus on the key differences between positive and negative training pairs. This design strategy significantly improves the performance in two aspects: Firstly, as a data augmentation technique, it increases the complexity and diversity of the dataset. Secondly, by enhancing the discrimination between positive and negative training pairs, it enhances the ability of model to discriminate between ambiguous sentiments and multiple sentiment categories.

3.4 Multi-hop Connected Syntactic Dependency Graph

The syntactic dependency graph highlights syntactic relationships in sentences. AESAL [Zhu *et al.*, 2024] evaluates these relationships by node proximity, suggesting that closer nodes are more significantly associated. However, in semantic comprehension, nodes with more connections are often more crucial [Borge-Holthoefer and Arenas, 2010], due to their involvement in multiple syntactic relationships and higher connectivity. Thus, we build a multi-hop syntactic dependency graph to analyze connectivity, capturing variations in syntactic relationships for a deeper semantic understanding of text.

We represent the significance of different syntactic relationships in the syntactic dependency tree through the connectivity of the graph. Specifically, we first use the spaCy[1] library to obtain the syntactic dependency tree and convert it into an undirected graph G. The adjacency matrix $M \in \mathbb{R}^{n \times n}$ of graph G represents its connectivity, where n is the number of words. M^k represents the number of paths of length k, and M^k_{ij} represents the number of paths of length k from node i to node j. The proof is given in Theorem 1.

Theorem 1. *For any graph G with adjacency matrix M, the element M^k_{ij} represents the number of paths of length k from vertex i to vertex j.*

We perform kth power operations on the syntactic dependency graph and normalize it according to the following Equations (4) and (5) to obtain the weights of syntactic relationships:

$$\text{Norm}(M) = \widetilde{D}^{\frac{1}{2}}(M + I)\widetilde{D}^{-\frac{1}{2}} \tag{4}$$

[1]https://spacy.io/

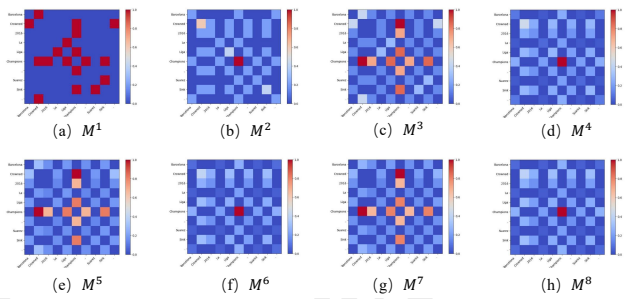

Figure 4: Normalized weight matrices from M^1 to M^8 for the syntactic dependency graph of Fig. 1 ⓐ. Blue indicates lower weight values, red denotes higher weight values.

$$\begin{aligned} \widetilde{M} &= \{\widetilde{M}^1, \widetilde{M}^2, \cdots, \widetilde{M}^k\} \\ &= \{\text{Norm}(M^1), \text{Norm}(M^2), \cdots, \text{Norm}(M^k)\} \end{aligned} \tag{5}$$

Here, I is the identity matrix, D is the degree matrix of $M + I$, and $\widetilde{D}$ is the diagonal matrix. $\widetilde{M}^1$, $\widetilde{M}^2$, ..., $\widetilde{M}^k$ represent the weights of syntactic dependency relationships from first-hop to kth-hop, $\widetilde{M}$ denotes the weight of the multi-hop syntactic dependency graph. Through experiment, we found that the length of syntactic relations typically ranges between 1 and 8, which corresponds to $k \in \{1, 2, ..., 8\}$.

As shown in Fig. 4, we present the normalized graph of M^1-M^8 for the example in Fig. 1 ⓐ. In the syntactic dependency tree, the importance of different nodes can be measured by their connectivity in the graph. It is observed that syntactic relationships like "*Champions-Champions*", "*Crowned-Champions*", "*Sink-Champions*" and "*Liga-Champions*" have larger values. This is because words like "*Champions*", "*Crowned*", "*Sink*" and "*Liga*" are involved in more syntactic relationships, resulting in stronger connectivity. In our method, by calculating the connectivity of different orders, we can more accurately capture these important syntactic relationships.

3.5 Multi-hop Connected Graph Attention

To effectively highlight the role of key syntactic relationships in the syntactic structure, particularly those that play a core role in aspect and sentiment expression, we propose a multi-hop connected graph attention mechanism based on the multi-head cross-attention mechanism. This mechanism enhances the focus on local contexts while also considering the flow of information in a broader context, aiding in the capture of deep semantic associations.

We employ a multi-head cross-attention mechanism to obtain a multi-head attention matrix MA for text and image, capturing their interactive relationships. We first define the single-head cross-attention matrix function (CF) in Equation (6). In Equations (7) and (8), we define the multi-head cross-attention matrix function (MCF) with k heads, and obtain a multi-head attention matrix MA that matches the $\widetilde{M}$.

$$CF(Q, K) = Softmax(\frac{QW_Q \times (KW_K)^T}{\sqrt{d_k}}) \tag{6}$$

$$MCF(Q,K) = \{CF^1(Q,K), CF^2(Q,K), \cdots, CF^k(Q,K)\} \tag{7}$$

$$MA = MCF(H^t, H^v) \tag{8}$$

where Q,K denotes the query and key vectors. W_Q and W_K are learnable parameters, and d_k denotes the dimension of k.

After that, in Equation (9), we use the multi-hop syntactic dependency graph weights $\widetilde{M}$ in combination with the multi-head attention matrix MA to obtain the multi-hop connected graph attention MC. In this way, the model enhances the refinement of the unimodal intra-features while also optimizing the interactions of the cross-modal features.

$$MC = \widetilde{M} \odot MA \tag{9}$$

where $\odot$ denotes element-by-element multiplication.

Finally, as shown in Equation (10), we multiply the multi-hop connected graph attention MC with the value vector to obtain the multi-modal fusion feature H^f with multi-hop syntactic dependency graph information.

$$H^f = MCH^vW_v \tag{10}$$

Here, W_v is the learnable parameter.

3.6 Prediction

As shown in Equations (11) and (12), we use a two-layer MLP as the predictor and use relu as the activation function and cross-entropy loss as the objective function.

$$\widehat{y} = Softmax(ReLU(H^fW_1 + b1)W_2 + b_2) \tag{11}$$

$$\mathcal{L} = -\sum y_i log(\widehat{y}_i) \tag{12}$$

where W_1,W_2,b_1,b_2 are the learnable parameters, $\widehat{y}$ is the predicted outcome, and y is the true label.

4 Experiment

In this section, we verify the validity and superiority of our proposed MCG through comparative experiments and ablation experiments on JMASA, MATE and MASC tasks.

4.1 Experiment Data and Setup

Dataset: we use two multimodal benchmark datasets, Twitter-2015 and Twitter-2017 [Yu and Jiang, 2019], for our experiments. These two Twitter datasets collect user posts published on Twitter during 2014-2015 and 2016-2017, respectively. Table 1 summarizes the datasets.

Experiment Setup: Our experiments are implemented in the PyTorch framework using NVIDIA 3090 GPUs, with the learning rate set to 2e-5, the hidden layer dimension to 768, and the dropout set to 0.1.

Evaluation Metrics: On the JMASA task and the MATE task, we evaluate our model using Micro-F1 (F1), Precision (P) and Recall (R). And on the MASC task, we use Accuracy (Acc) and Micro-F1 (F1) following previous studies [Zhu *et al.*, 2024; Zhou *et al.*, 2023].

Label	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1,508	515	493
Neutral	1,883	670	607	1,638	517	573
Negative	368	149	113	416	144	168
Total Aspects	3,179	1,122	1,037	3,562	1,176	1,234
Total Sentence	2,101	727	674	1,746	577	587
# multi sentiment		1,257			1,690	

Table 1: Statistics of the Twitter-2015 and Twitter-2017 datasets. "# multi sentiment" in the last line denotes the number of sentences with multiple sentiment.

4.2 Baselines

We compare MCG with four types of methods.

Methods for textual ABSA. 1) **SPAN** [Hu *et al.*, 2019] identifies opinion targets with their sentiments. 2) **D-GCN** [Chen *et al.*, 2020] utilizes GCN [Wu *et al.*, 2020a] to model syntactic dependencies. 3) **BART** [Yan *et al.*, 2021] addresses seven ABSA subtasks.

Methods for JMASA. 1) **UMT-collapse** [Yu *et al.*, 2020], **OSCGA-collapse** [Wu *et al.*, 2020b] and **Rpbert-collapse** [Sun *et al.*, 2021] use the same visual input to collapse individual tokens. 2) **UMT+TomBERT**, **OSCGA+TomBERT** are two pipelines that combine UMT, OSCGA and TomBERT [Yu and Jiang, 2019]. 3) **JML** [Ju *et al.*, 2021] introduces hierarchical text-image relation detection with auxiliary modules for joint MATE-MASC optimization. 4) **VLP- MABSA** [Ling *et al.*, 2022] is a multimodal framework employing five dedicated pretraining tasks to model aspect features, opinion expressions and cross-modal alignment. 5) **CMMT** [Yang *et al.*, 2022] is a cross-modal learning approach that regulates multimodal information interaction through gating mechanisms. 6) **AoM** [Zhou *et al.*, 2023] is an aspect-oriented method that simultaneously detects aspect-relevant semantic information and sentiment features from multimodal inputs. 7) **Atlantis** [Xiao *et al.*, 2024] is a framework that augments multimodal representations using visual aesthetic characteristics. 8) **AESAL** [Zhu *et al.*, 2024] is a method featuring an aspect correlation learning module and a syntax-adaptive mechanism for multimodal input processing. 9) **GLM-4V-Plus** [Hong *et al.*, 2024], **Llama-3.2-11B-Vision-Instruct** (Llama-3.2) [AI@Meta, 2024], and **llama3-llava-next-8b-hf** (LLaVA-NeXT) [Li *et al.*, 2024a], all advanced large language models(LLMs), were meticulously directed to perform specific tasks in our experiments through carefully crafted prompt information.

Methods for MATE. 1) **RAN** [Wu *et al.*, 2020b] specifically emphasizes aligning text with object regions. 2) **UMT** [Yu *et al.*, 2020] effectively uses text-based entity span detection as an auxiliary task. 3) **OS-CGA** [Wu *et al.*, 2020b] primarily focus on aligning visual objects with entities.

Methods for MASC. 1) **ESAFN** [Yu *et al.*, 2019] employs LSTM networks to perform sentiment analysis at the entity level. 2) **TomBERT** [Yu and Jiang, 2019] is a target-oriented multimodal utilizes BERT architecture to generate aspect-aware text representations. 3) **CapTrBERT** [Khan and Fu, 2021] translates images into text and constructs an auxiliary sentence for fusion, enhancing cross-modal interac-

	Methods	Venue	Twitter-2015			Twitter-2017		
			P	R	F1	P	R	F1
Text-based	SPAN*	ACL 2020	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN*	COLING 2020	58.3	58.8	59.4	64.2	64.1	64.1
	BART*	ACL 2021	62.9	65.0	63.9	65.2	65.6	65.4
Multimodal	UMT+TomBERT*	ACL 2020, IJCAI 2019	58.4	61.3	59.8	62.3	62.4	62.4
	OSCGA+TomBERT*	ACM MM 2020, IJCAI 2019	61.7	63.4	62.5	63.4	64.0	63.7
	OSCGA-collapse*	ACM MM 2020	63.1	63.7	63.1	63.5	63.5	63.5
	RpBERT-collapse*	AAAI 2021	49.3	46.9	48.0	57.0	55.4	56.2
	UMT-collapse*	ACL 2020	61.0	60.4	61.6	60.8	60.0	61.7
	JML	EMNLP 2021	65.0	63.2	64.1	66.5	65.5	66.0
	VLP-MABSA	ACL 2022	65.1	68.3	66.6	66.9	69.2	68.0
	CMMT	IPM 2022	64.6	68.7	66.5	67.6	69.4	68.5
	AoM	ACL 2023	67.9	69.3	68.6	68.4	71.0	69.7
	Atlantis	Inf. Fusion 2024	65.6	69.2	67.3	68.6	70.3	69.4
	AESAL	IJCAI 2024	68.7	70.4	69.5	69.4	74.8	72.0
	MCG	**Ours**	**71.3**	**70.5**	**70.9**	**77.2**	**78.8**	**77.1**
LLMs	GLM-4V-Plus	Zhipu AI 2024	55.4	64.2	56.9	61.7	61.0	59.9
	Llama-3.2	Meta AI 2024	42.5	52.0	44.3	50.8	54.9	50.9
	LLaVA-NeXT	Meta AI 2024	38.4	56.1	42.1	48.2	59.9	50.7

Table 2: Results of different models on JMASA on the two Twitter datasets. Our model MCG almost achieves the current optimal results on JMASA. The best results are bold-typed and the second best ones are underlined. * denotes the results from [Zhu *et al.*, 2024].

Methods	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
RAN*	80.5	81.5	81.0	90.7	90.7	90.0
UMT*	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA*	81.7	82.1	81.9	90.2	90.7	90.4
JML	83.6	81.2	82.4	92.0	90.7	91.4
VLP-MABSA	83.6	87.9	85.7	90.8	92.6	91.7
CMMT	83.9	88.1	85.9	92.2	93.9	93.1
AoM	84.6	87.9	86.2	91.8	92.8	92.3
Atlantis	84.2	87.7	86.1	91.8	93.2	92.7
AESAL	90.2	90.6	90.4	93.1	96.4	94.7
MCG	**91.5**	**91.1**	**91.3**	**93.6**	**97.3**	**95.4**
GLM-4V-Plus	57.2	74.1	64.6	68.4	79.2	73.4
Llama-3.2	44.4	67.4	53.5	53.1	66.6	59.1
LLaVA-NeXT	39.0	71.2	50.4	55.9	73.1	63.3

Table 3: Results of different methods for MATE. Our model MCG achieves the current optimal results on MATE. * denotes the results from [Zhu *et al.*, 2024].

Methods	Venue	Twitter-2015		Twitter-2017	
		Acc	F1	Acc	F1
ESAFN*	ACM 2019	73.4	67.4	67.8	64.2
TomBERT*	IJCAI 2019	77.2	71.8	70.5	68.0
CapTrBERT*	ACM 2021	78.0	73.2	72.3	70.2
JML	EMNLP 2021	78.7	-	72.3	-
VLP-MABSA	ACL 2022	78.6	73.8	73.8	71.8
CMMT	IPM 2022	77.9	-	73.8	-
AoM	ACL 2023	80.2	75.9	76.4	75.0
Atlantis	Inf. Fusion 2024	79.3	-	74.2	-
AESAL	IJCAI 2024	80.1	75.2	78.8	75.9
MCG	**Ours**	**81.4**	**79.3**	**80.1**	**79.0**
GLM-4V-Plus	Zhipu AI 2024	68.7	68.8	62.6	62.5
Llama-3.2	Meta AI 2024	59.9	59.8	51.2	51.1
LLaVA-NeXT	Meta AI 2024	65.0	64.9	57.5	57.6

Table 4: Results of different models on MASC task. Our model MCG achieves the current optimal results on MASC. * denotes the results from [Zhu *et al.*, 2024].

tion through generated captions.

4.3 Experiment Results

In this section, we show the excellent performance of MCG.

Performance on JMASA. As shown in Table 2, **first**, MCG significantly surpasses all text-based models, indicating that detecting richer visual and textual information is effective. **Second**, MCG also outperforms all multimodal methods across various metrics. In particular, compared to the second-best AESAL, the P value increased by 2.6% on Twitter-2015 and by 7.8% on Twitter-2017. These results demonstrate that the identification of syntactic relationships based on syntactic connectivity is more effective than AESAL that rely solely on distance. **Last**, compared to LLMs, which requires aspect identification before sentiment determination, MCG can more accurately recognize both aspect and sentiment simultaneously, significantly reducing the task burden and demonstrating stronger task coherence.

Performance on MATE. As shown in Table 3, MCG achieved satisfactory results on the MATE task. **First**, MCG achieved optimal results across all metrics. These results demonstrate that MCG more accurately identifies core syntactic nodes through connectivity, facilitating the capture of syntactic structures that contain aspect terms and thereby en-

Methods	JMASA						MATE						MASC			
	Twitter-2015			Twitter-2017			Twitter-2015			Twitter-2017			Twitter-2015		Twitter-2017	
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc	F1	Acc	F1
Full	**71.3**	70.5	**70.9**	**77.2**	**78.8**	**77.1**	**88.2**	**94.1**	**90.9**	**93.3**	96.3	**94.8**	**81.4**	**79.3**	**80.1**	**79.0**
W/o Img	63.4	68.1	65.2	69.6	76.2	72.6	86.4	92.1	89.7	88.2	97.3	92.3	80.3	68.2	79.0	72.7
W/o SEP	69.9	**70.7**	70.3	74.2	74.6	73.3	87.2	94.0	90.3	91.7	**97.5**	94.4	79.4	74.6	78.8	73.0
W/o MCGA	68.5	70.1	67.6	76.6	77.9	74.8	87.6	93.2	89.7	90.3	95.4	93.5	80.6	72.7	79.1	77.3

Table 5: Comparison of the performance of the complete model and its ablation methods on JMASA, MATE and MASC.

hancing aspect extraction. **Second**, compared to LLMs, our multi-hop syntactic dependency graph better leverages contextual information to identify correct aspects.

Performance on MASC. Table 4 highlights MCG's top performance on the MASC. **First**, on Twitter-2015, it boosts the F1 score by 3.4% over the runner-up AoM model, and by 3.1% compared to AESAL on Twitter-2017. This improvement is due to sentiment-enhanced pre-training, enhancing the model's sentiment differentiation. **Second**, MCG surpasses LLMs in identifying fine-grained sentiments.

4.4 Ablation Study

In this section, to demonstrate the effectiveness of each module of MCG, we compare the variants of MCG in terms of image (Img), sentiment-enhanced pre-training (SEP), multi-hop connected graph attention (MCGA).

W/o Img is a variant of MCG that removes image information but utilizes textual information. **W/o SEP** is a variant of MCG that does not use sentiment-enhanced pre-training. **W/o MCGA** is a variant of MCG that removes multi-hop connected graph attention and only uses cross attention.

As shown in Table 5, we found that: **1)** Visual information enhances the accuracy of aspect word and sentiment recognition when interacting with textual data. MCG outperforms W/o Img across all three tasks, highlighting the effectiveness of visual inputs. **2)** SEP can improve the capacity of multiple sentiment discrimination. Taking the MASC as an example, on the two datasets, MCG improves the Acc and F1 values by 2%, 4.7%, 1.3%, and 6%, respectively, compared to W/o SEP, demonstrating its effectiveness. **3)** MCGA captures the importance of different syntactic relationships, enhancing focus on both local context and global information. On MATE, metrics for W/o MCGA decrease on both datasets, proving MCGA's effectiveness in highlighting key syntactic relations for text semantic understanding.

4.5 Case Study

Fig. 5 illustrates two examples of predictions using JML, CMMT, and MCG to further substantiate the efficacy of MCG. In Example (a), JML failed to fully recognize "*Fifth Harmony*" and misjudged the sentiment for the "*BBMAs*". CMMT incorrectly predicted the sentiment for the "*BBMAs*". However, MCG successfully determine this sentiment, which is attributed to our sentiment-enhanced pre-training task. In Example (b), both JML and CMMT provided incorrect sentiment judgments for "*Darrelle Revis*", with JML failing to identify the aspect term "*cowboys*" and its sentiment, and CMMT erroneously identifying "*CBs*" as "*rookie CBs*". Ours

	(a) **Fifth Harmony** ' s seats at the @ **BBMAs** ! They are in front of **Kelly** Rowland and behind **Kesha** !	(b) **Dez Bryant** believes in rookie **CBs** despite **Darrelle Revis** tweet # **cowboys** # NFL .
Text Image		
True Label	**(Fifth Harmony, POS)** **(BBMAs, NEU)** **(Kelly Rowland, POS)** **(Kesha, POS)**	**Dez Bryant, NEU)** **(CBs, POS)** **(Darrelle Revis, NEU)** **(cowboys, NEU)** **(NFL, NEU)**
JML	(Harmony, POS)(×,√) (BBMAs, POS)(√,×) (Kelly Rowland, POS)(√,√) (Kesha, POS)(√,√)	(Dez Bryant, NEU)(√,√) (CBs, POS)(√,√) (Darrelle Revis, NEG)(√, ×) (-, -)(×, ×) (NFL, NEU)(√,√)
CMMT	(Fifth Harmony, POS)(√,√) (BBMAs, POS)(√, ×) (Kelly Rowland, POS)(√,√) (Kesha, POS)(√,√)	(Dez Bryant, NEU)(√,√) (rookie CBs, POS)(×,√) (Darrelle Revis, NEG)(√, ×) (cowboys, NEU)(√,√) (NFL, NEU)(√,√)
Ours	(Fifth Harmony, POS) (√,√) (BBMAs, NEU) (√,√) (Kelly Rowland, POS) (√,√) (Kesha, POS) (√,√)	(Dez Bryant, NEU) (√,√) (CBs, POS) (√,√) (Darrelle Revis, NEU) (√,√) (cowboys, NEU) (√,√) (NFL, NEU) (√,√)

Figure 5: Predictions of different methods on two test samples.

MCG, however, identified all aspect terms in both cases and provided correct sentiment predictions. This demonstrates the effectiveness of utilizing multi-hop connected graph attention to focus on key syntactic relationships for MABSA.

5 Conclusion

This paper introduces a sentiment-enhanced multi-hop connected graph attention network for MABSA. Firstly, we construct positive and negative samples to perform comparative sentiment pre-training, thereby enhancing the diversity of the training samples and improving the multiple sentiment discrimination capability. Secondly, we build a multi-hop connected syntactic dependency graph to capture diverse syntactic relationships, and utilize the multi-hop connection graph attention mechanism to emphasize significant syntactic relationships within the syntactic structure. Lastly, we conduct three tasks on two widely used datasets, and the experimental results validate the effectiveness of our method. Additionally, we examine the model's capability to address nuanced sentiment expressions through extensive empirical evaluations.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant 62472348, in part by the Aviation Science Foundation 2023M071070002 and 2024M071070001, in part by the Key Research and Development Program of Shaanxi under Grant 2023-YBGY-230 and 2024GX-YBXM-533, in part by the Innovation Capability Support Plan of Shaanxi under Grant 2022PT-33, in part by the Xi'an Science and Technology plan Key industrial chain technology research project under Grant 23ZDCYJSGG0007, and in part by the Xi'an Science and Technology plan Key industrial chain, key core technology research project under Grant 23LLRH0022, and Qinchuangyuan Construction of Two Chain Integration Important Project 23LLRHZDZX0006, and Sichuan Science and Technology Program 2025ZNSFSC0468. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[AI@Meta, 2024] AI@Meta. Llama 3 model card. 2024.

[Borge-Holthoefer and Arenas, 2010] Javier Borge-Holthoefer and Alex Arenas. Semantic networks: Structure and dynamics. *Entropy*, 12(5):1264–1302, 2010.

[Chen *et al.*, 2020] Guimin Chen, Yuanhe Tian, and Yan Song. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th international conference on computational linguistics*, pages 272–279, 2020.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Hong *et al.*, 2024] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

[Hu *et al.*, 2019] Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546, July 2019.

[Ju *et al.*, 2021] 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 *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4395–4405, 2021.

[Khan and Fu, 2021] Zaid Khan and Yun Fu. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3034–3042, 2021.

[Li *et al.*, 2024a] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024.

[Li *et al.*, 2024b] You Li, Han Ding, Yuming Lin, Xinyu Feng, and Liang Chang. Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. *Artificial Intelligence Review*, 57(4):78, 2024.

[Ling *et al.*, 2022] Yan Ling, Jianfei Yu, and Rui Xia. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2149–2159, May 2022.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Sun *et al.*, 2021] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868, 2021.

[Wu *et al.*, 2020a] Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. Multimodal aspect extraction with region-aware alignment network. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9*, pages 145–156. Springer, 2020.

[Wu *et al.*, 2020b] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1038–1046, 2020.

[Xiao *et al.*, 2024] Luwei Xiao, Xingjiao Wu, Junjie Xu, Weijie Li, Cheng Jin, and Liang He. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Information Fusion*, 106:102304, 2024.

[Yan *et al.*, 2021] Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online, August 2021.

[Yang *et al.*, 2022] Li Yang, Jin-Cheon Na, and Jianfei Yu. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing & Management*, 59(5):103038, 2022.

[Ye *et al.*, 2022] Junjie Ye, Jie Zhou, Junfeng Tian, Rui Wang, Jingyi Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. Sentiment-aware multimodal pre-training for multimodal sentiment analysis. *Knowledge-Based Systems*, 258:110021, 2022.

[Yu and Jiang, 2019] Jianfei Yu and Jing Jiang. Adapting bert for target-oriented multimodal sentiment classification. IJCAI, 2019.

[Yu *et al.*, 2019] Jianfei Yu, Jing Jiang, and Rui Xia. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439, 2019.

[Yu *et al.*, 2020] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Meeting of the Association for Computational Linguistics*, 2020.

[Yu *et al.*, 2022] Zhewen Yu, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. Dual-encoder transformers with cross-modal alignment for multimodal aspect-based sentiment analysis. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 414–423, 2022.

[Zhang *et al.*, 2023] Xulang Zhang, Rui Mao, Kai He, and Erik Cambria. Neuro-symbolic sentiment analysis with dynamic word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8772–8783, 2023.

[Zhou *et al.*, 2023] 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 *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8184–8196, July 2023.

[Zhu *et al.*, 2024] Linlin Zhu, Heli Sun, Qunshu Gao, Tingzhou Yi, and Liang He. Joint multimodal aspect sentiment analysis with aspect enhancement and syntactic adaptive learning. In *IJCAI*, 2024.

[Zhu *et al.*, 2025] Linlin Zhu, Heli Sun, Qunshu Gao, Yuze Liu, and Liang He. Aspect enhancement and text simplification in multimodal aspect-based sentiment analysis for multi-aspect and multi-sentiment scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.