



Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis

You Li¹ · Han Ding¹ · Yuming Lin¹ · Xinyu Feng¹ · Liang Chang¹

Accepted: 30 December 2023 / Published online: 1 March 2024
© The Author(s) 2024

Abstract

Multimodal Aspect-Based Sentiment Analysis (MABSA) is an essential task in sentiment analysis that has garnered considerable attention in recent years. Typical approaches in MABSA often utilize cross-modal Transformers to capture interactions between textual and visual modalities. However, bridging the semantic gap between modalities spaces and addressing interference from irrelevant visual objects at different scales remains challenging. To tackle these limitations, we present the Multi-level Textual-Visual Alignment and Fusion Network (MTVAF) in this work, which incorporates three auxiliary tasks. Specifically, MTVAF first transforms multi-level image information into image descriptions, facial descriptions, and optical characters. These are then concatenated with the textual input to form a textual+visual input, facilitating comprehensive alignment between visual and textual modalities. Next, both inputs are fed into an integrated text model that incorporates relevant visual representations. Dynamic attention mechanisms are employed to generate visual prompts to control cross-modal fusion. Finally, we align the probability distributions of the textual input space and the textual+visual input space, effectively reducing noise introduced during the alignment process. Experimental results on two MABSA benchmark datasets demonstrate the effectiveness of the proposed MTVAF, showcasing its superior performance compared to state-of-the-art approaches. Our codes are available at <https://github.com/MKMaS-GUET/MTVAF>.

Keywords Multimodal aspect-based sentiment analysis · Textual-visual alignment · Multi-scale fusion · Multi-granularity translation

1 Introduction

Sentiment analysis, also known as opinion mining, plays a crucial role in natural language processing. Its objective is to analyze sentiments, opinions, evaluations, attitudes, and emotions expressed in user-generated content online, such as tweets. Aspect-Based Sentiment Analysis (ABSA) is a fine-grained sentiment analysis task that attracts significant attention

✉ Yuming Lin
ymlin@guet.edu.cn

¹ Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Jinji Road, Guilin 541004, Guangxi, China

due to its ability to offer detailed sentiment information, making it applicable to various scenarios. Many previous works on ABSA have focused on analyzing entities expressing user sentiment and their interrelationships from text, such as aspect terms, opinion terms, and sentiment polarities (Chen et al. 2020; Li et al. 2022; Liang et al. 2023). Users often express opinions through multimodal posts with both text and images, rather than just text. Analyzing these multimodal inputs enables more effective aspect-based sentiment analysis. The recent introduction of the multimodal aspect-based sentiment analysis (MABSA) task determines sentiment polarities towards different aspects mentioned in text-image pairs (Xu et al. 2019).

Previous research of MABSA is typically divided into three subtasks: Multimodal Aspect Term Extraction (MATE) (Wang et al. 2022; Chen et al. 2022), Multimodal Aspect-oriented Sentiment Classification (MASC) (Khan and Fu 2021), and Joint Multimodal Aspect-Sentiment Analysis (JMASA) (Ju et al. 2021; Ling et al. 2022). Given a text-image pair as input, MATE aims to extract all the aspect terms mentioned in the text, MASC focuses on detecting the sentiment corresponding to specific aspect terms, and JMASA is designed to jointly extract aspect terms and their corresponding sentiments. For example, considering the two image-text pairs shown in Fig. 1, the goal of JMASA is to identify all aspect-sentiment pairs, such as (Brent Seabrook, Positive) and (Blackhawks, Negative) in (a), as well as (Jesse Eisenberg, Positive) in (b).

Since the output of the JMASA task includes both the results of MATE and MASC, it typically poses a greater challenge compared to either of them individually. Ju et al. (2021) regarded the global correlation between sentences and images as the degree to integrate visual cues into textual representations, almost completely neglecting the role of object-level visual information. In contrast to coarse-grained text-image relevance judgments, the text in JMASA tasks comprises multiple complete aspect terms, while images often provide information related to only one or a few of these aspects. The information across different modalities is not always entirely consistent. Subsequently, Yang et al. (2022) leverage auxiliary tasks to capture highly sensitive multimodal representations. However, focusing too much on emotionally sensitive areas in visuals may introduce noise from unrelated images

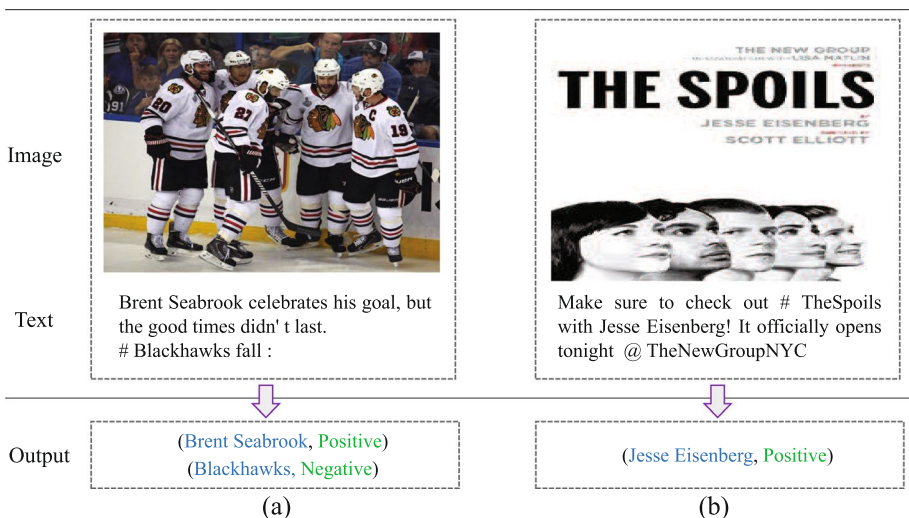


Fig. 1 Two examples of the MABSA task

during cross-modal interactions. As shown in Fig. 1(a), the celebration action, serving as a region of interest (ROI) feature, exhibits low relevance to the *Blackhawks*, yet it still impacts the sentiment prediction with negative ground truth. Recently, Ling et al. (2022) devised a task-specific pre-training framework for MABSA closely associated with downstream tasks. Nevertheless, their work solely considers aligning fine-grained object visual information with text (Chen et al. 2022), overlooking hierarchical alignment of vision and language modalities at multiple granularities. For instance, in Fig. 1(b), visual information is distributed across the global images, local faces, and text in the image, each corresponding to image descriptions at different levels. In this complex multimodal context, achieving cross-modal alignment and fusion between text and visual information at various levels is a significant challenge.

In this paper, we propose a multi-level textual-visual alignment and fusion network called MTVAF for the JMASA task to address the aforementioned limitations. We first construct an image-text alignment module that translates hierarchical visual information into the textual space by leveraging multi-granularity visual information. Next, the textual (\mathbf{T}) input is connected with multi-level visual context as textual+visual ($\mathbf{T}+\mathbf{V}$) input and fed into a text-based pre-trained language model. To inject visual sentiment knowledge from low to high levels into the network, we design a multi-scale visual aspect-opinion fusion module that dynamically integrates additional visual features into the text modal model using a dynamic attention mechanism. Additionally, we process and obtain top-N visual aspect-opinion, providing explicit fine-grained visual cues. Finally, we adopt a text-centered multimodal training approach using multi-scale visual data, enhancing the robustness of the proposed model by minimizing the KL divergence between the probability distributions of the \mathbf{T} input space and the $\mathbf{T}+\mathbf{V}$ input space. Our main contributions are summarized as follows:

1. We propose a multi-level textual-visual alignment and fusion network that bridges the semantic gap between text and images, which aligns them at multiple granularities and fully integrates hierarchical visual features into the textual space as context.
2. Building upon modal alignment, we further devise a multi-scale visual aspect-opinion fusion module that effectively incorporates significant visual information into the Transformer model, adaptively learning fine-grained knowledge from images.
3. Extensive experiments on two MABSA benchmark datasets demonstrate that the proposed model consistently outperforms existing unimodal and multimodal approaches and achieves state-of-the-art results.

The remainder of this paper is organized as follows. Section 2 gives a brief review of related work. Section 3 formally defines the task and describes our proposed MTVAF model. In Sect. 4, we compare and discuss the experimental results to verify the effectiveness of the proposed models. Finally, Sect. 5 concludes the paper and outlines future work.

2 Related work

2.1 Unimodal aspect-based sentiment analysis

Unimodal aspect-based sentiment analysis involves classifying the sentiment polarity of aspect terms extracted within a single modality, typically text (Chen and Qian 2019; Chen

et al. 2020). Early studies primarily concentrated on two independent subtasks: Aspect Term Extraction (ATE) and Aspect Sentiment Classification (ASC). ATE was initially formulated as a sequence labeling task, with approaches emphasizing representation learning methods to enhance word embeddings and employ Conditional Random Field (CRF) models (Luo et al. 2019). Advancements in deep learning led to the increased popularity of neural networks like Convolutional Neural Networks (CNN) (Xue and Li 2018), Recurrent Neural Networks (RNN) (Ding et al. 2017), and Recursive Neural Networks (RecNN) (Wang and Pan 2020) in ATE research.

ASC tasks are typically categorized based on the aspect type: aspect term sentiment classification and aspect category sentiment classification. Many existing approaches address these problems by designing models that leverage attention mechanisms to capture positional information and obtain aspect-specific representations for sentiment detection (Tang et al. 2016). These models effectively integrate the relationships between aspects (terms/categories) and sentence contexts. Additionally, researchers have explored methods based on Graph Neural Network (GNN) to explicitly leverage the syntactic structural relations between aspects and corresponding opinions for sentiment polarity detection (Sun et al. 2019).

Recognizing that aspects provide valuable cues for sentiment classification and polarity detection, the joint ABSA task, which extracts both aspect terms and their corresponding sentiment polarities, has gained substantial attention (Chen et al. 2020). Recent ABSA research has increasingly focused on extracting more fine-grained information, encompassing aspect sentiment triplet extraction, aspect category sentiment detection, and aspect sentiment quad prediction. While unimodal ABSA has made significant progress, it overlooks a key fact: real-world sentiments are often expressed through integrating information from multiple modalities. This includes not just words, but also visual cues.

2.2 Multimodal aspect-based sentiment analysis

In addition to the coarse-grained multimodal sentiment analysis conducted at the sentence level (Zadeh et al. 2017; Yu et al. 2020; Gandhi et al. 2023), the MABSA task aims to extract more detailed information from text-image pairs. This includes tasks such as multimodal aspect term extraction (MATE), multimodal aspect-sentiment classification (MASC), and joint multimodal aspect-sentiment analysis (JMASA). Regarding the MATE task, similar to information extraction, sequential annotation methods such as CRF (Wang et al. 2022) and GNN (Zhang et al. 2021) are employed for entity extraction. In contrast, the MASC task is typically approached as a sequence classification problem, and various neural network-based models, including Bidirectional Long Short-Term Memory (BiLSTM) (Zhou et al. 2021), BERT (Khan and Fu 2021), have been proposed.

Recent MABSA research has been focused on acquiring crucial visual features to enhance the semantic representation of entities in complex scenes. Yu et al. (2022) introduce a hierarchical interaction module with assisted image reconstruction. However, this module primarily emphasizes local interactions and overlooks global semantic relationships and modality heterogeneity. Another approach, pioneered by Khan and Fu (2021), involves converting images to textual descriptions to avoid bias arising from cross-modal interactions. Despite aiding cross-modal alignment, this approach often results in neutral descriptions of images, introducing noise when learning the diverse emotional cues in the visual modality. Consequently, Yang et al. (2022) employ facial emotions as a supervised signal for learning visual emotions. Nonetheless, this method overlooks scenarios where

facial expressions are absent in images, making it challenging to capture emotional visual cues.

The JMASA task proposed by Ju et al. (2021) aims to simultaneously address both sub-tasks: extracting aspect terms and classifying their corresponding sentiments while also considering text-image relation detection through annotated datasets. Nonetheless, this approach primarily emphasizes cross-modal global interactions and does not extensively examine fine-grained aspects and sentiments. VLP-MABSA (Ling et al. 2022) simulates text-based ABSA and designs specific pre-trained tasks for images to achieve cross-modal alignment. However, this approach demands significant computational resources and relies on extensive pre-trained labeled data. To tackle this challenge, CMMT (Yang et al. 2022) introduces multi-aspect and sentiment detection tasks for cross-modal interaction learning, using a unified label. However, few studies have delved into bridging the semantic gap between modalities and efficiently harnessing visual information for coarse-grained to fine-grained alignment. Different from them, the goal of our model is to effectively translate different granularities of image information into the textual space, thereby achieving accurate alignment between images and text.

2.3 Pretrained vision-language models

Inspired by the success of foundational pre-trained language models like BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), BART (Yan et al. 2021), and GPT (Radford et al. 2019) across various domains in natural language processing, Vision-Language models have gained prominence. Pre-trained vision-language models have been trained on large-scale image-text pair datasets. These models have shown impressive generalization abilities in various vision-language tasks. For example, in image-text retrieval, visual question answering, and image captioning (Tu et al. 2021, 2022), pre-trained models have achieved state-of-the-art performance.

Early Vision-Language Pretraining (VLP) approaches, such as Oscar (Li et al. 2020) and Uniter (Chen et al. 2020), heavily relied on pre-trained Object Detectors for visual feature extraction during pretraining. This reliance resulted in limited generalization capabilities and a strong dependence on prior knowledge. To address these issues, CLIP (Radford et al. 2021), ALBEF (Li et al. 2021), and others proposed incorporating the Vision Transformer architecture in VLP (Li et al. 2022; Zhan et al. 2023), enabling the learning of more abstract and versatile visual representations. More recently, models like BLIP (Li et al. 2022) and BEiT-3 (Wang et al. 2022) introduced Transformer-based encoder-decoder architectures that almost unified all aspects of visual-language understanding and generation tasks. In this paper, we utilize the CLIP for image captioning tasks, translating global image information into the textual feature space.

Vision-Language models typically adopt coarse-grained pre-training tasks like Masked Language Modeling (MLM), Masked Region Classification (MRC), and Image-Text Matching (ITM). These tasks aim to equally understand and learn from both images and text. However, in the case of MABSA tasks, the focus shifts to prioritizing using aligned visual information to enhance text representations, especially for aspects and emotions. The VAL (Chen et al. 2020) and VLP-MABSA (Ling et al. 2022) respectively align visual and language information at the levels of multi-grained and fine-grained to obtain expressive representations. Considering the limited data resources for MABSA tasks, in order to effectively utilize alignment information, our work proposes a method of integrating multi-scale image information as prompts into text representations.

3 Methodology

Task definition: Following previous works (Ju et al. 2021; Yang et al. 2022), we formulate the JMASA task as a sequence labeling problem with a unified tagging scheme. Formally, considering a sentence $S = (s_1, s_2, \dots, s_n)$ comprising n words, along with its corresponding global image G . The objective of JMASA is to identify the aspect terms within the sentence S and their corresponding sentiment polarities. In particular, we aim to derive a joint label sequence $y = (y_1, y_2, \dots, y_n)$, where $y_i \in \{\text{B-POS}, \text{B-NEU}, \text{B-NEG}, \text{I-POS}, \text{I-NEU}, \text{I-NEG}, \text{O}\}$. The labels B, I, and O signify the beginning, inside and out of an aspect, respectively. POS, NEU, and NEG represent positive, neutral, and negative sentiments towards the aspect. For example, when the word s_i is tagged with the label B-POS, it indicates that the word represents the beginning of an aspect with a positive sentiment.

3.1 Model overview

As mentioned before, two challenges of MABSA are to extract effective emotional visual features at various levels of granularity and bridge the gap between different modalities. To tackle these challenges, we propose a multi-level textual-visual alignment and fusion network. Figure 2 provides an overview of the proposed framework. Our approach leverages hierarchical visual information from ResNet (He et al. 2016) and transforms it into the visual prompt, facilitating the seamless integration of richer visual information into both the **T** input space and **T+V** input space.

As shown in Fig. 2, the proposed MTVAF consists of the following three main components:

1. Multi-granularity visual translation alignment: This component aims to align textual and visual spaces by translating the image into visual context.
2. Multi-scale visual aspect-opinion fusion: It is designed to project multi-scale visual features into the same low-dimensional space and dynamically fuse them into each layer of the Transformer. By incorporating visual aspect-opinion supervision, it enables the acquisition of fine-grained visual information throughout the model.
3. Text-centered multimodal training: To enhance robustness, this component minimizes the KL divergence over the output distributions of two inputs, effectively denoising the visual modality. Furthermore, it employs CRF for generating JMASA output.

3.2 Multi-granularity visual translation alignment

Despite recent multimodal research has explored different ways to align textual and visual spaces (Yang et al. 2022; Ling et al. 2022), these studies have consistently neglected the need for precise alignment, particularly in the ranging from coarse to fine-grained levels. This oversight might hinder the recognition of visual clues that could enhance textual representations.

To address this issue, we propose a multi-granularity visual translation alignment module, as shown in Fig. 3, which transforms images into visual context inputs at different granularity levels. These inputs are combined with the **T** input and subsequently fed into a stacked bidirectional Transformer with multi-layer attention modules. This method

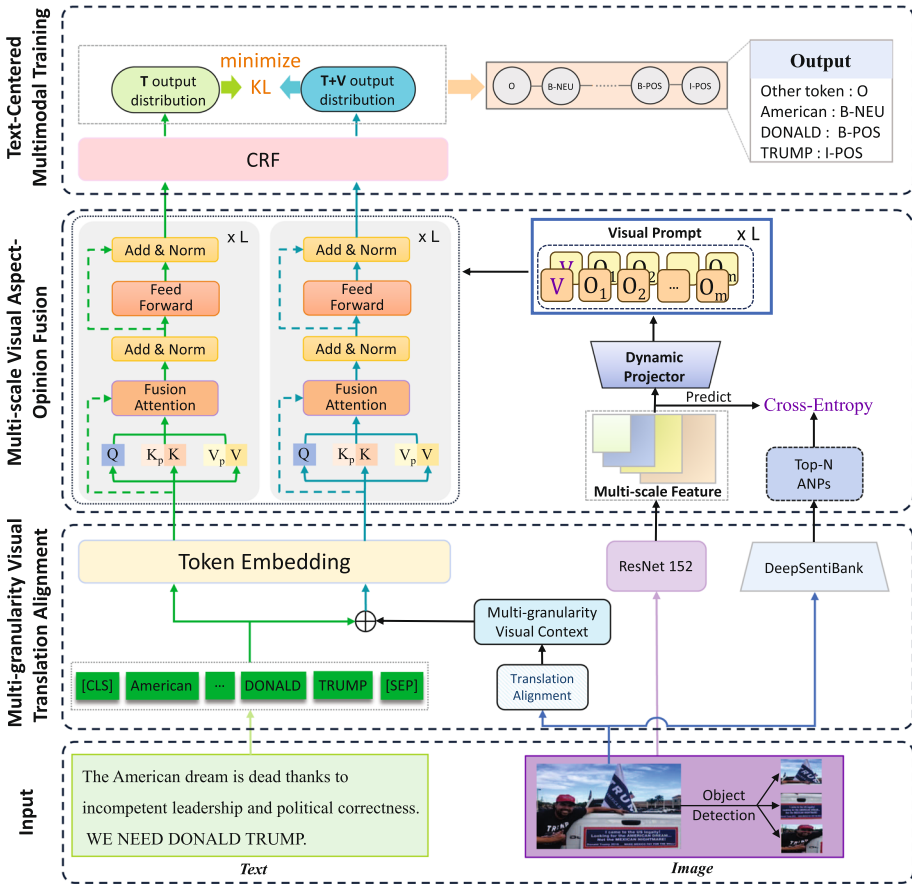


Fig. 2 The overall architecture of our proposed MTVAF

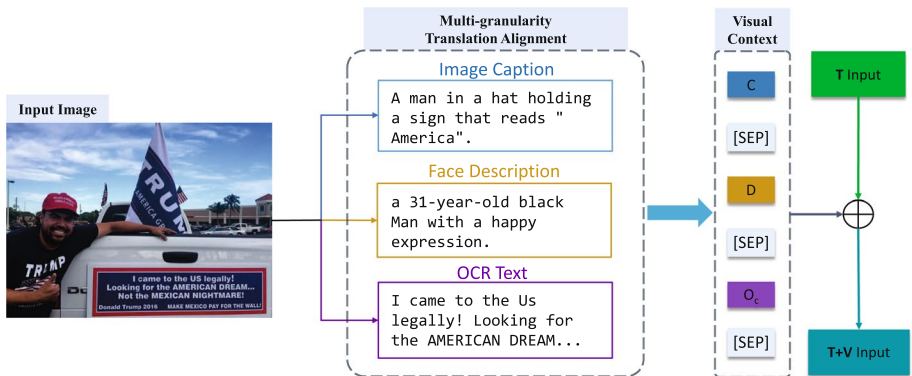


Fig. 3 The workflow of multi-granularity visual translation alignment

indirectly accomplishes multi-granularity alignment between images and text, covering three levels of granularity alignment: global, local, and character-level.

3.2.1 Global coarse-grained alignment

Our primary objective is to establish a coarse-grained alignment between global images and text. We use an image captioning model to create a comprehensive connection between visual and textual modalities, reducing the impact of irrelevant images. This process strives to generate meaningful and accurate image descriptions that convey the semantic information of visual content at a coarse-grained level. In particular, we apply an image captioning tool *ClipCap* (Mokady et al. 2021), which generates the high-quality caption for scenes, denoted as C :

$$C = \text{Caption}(G) \quad (1)$$

where C denotes the comprehensive description of the image generated from *Caption*, serving as a coarse-grained text-aligned mapping of the entire image.

3.2.2 Local fine-grained alignment

Building upon the global coarse-grained alignment, local fine-grained alignment focuses on translating local facial features into the textual space to obtain aligned descriptions of faces. Facial expressions, being a direct means for humans to convey emotions, are invaluable for the exact identification of emotions at the object level in images (Fan et al. 2018). Our observations of the Twitter-2017 dataset reveal that facial expressions appear in over half of the images within tweets.

Hence, for the extraction of fine-grained emotional visual information, we initiate the process by utilizing the LightFace¹ face detector to identify all faces and transform them into textual facial attributes. Following this, we adopt the facial expression description template proposed by Yang et al. (2022) to generate face descriptions:

$$D = \text{Face_Description}(G) \quad (2)$$

where $D = \{D_1, D_2, \dots, D_d\}$, and d represents the number of facial descriptions.

After obtaining the facial attributes in textual form, we sort them in descending order based on the prediction confidence of the face detector, allowing us to filter out attributes with low prediction confidence. This step is essential for focusing on emotionally relevant information from local regions in images and achieving fine-grained alignment between different modalities.

3.2.3 Optical Character-grained Alignment

In addition, images that include text offer valuable semantic information that enhances the visual content (Wang et al. 2022; Yao et al. 2023). Conventional image encoders often face challenges in comprehending this information such as slogans in advertisements, text within emojis, famous quotes on posters, and so on. Thus, we apply Google's

¹ <https://github.com/serengil/lightface>

Tesseract OCR engine², an advanced lightweight open-source OCR system to extract text from images. Through accurate text identification and extraction from images, our method achieves character-level alignment, greatly improving its sensitivity to the emotional information conveyed by these images.

$$O_c = OCR(G) \quad (3)$$

where O_c represents the concatenated sequence of English words extracted by the OCR model.

To mitigate the noise caused by irrelevant images, we concatenate O_c with C and D to generate a textual-visual alignment of visual context called $V_c = (C, [SEP], D, [SEP], O_c, [SEP])$. At this stage, visual information is mapped into the text space, and after concatenating it with the \mathbf{T} input, a $\mathbf{T}+\mathbf{V}$ input is formed. Just like before, we insert the $[SEP]$ token between the \mathbf{T} input (S) and the visual context (V_c). Both $\mathbf{T}+\mathbf{V}$ and \mathbf{T} pass through a Transformer-based model to acquire the final hidden representations H_{T+V}^L and H_T^L , which are fed into the CRF layer. For a label sequence $y = (y_1, y_2, \dots, y_n)$, we define the probability of the tag sequence y based on the hidden representations H^L as follows:

$$s(H^L, y) = \sum_{j=0}^n M_{y_j, y_{j+1}} + \sum_{j=1}^n P_{j, y_j} \quad (4)$$

$$p(y|H^L) = \text{Softmax}(s(H^L, y))$$

where $M_{y_j, y_{j+1}}$ is the randomly initialized transition matrix from label y_j to y_{j+1} , P_{j, y_j} denotes the emission matrix of label y_j linearly transformed from the H_L .

3.3 Multi-scale visual aspect-opinion fusion

In this module, three primary subtasks are addressed. Firstly, the multi-scale visual feature subtask is dedicated to converting the input image into visual features at various scales. Secondly, the top-N visual aspect-opinion subtask aims to acquire detailed aspect-opinion information from visual data. To achieve this goal, it employs Adjective-Noun Pairs (ANPs) (Borth et al. 2013) and predicts their top-N probabilities, which serve as supervision signals. Finally, the prompt-based dynamic visual fusion subtask primarily focuses on dynamically integrating multi-scale visual cues, acting as key-value prompt information in the multi-layer bidirectional Transformer (BERT) (Devlin et al. 2019) for two inputs: \mathbf{T} input and $\mathbf{T}+\mathbf{V}$ input.

3.3.1 Multi-scale visual feature

Recent research (Tian et al. 2023) has shown the ability of convolutional neural networks (CNN) to hierarchically extract target features. It's been shown that different layers of these networks, both shallow and deep, possess distinct receptive fields suitable for processing objects of various sizes. This is particularly important when dealing with various-grained images. Our module's objective is to capture multi-scale visual features and obtain corresponding hierarchical visual representations.

² <https://github.com/madmaze/pytesseract>

To accomplish this, we incorporate global and regional images as supplementary visual information. Global images help capture large-scale abstract concepts such as entity context and overall emotional clues. On the other hand, regional images act as vital visual cues for smaller-scale details, guiding visual feature learning. By combining semantic and spatial information from deep and shallow features, we obtain multi-scale visual features. Specifically, we utilize a four-block structured ResNet (He et al. 2016) as the visual encoder and YOLOv5x6³ as the object detector. We retain at most z regions $O_b = (O_1, O_2, \dots, O_z)$ with the highest confidence scores.

The multi-scale image inputs are fed into the visual encoder, where deep information is upsampled and element-wise added with shallow information. This process extracts multi-scale feature maps $F = (F_1, F_2, \dots, F_r)$, which are then fused. Following this, an average pooling operation is executed to enhance the recognition capability of visual aspects within the image:

$$\begin{aligned} [F_1, F_2, \dots, F_r]_G; [F_1, F_2, \dots, F_r]_{O_b} &= \text{Visual_Encoder}([G]; [O_b]) \\ \hat{F}_i &= \text{Ave}(F_i) \end{aligned} \quad (5)$$

where $[F_1, F_2, \dots, F_r]_G$ and $[F_1, F_2, \dots, F_r]_{O_b}$ represent the visual features obtained from the fusion of multi-scale feature maps, including global image features and object features. *Ave* denotes the average pooling layer, which transforms F_i into same dimension.

3.3.2 Top-N visual aspect-opinion

When integrating a significant amount of fine-grained visual information into two inputs, it's clear that exploring the fine-grained relationships of the visual features obtained from the multi-scale network is essential. Inspired by VLP-MABSA (Ling et al. 2022), we utilize Adjective-Noun Pairs (ANPs) as supervision for visual aspects and opinions. These ANPs are derived from the pre-trained ANP detector, DeepSentiBank⁴ (Chen et al. 2014), which predicts the class distribution of 2089 ANPs within the entire image, reflecting visual aspect-opinion information.

Since depending just on the predicted ANP causes an error propagation issue and utilizing the entire distribution for supervision creates redundant noise, we propose using adjective-noun pairs with top-N prediction probabilities for the guided model. For instance, in the middle right part of Fig. 2, the top-N adjective-noun pairs from the input image contain an ordered list relevant to the visual aspect-opinion, potentially guiding our model's focus on fine-grained visual information. The distribution of top-N prediction P is computed as:

$$P = \text{Softmax}(W^T (\frac{1}{r} \sum_{i=1}^r (\hat{F}_i)) + b) \quad (6)$$

with $r=4$, where $W \in \mathbb{R}^{d \times N}$ and $b \in \mathbb{R}^N$ represent trainable parameters, d represents the dimension of the text representation in BERT.

To bring the predicted distribution P closer to the ground-truth top-N adjective-noun pairs distribution A , we employ the standard Cross-Entropy loss to find fine-grained information from image input:

³ <https://docs.ultralytics.com/yolov5/>

⁴ <https://github.com/stephen-pillli/DeepSentiBank>

$$L_V = -A \log(P) \quad (7)$$

This loss function is designed to minimize the discrepancy between the predicted and ground-truth distributions, thereby enhancing the model's capability to capture the needed visual aspect-opinion relationship.

3.3.3 Prompt-based dynamic visual fusion

In multimodal aspect-based sentiment analysis, the textual modality remains crucial for identifying entities and sentiments, even as the visual modality plays a significant role (Zhan et al. 2023). Therefore, we propose a submodule that employs dynamic attention mechanism (Chen et al. 2018) to project multi-level visual information, including not only full image details but also object-level information, as prompts to the l -th layer of BERT within the textual modality. These visual prompts concatenate keys and values in each layer during multi-head attention calculations, mitigating noise interference from irrelevant visual information. This dynamic projector calculates multiple normalized vectors that control the extent of visual feature transformation for each BERT block. Firstly, we calculate the logits α_i^l for the projecting signal:

$$\begin{aligned} e_i &= MLP(\hat{F}_i) \\ \alpha_i^l &= \frac{\exp(e_i)}{\sum_{k=1}^r \exp(e_k)} \end{aligned} \quad (8)$$

where MLP represents a layer that appropriately reduces the feature dimensionality.

As for integrating textual and visual features, we employ multi-head self-attention. This process combines the visual prompts with the key/value vectors of contextual representations in each BERT layer. Here, V^l represents the transformed visual features, which are fed into the l -th layer of BERT.

$$\begin{aligned} V^l &= [V_G^l; V_O^l] = \sum_{i=1}^r (\alpha_i^l \cdot \hat{F}_i) \\ [\delta_k^l; \delta_v^l] &= W_\delta^l V^l \end{aligned} \quad (9)$$

where visual prompts $\delta_k^l, \delta_v^l \in \mathbb{R}^{(z+1)hw \times d}$, $(z+1)hw$ denotes the length of visual features. The multi-scale visual features undergo a linear transformation $W_\delta^l \in \mathbb{R}^{d \times 2 \times d}$ projects them into the same embedding space as the textual representations.

The equally-length visual prompts are then concatenated with the origin key and value vector from the previous BERT layer, which acts as the new key and value during the attention process. Formally, the fusion of visual prompts with text-based attention is calculated as follows:

$$Fusion_Attention = \frac{\text{Softmax}(W_Q^l H^{l-1} \cdot [\delta_k^l; W_K^l H^{l-1}])}{\sqrt{d}} [\delta_v^l; W_V^l H^{l-1}] \quad (10)$$

where $W_Q^l H^{l-1}$, $[\delta_k^l; W_K^l H^{l-1}]$ and $[\delta_v^l; W_V^l H^{l-1}]$ represent the query, key, and value in the new attention matrices.

3.4 Text-centered multimodal training

Given the diverse and multi-level visual information, when input to multi-layer bidirectional Transformers, there's a risk of excessive attention toward lengthy visual context. This can overshadow the primary role of textual information during gradient loss back-propagation. Furthermore, the absence of annotated labels for supervising the alignment between textual and multimodal information poses a challenge. Thus, we introduce minimizing the KL-divergence over two probability distributions, which is obtained from feeding two inputs into the Transformer-based model in Eq. 10. It is equivalent to calculating the cross-entropy loss between the two distributions:

$$L_{T+V} = KL(p(y|H_{T+V}^L) || p(y|H_T^L)) = \sum_{y \in Y} p(y|H_{T+V}^L) \log(p(y|H_T^L)) \quad (11)$$

where $p(y|H_{T+V}^L)$ and $p(y|H_T^L)$ are **T+V** and **T** probability distributions derived from two inputs through Eq. 4.

As **T+V** information introduces noise in the aspect-sentiment pairs process of the MABSA model, we adopt a text-centered approach to transfer crucial information from the **T+V** context. Thereby, only $p(y|H_T^L)$ is backpropagated. This loss function is the negative log probability of the ground truth label sequence as follows:

$$L_T = - \sum_{i=1}^n \log(p(y|H_T^L)) \quad (12)$$

The final combined objective function is defined as follows:

$$L_{MTVAF} = \lambda \cdot L_T + \mu \cdot L_V + \gamma \cdot L_{T+V} \quad (13)$$

where λ , μ , and $\gamma \in [0, 1]$ are trade-off hyper-parameters to control the contribution of each module.

4 Experiments

4.1 Experimental settings

Datasets: We demonstrate the effectiveness of our approaches on the Twitter-2015 and Twitter-2017 datasets (Yu et al. 2019) for MABSA. These datasets comprise text-image pairs extracted from tweets spanning the years 2014 to 2017. The statistics of these two datasets are presented in Table 1. And the detailed statistics of the multi-granularity visual translation alignment are shown in Table 2.

Evaluation metrics: Following previous works (Ju et al. 2021; Ling et al. 2022; Yang et al. 2022; Yang et al. 2023), we adopted Precision (P), Recall (R), and F1 score as the evaluation metrics to assess the performance of different methods in the MABSA task.

$$Precision = \frac{\#true}{\#prediction} \quad (14)$$

Table 1 Statistics of two benchmark datasets (All: number of all aspects, #S: number of sentences, Mean: mean length of sentences, Max: maximum length of sentences)

Datasets		Postive	Neutral	Negative	All	#S	Mean	Max
Twitter-2015	Train	928	1883	368	3179	2101	15	35
	Dev	303	670	149	1122	727	16	40
	Test	317	607	113	1037	674	16	37
Twitter-2017	Train	1508	1638	416	3562	1746	15	39
	Dev	515	517	144	1176	577	16	31
	Test	493	573	168	1234	587	15	38

Table 2 A statistic about the number of sentences with different-granularity visual contexts and their mean and maximum length

	Twitter-2015	Twitter-2017
Num of image caption / Total Sentences	3502 / 3502 (100%)	2910 / 2910 (100%)
Mean / Max of image caption	51.0 / 105	51.5 / 160
Num of face description / Total Sentences	1205 / 3502 (34.41%)	1646 / 2910 (56.56%)
Mean / Max of face description	63.0 / 163	72.1 / 162
Num of OCR text / Total Sentences	861 / 3502 (24.59%)	786 / 2910 (27.01%)
Mean / Max of OCR text	61.2 / 100	59.3 / 100

We translate from the image by the multi-granularity alignment methods introduced in Sect. 3.2

$$Recall = \frac{\#true}{\#ground\ truth} \quad (15)$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (16)$$

where $\#prediction$ and $\#ground\ truth$ denote the number of predicted and ground truth aspect-sentiment pairs, respectively. The amount of correct predictions in aspect-sentiment pairs is represented by $\#true$, implying that both the aspect boundary and sentiment classification are correct.

Implementation details: To ensure fair comparisons, we employed BERT-base-uncased⁵ as our textual backbone and ResNet152 as the visual encoder, consistent with recent studies. The Transformer model had 12 attention heads and a dropout rate of 0.1. Training lasted for 30 epochs, with evaluation after the 16th epoch. In optimization, we used the AdamW optimizer with a weight decay of 0.01. Additionally, the learning rate was linearly warmed up to its maximum value during the first 1% of training steps. For the Twitter-2015 dataset, we used a learning rate of $2e-5$ and a batch size of 16. For the Twitter-2017 dataset, a learning rate of $1.5e-5$ and a batch size of 4 were employed. The length of the prompt was set to 4, and its dimensionality was reduced to 800. The number of image objects was

⁵ <https://huggingface.co/bert-base-uncased>

limited to no more than 3, and the OCR text length was limited to 100 characters. Besides, we consider the top- N visual aspect-opinion as its fine-grained concepts, i.e., N set to 10. The impact of tradeoff hyperparameters λ and μ impact on performance is not particularly sensitive, as both are set to 1 in our model. Conversely, another hyperparameter γ demonstrates high sensitivity to performance. In subsection 4.5.2, we via a grid-search strategy determine the optimal trade-off values for γ , resulting in values of 0.3 and 0.2 for the Twitter-2015 and Twitter-2017 datasets, respectively. We implemented all our methods using PyTorch and executed them on a single NVIDIA Tesla V100 GPU.

4.2 Baselines

In this subsection, we conduct a comprehensive comparison by comparing our MTVAF model with two groups of competitive models. Specifically, we consider the unimodal and multimodal baselines.

Text-based methods:

1. **SpanABSA** (Hu et al. 2019) is a span-based hierarchical method for textual ABSA.
2. **D-GCN** (Chen et al. 2020) proposes the concept of second-order proximity information, which is used to extend the convolution operation receptive field to extract more features on the directed graph.
3. **GPT-2** (Radford et al. 2019) adopts a Transformer-based architecture using only a decoder structure, enabling end-to-end applications in textual ABSA through text generation.
4. **RoBERTa** (Liu et al. 2019) is an advanced pre-trained transformer-based model, which feeds the contextualized text representation into a CRF layer for sequence labeling.
5. **BART** (Yan et al. 2021) is an Encoder-Decoder Transformer architecture that combines contextual information and autoregressive features, formulating textual ABSA as an index generation task.

Multimodal methods:

1. **UMT-collapsed** (Yu et al. 2020), **OSCGA-collapsed** (Wu et al. 2020) and **RpBERT-collapse** (Sun et al. 2021) are model the multimodal named entity recognition (MNER) task originally, replaced with collapsed tagging for MABSA task. Note that **UMT-collapsed** uses the cross-modal Transformer to model the interaction between text and images, and **OSCGA-collapsed** combines object-level visual information with textual information. The **RpBERT-collapsed** uses the confidence of the image-text relationship to fuse the two modalities.
2. **CLIP** (Radford et al. 2021) employs contrastive pretraining to encode rich semantic representations both images and text, which can be applied in MABSA.
3. **JML** (Ju et al. 2021) performs a span-based hierarchical joint learning approach while introducing an auxiliary cross-model relationship detection task to integrate appropriate visual information.
4. **CMMT** (Yang et al. 2022) uses a gating mechanism to control the contribution of text and image representations and to capture the interaction between them, with two unimodal auxiliary tasks.
5. **VLP-MABSA** (Ling et al. 2022) designs multiple distinct vision-language pre-training tasks on an extra pre-labeled dataset containing over 17,500 image-text pairs. This

Table 3 A comparison of our MTVAF model and other competitive baselines for MABSA

Modality	Approaches	Twitter-2015			Twitter-2017		
		P	R	F1	P	R	F1
Text	SpanABSA	53.7	53.9	53.8	59.6	61.7	60.6
	D-GCN	58.3	58.8	59.4	64.2	64.1	64.1
	GPT-2	66.6	60.9	63.6	55.3	59.6	57.4
	RoBERTa	62.4	64.5	63.4	65.3	66.6	65.9
	BART	62.9	65.0	63.9	65.2	65.6	65.4
Text-Image	UMT-collapse	60.4	61.6	61.0	60.0	61.7	60.8
	OSCGA-collapse	63.1	63.7	63.2	63.5	63.5	63.5
	RpBERT-collapse	49.3	46.9	48.0	57.0	55.4	56.2
	CLIP	44.9	47.1	45.9	51.8	54.2	53.0
	JML	65.0	63.2	64.1	65.8	65.2	65.5
	CMMT	63.5	66.6	65.0	66.1	69.0	67.5
	VLP-MABSA	65.7	69.0	67.3	67.4	68.2	67.8
	GMP	65.5	68.8	67.1	66.8	68.0	67.4
	AoM	67.5	69.0	68.2	67.6	67.0	67.3
	MTVAF(Ours)	69.3	72.8	71.0	68.1	68.3	68.2

The best scores on each metric are indicated in bold

approach aims to bridge the gap between a fine-grained MABSA task with limited resources and a general pre-training task.

6. **GMP** (Yang et al. 2023) utilizes multimodal encoders and decoders to automatically generate aspect-oriented and sentiment-oriented prompts for MABSA in text-image few-shot scenarios.
7. **AoM** (Zhou et al. 2023) reduces inter-modal noise for fine-grained sentiment analysis by jointly modeling aspect level semantics and guided sentiment aggregation.

4.3 Main results

In Table 3, we compare our approach with the baselines on the Twitter-2015 and Twitter-2017 datasets to demonstrate superior performance, and the following observations can be made.

Incorporating additional visual content enhances the model's understanding of the correlation between aspects and sentiments, resulting in promoting performance in the MABSA task. Multimodal models that leverage visual information clearly outperform text-only methods in MABSA. A notable comparison can be drawn by contrasting the unimodal baseline BART with its corresponding multimodal method VLP-MABSA, both based on pre-training. The latter outperforms the former, with F1 scores increased by 3.4% and 2.4% on the two datasets, respectively. This finding further emphasizes the valuable role and significance of visual information in aspect-level sentiment analysis in a multimodal setting.

The representation capacity of the model can be improved by incorporating relevant auxiliary tasks to filter out irrelevant image and text information, like multimodal relation detection and late-stage weighted fusion. Among the multimodal baseline models, UMT-collapse and OSCGA-collapse perform poorly because they do not consider the relevance

Table 4 Ablation study of the MTVAF

Approaches	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
MTVAF (Full)	69.3	72.8	71.0	68.1	68.3	68.2
only global coarse-grained alignment	68.5	71.4	69.9	66.3	67.2	66.7
only local fine-grained alignment	67.2	71.2	69.1	66.5	67.6	67.0
only optical character-grained alignment	66.2	70.1	68.5	65.9	66.4	66.1
w/o global coarse-grained alignment	67.6	71.4	69.4	66.2	68.2	67.2
w/o local fine-grained alignment	68.0	72.4	70.1	66.4	68.4	67.4
w/o optical character-grained alignment	69.3	72.1	70.7	67.8	68.8	68.3
w/o multi-granularity visual translation alignment	65.9	68.8	67.3	64.8	66.5	65.6
w/o top-N visual aspect-opinion	66.9	70.7	68.8	66.0	66.7	66.3
w/o prompt-based dynamic visual fusion	66.0	69.4	67.7	63.6	64.2	63.9
w/o multi-scale visual aspect-opinion fusion	65.2	64.3	64.7	67.8	53.7	60.0
rep. text-centered multimodal training	67.5	70.9	69.1	67.1	67.8	67.5

to downstream tasks and simply fuse visual features without alignment. Direct fusing of image and text features into RpBERT-collapse through naive fusion results in remarkably inferior performance. On the other hand, JML and CMMT design auxiliary tasks to explore coarse-grained or fine-grained alignment information in visual content, yielding better results. However, they do not fully exploit the diverse range of semantic information present in images, resulting in the omission of crucial visual cues. In contrast, MTVAF leverages multi-level visual-textual alignment and visual representations as prompts. The table results show that our method effectively mitigates interference from irrelevant global and local image information, providing a more comprehensive approach for visual-language alignment and fusion in MABSA.

Multimodal pre-trained models generally require task-specific pre-training or prompt-based learning to provide supervised signals for model fine-tuning. Although CLIP can capture semantically powerful multimodal representations through contrastive learning, its performance is significantly lower than VLP-MABSA guided by proper pre-training tasks. In comparison, GMP may help model the correlations between input images, text, and output sentiments better by automatically generating customized multimodal prompts for MABSA. This also highlights the effectiveness of visual prompt-based fusion for inter-modality interaction in MTVAF.

As shown in Table 3, it is evident that the proposed MTVAF outperforms the state-of-the-art AoM model by 1.8%, 3.8%, and 2.8% in terms of Precision, Recall, and F1 score, respectively, on the Twitter-2015 dataset. This observation indicates that even selectively attending to aspect-relevant visual-textual contents and aggregating associated sentiment signals may still be insufficient to completely filter out misaligned visual noise. For the MABSA task, our approach leverages multi-level visual knowledge aligned with text, which proves to be more beneficial in training text-based models.

4.4 Ablation study

In Table 4, we conducted ablation experiments to further analyze the contributions of each module in our model. (1) only global/local/character-grained alignment: Retain only the alignment translation at the specified level. (2) w/o global/local/character alignment: Remove alignment translation at the specified level (various combinations of multi-granularity alignment). (3) w/o multi-granularity visual translation alignment: Completely remove $\mathbf{T}+\mathbf{V}$ input and KL divergence in Fig. 2 top left. (4) w/o top-N visual aspect-opinion: Remove the visual aspect-opinion loss in Eq. 7. (5) w/o prompt-based dynamic visual fusion: Exclude the fusion of image information via visual prompts, only use translated visual contexts as visual modality input. (6) rep. text-centered multimodal training: Replace text probability distribution with $\mathbf{T}+\mathbf{V}$ distribution in Eq. 12 for backpropagation, eliminating the two-distribution loss in Eq. 11.

We conducted a series of experiments to validate the vital role played by various granularity alignment methods and their combinations. The performance drop of *only global coarse-grained alignment* on Twitter-2015 is less significant than on Twitter-2017. This difference could be due to the dominance of neutral samples in Twitter-2015, as shown in Table 1 and Table 2, where face descriptions, accounting for sentiment clues, only make up 34.41%. During training, the model may tend to focus more on objective global coarse-grained alignment in such cases. Meanwhile, on Twitter-2017, F1 of *w/o local fine-grained alignment* is slightly higher than the complete MTVAF model. There could be two reasons for this result. First, OCR has weaker recognition capability on low-quality or complex images, which may introduce irrelevant information during multi-granularity visual translation alignment. Second, a great deal of OCR text in the Twitter datasets exhibit little semantic association with linked sentences. OCR text not only fail to enhance visual alignment, but they also introduce noise. Subsequently, we removed all visual context (*w/o multi-granularity visual translation alignment*). Performance degradation was observed on both datasets, with a larger decrease of 3.7% and 2.6% when the visual context was removed. This reveals that translating visual information from coarse-grained to fine-grained levels into the textual space can bridge the semantic gap between the image and text modalities.

To demonstrate the practicality of multi-scaled visual fusion, we conducted experiments involving the removal of two components: top-N adjective-noun pairs (*w/o top-N visual aspect-opinion*), and multi-scaled fusion (*w/o prompt-based dynamic visual fusion*) both as ablative models. When we eliminated auxiliary supervision, there was a decrease of 2.2% and 1.9% in F1 scores on the two Twitter datasets, indicating that the predictions of ANPs may capture the visual aspect-opinion information. Furthermore, when we excluded multi-scale visual cues, there was a significant decline in performance. This emphasizes the importance of our proposed MTVAF, which facilitates comprehensive and informative interactions within the textual model by integrating visual cues of different scales.

Besides, the decline observed in both datasets by removing the loss of the two distributions in Eq. 11 (*w/o text-centered multimodal training*), as it effectively reduces noise introduced by redundant visual context.

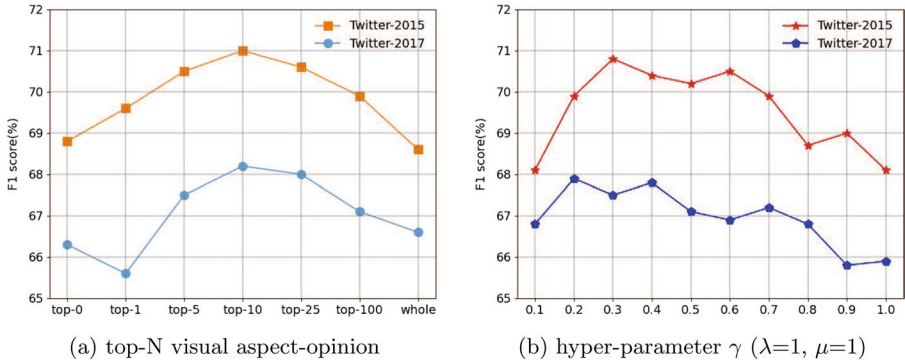


Fig. 4 The result of different top-n adjective-noun pairs and contribution of multi-granularity visual translation alignment module for MTVAF

4.5 In-depth analysis

We perform the in-depth analysis on two Twitter datasets to investigate the effect of the number of adjective-noun pairs, hyper-parameter settings, and layers of the Image Encoder on the performance, which further demonstrates the validity of the proposed MTVAF model.

4.5.1 Analysis of the sensitivity of the top-N visual aspect-opinion

At the top-N adjective-noun pairs stage, we investigated the influence of different settings for N. We explored values of 0, 1, 5, 10, 25, 100, as well as the entire distribution (2089) to understand their impact. Our tests revealed the sensitivity of the number of adjective-noun pairs. Figure 4(a) manifests that using adjective-noun pairs generally enhances the performance of MABSA compared to the top-0 (w/o top-N visual aspect-opinion).

Moreover, using only the adjective-noun pair with the highest predicted probability ($N = 1$) may lead to error propagation, while employing the distribution data of a large number of adjective-noun pairs (whole distribution) may introduce noise. Through our observations, setting the number of adjective-noun pairs to 10 yielded the best result. This finding highlights that adopting an appropriate number of adjective-noun pairs can effectively alleviate error propagation and visual noise problems to a certain extent, consequently leading to improved performance.

4.5.2 Analysis of the sensitivity of the hyper-parameter

We conduct a hyperparameter experiment of MTVAF to effectively utilize aligned image information through translating input image into three granularity, as depicted in Fig. 4(b), based on the final performance on the development set. For the hyperparameter γ in Eq. 13, we experimented with different settings ranging from 0.1 to 1.0 in increments of 0.1. We found that the optimal values for γ were 0.3 and 0.2 for the two Twitter datasets, yielding the best performance.

As previously emphasized, reducing irrelevant image information is crucial in the context of the MABSA task. The tradeoff parameter γ inherently signifies the degree of

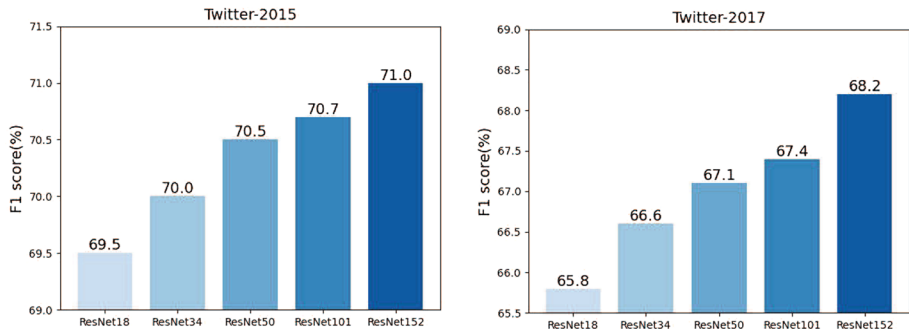


Fig. 5 Performance comparison of image encoder on twitter datasets

influence exerted by the aligned image content. When γ sets a larger value, it implies the introduction of redundant **T+V** input, potentially biasing the model training process and resulting in decreased robustness. On the contrary, selecting a smaller γ reduces the number of **T+V** inputs, negatively impacting overall performance. Therefore, to strike a balance between incorporating sufficiently aligned visual information and mitigating the impact of noise, it is advisable to assign a relatively smaller value to the contribution level of **T+V** input, optimizing the model's performance.

4.5.3 Analysis of the effectiveness of the image encoder




The importance of layers in the image encoder is emphasized in Fig. 5, we conducted experiments by replacing ResNet152 with alternative ResNet variants featuring varying numbers of layers. Notably, we observed a consistent decrease in F1 scores across the Twitter-2015 and Twitter-2017 datasets as the number of layers diminished. This observation underscores the importance of increasing the number of residual blocks within each module, as it enables the capture of more comprehensive bottom-up syntactic and semantic information within the images, despite the overall similarity in structure.

4.6 Case study

In this subsection, we selected three representative test examples in Table 5 for comparison among four models: the textual baseline BART, the multimodal benchmark model JML, VLP-MABSA, and our MTVAF framework.

As shown in Table 5, in example (a), we find that JML failed to identify the aspect term (Brandon Carr), while BART and VLP-MABSA additionally identified an aspect word *the* and BART also made an incorrect sentiment prediction. MTVAF might takes advantage of the image caption and relevant object-level images, which do not have specific emotional tendencies, aiding in entity recognition and sentiment prediction. In example (b), owing to the lack of image input, BART did not identify the aspect term (LFC). JML only extracted one aspect word *Steven*, and VLP-MABSA made an incorrect sentiment prediction. MTVAF, on the other hand, aligns the second entity, which combines the facial description of *angry expression* with adjective-noun pairs such as *tough face* and *extreme violence* to capture emotional visual contextual cues. As a result, MTVAF correctly extracts two aspect terms and classifies their sentiment as neutral and negative. In example (c), BART

Table 5 Three examples of the predictions by different methods

GroundTruth	(Brandon Carr, NEU), (Dallas Cowboys, NEU)	(#LFC, NEU), (Steven Gerrard, NEG)	(The Seth Leibsohn Show, NEU), (Steve Hayward, POS)
VisualModality			
TextualModality	(a) RT @ momandwife3: Brandon Carr of the Dallas Cowboys speaking to the kids at the sports camp at @ ElevateLC	(b) LFC can confirm that Steven Gerrard is to leave the club at the end of the 2014 - 15 season # LFCicon	(c) The Seth Leibsohn ShowTONIGHT: Attorney SheilaPolk on Legalizing # Marijuanaandamp Steve Hayward !
Top-NANPs	Young friends,fat body, young fashion, great street, cute shoes	Tough face, hot guy,dark skin, hot male,extreme violence	Happy guy, great smile, nice guy, attractive face, handsome face
VisualTranslation	Captions: a man standingin front of a brick wallholding a microphone. FD: None. OCR: None.	Captions: a man in a red jacket looking downat the camera. FD: The man show with aangry expression. OCR: None.	Captions: a poster of aman in a suit and tie. FD: The man show with alhappy expression. OCR: PRINCIPLES notPOLITICS The Seth LeibsohnShow Weeknights 9 11 PM.
BART	(Brandon Carr, NEU) ✓, (the Dallas Cowboys, NEG) X	- X, (Steven Gerrard, NEU) X	(Seth Leibsohn, NEU) X, (Steve Hayward, POS) ✓
JML	(Brandon Carr, NEU) ✓, - X	(LFC, POS) X, (Steven, NEG) X	(Seth Leibsohn Show, NEU) X, (Steve Hayward, POS) ✓
VLP- MABSA	(Brandon Carr, NEU) ✓, (the Dallas Cowboys, NEU) X	(LFC, NEU) ✓, (Steven Gerrard, NEU) X	(The Seth Leibsohn Show, POS) X, (Steve Hayward, POS) ✓
MTVAF	(Brandon Carr, NEU) ✓, (Dallas Cowboys, NEU) ✓	(LFC, NEU) ✓, (Steven Gerrard, NEG) ✓	(The Seth Leibsohn Show, NEU) ✓, (Steve Hayward, POS) ✓

Note that Top-N ANPs and Visual Translation denote the top-5 adjective-noun pairs predicted by DeepSentBank and the visual text translated by MTVAF, respectively. FD: face description, POS: positive, NEU: neutral, NEG: negative

Table 6 Agreement between every pair of three graduate annotators (G1, G2, G3)

	Graduate annotators		
	G1–G2	G1–G3	G2–G3
Cohen's kappa	0.59	0.68	0.65

Table 7 Performance on human-annotated datasets for MTVAF and its ablative variants

Dataset	MTVAF-unimodal	MTVAF-multimodal
supported (81%)	61.9	72.2
unsupported (19%)	62.4	66.7

MTVAF-Unimodal refers to the approach that utilizes text as input exclusively

Supported and unsupported refer to whether the image has a positive or potentially hindering impact on MABSA

and JML only extracted partial entities, with JML identifying an additional aspect word *Show* compared to BART. Although VLP-MABSA accurately predicted the correct aspect term (The Seth Leibsohn Show), it failed to classify sentiment, probably due to its equal treatment of images and text.

In the case of using images as an auxiliary modality, the MTVAF model not only correctly recognizes the aspect term (The Seth Leibsohn Show) through character-grained visual context, but also predicts a positive sentiment in combination with image information. These case results demonstrate that our MTVAF model can obtain all correct aspect terms and their associated sentiments by comprehensively aligning and fusing textual information with relevant images at different scales.

4.7 Comparison with human assessment

Is MTVAF consistent with human assessment? To gain a deeper understanding of the correlation and discrepancies between our approach and human perception of multimodal data, we engaged three graduate students (majoring in computer science and technology) to independently annotate each text along with its associated image. Their task was to assess whether the image enhances aspect recognition and sentiment detection in the text. To reduce semantic discrepancies and ensure dataset quality, we randomly sample three groups of 100 samples from the Twitter-2015 test set using different seeds (12, 43, 100). Cohen's Kappa (Cohen 1960) is adopted to measure inter-annotator agreement, and the highest average Kappa score among the three groups is presented in Table 6. This subset, demonstrating agreement at a fundamental level, was chosen as the evaluation dataset. The majority label among the three annotations was then adopted as the ground truth label.

We evaluated our model's performance in both unimodal and multimodal settings on the data subset, as shown in Table 7. When annotators considered the visual modality supportive for ABSA tasks, our model exhibited a significant improvement upon incorporating the visual modality, indicating its effective utilization of additional image information, in alignment with human judgment. However, in the evaluation of 19 unsupported samples, MTVAF showed marginal improvements of 4.3%. This observation underscores that even for human annotators, subject to inherent subjectivity and limited prior knowledge,

the assessment of image support in image-text pairs can vary, as discussed by Lake et al. (2017) and others.

Our model may mitigate the semantic gap between images and text during modality alignment and reduce interference from irrelevant images during fusion. Although there is a modest decline in performance compared to cases where judgments align with image-text correlation, the multi-level image information may still provide implicit semantic knowledge for MABSA. Therefore, aligning and fusing information from different modalities into an effective and robust multimodal representation holds great potential. This is not only because it aligns with human cognitive processes but also because it enables a more comprehensive understanding and richer representation.

5 Conclusions

In this paper, we present a novel multimodal textual-visual alignment and fusion network for performing joint multimodal aspect-sentiment analysis (JMASA). Our approach enables comprehensive interaction between textual and visual modalities by integrating multi-granularity alignment and multi-scale fusion techniques. Moreover, we introduce a text-centered multimodal training strategy to effectively address the noise introduced by the extensive visual context. Experimental results on two benchmark MABSA datasets demonstrate that our proposed model outperforms the state-of-the-art baselines in the MABSA task. Furthermore, in-depth analysis validates the effectiveness of our proposed model and the appropriateness of our chosen hyperparameters, highlighting its ability to accurately handle the complexities of JMASA in a comprehensive manner.

In future work, we aim to delve into more refined modeling approaches, extending the proposed method to cater to a broader range of multimodal tasks in practical applications, such as multimodal aspect sentiment triplet extraction. Furthermore, we plan to design alignment mechanisms for incorporating relevant multi-granularity visual contexts into our model training to reduce reliance on external alignment tools. This will enhance the overall robustness of our approach and enable more sophisticated integration of visual information into the analysis process.

Author Contributions YL: Methodology, Conceptualization, Investigation, Validation, Writing—original draft, Writing—review & editing. HD: Methodology, Software, Writing—Original Draft, Writing—review & editing, Funding acquisition, Validation. YL: Methodology, Writing—review & editing, Funding acquisition. XF: Validation, Writing—review & editing. LC: Resources, Funding acquisition, Supervision. All authors reviewed the manuscript.

Funding This work was supported by National Natural Science Foundation of China (Nos. 62362015, 62062027 and U22A2099), Innovation Project of GUET Graduate Education (No. 2023YCX5045) and the project of Guangxi Key Laboratory of Trusted Software.

Data availability The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Borth D, Ji R, Chen T, Breuel T, Chang S-F (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. ACM multimedia conference. Association for Computing Machinery, New York, pp 223–232
- Chen Y-C, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu J (2020) Uniter: Universal image-text representation learning. In: European conference on computer vision, pp. 104–120 . https://doi.org/10.1007/978-3-030-58577-8_7
- Chen Q, Ling Z-H, Zhu X (2018) Enhancing sentence embedding with generalized pooling. Proceedings of the 27th international conference on computational linguistics. Association for Computational Linguistics, Santa Fe
- Chen T, Borth D, Darrell T, Chang S (2014) DeepSentibank: Visual sentiment concept classification with deep convolutional neural networks. CoRR **abs/1410.8586**
- Chen Y, Gong S, Bazzani L (2020) Image search with text feedback by visiolinguistic attention learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) . <https://doi.org/10.1109/CVPR42600.2020.00307>
- Chen S, Liu J, Wang Y, Zhang W, Chi Z (2020) Synchronous double-channel recurrent network for aspect-opinion pair extraction. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 6515–6524. Association for Computational Linguistics, Online . <https://doi.org/10.18653/v1/2020.acl-main.582>
- Chen Z, Qian T (2019) Transfer capsule network for aspect level sentiment classification. In: Proceedings of the 57th annual meeting of the association for computational linguistics, pp. 547–556. Association for Computational Linguistics, Florence, Italy . <https://doi.org/10.18653/v1/P19-1052>
- Chen G, Tian Y, Song Y (2020) Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In: Proceedings of the 28th international conference on computational linguistics, pp. 272–279. International Committee on Computational Linguistics, Barcelona, Spain (Online). <https://doi.org/10.18653/v1/2020.coling-main.24>
- Chen X, Zhang N, Li L, Yao Y, Deng S, Tan C, Huang F, Si L, Chen H (2022) Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In: Findings of the association for computational linguistics: NAACL 2022, pp. 1607–1618. Association for Computational Linguistics, Seattle, United States . <https://doi.org/10.18653/v1/2022.findings-naacl.121>
- Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46. <https://doi.org/10.1177/001316446002000104>
- Devlin J, Chang M-W, Lee K, 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota . <https://doi.org/10.18653/v1/N19-1423>
- Ding Y, Yu J, Jiang J (2017) Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. Proc AAAI Conf Artif Intell. <https://doi.org/10.1609/aaai.v31i1.11014>
- Fan S, Shen Z, Jiang M, Koenig BL, Xu J, Kankanhalli M, Zhao Q (2018) Emotional attention: a study of image sentiment and visual attention. IEEE/CVF Conference on computer vision and pattern recognition 2018:7521–7531. <https://doi.org/10.1109/CVPR.2018.00785>

- Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A (2023) Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inform Fusion* 91:424–444
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp. 770–778 . <https://doi.org/10.1109/CVPR.2016.90>
- Hu M, Peng Y, Huang Z, Li D, Lv Y (2019) Open-domain targeted sentiment analysis via span-based extraction and classification. Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence
- Ju X, Zhang D, Xiao R, Li J, Li S, Zhang M, Zhou G (2021) 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, pp. 4395–4405. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic . <https://doi.org/10.18653/v1/2021.emnlp-main.360>
- Khan Z, Fu Y (2021) Exploiting bert for multimodal target sentiment classification through input space translation. In: Proceedings of the 29th acm international conference on multimedia. MM '21, pp. 3034–3042. Association for Computing Machinery, New York, NY, USA . <https://doi.org/10.1145/3474085.3475692>
- Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. *Behav Brain Sci* 40:253. <https://doi.org/10.1017/S0140525X16001837>
- Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi SCH (2021) Align before fuse: vision and language representation learning with momentum distillation. *Adv Neural Inform Process Syst* 34:9694–9705
- Li Y, Lin Y, Lin Y, Chang L, Zhang H (2022) A span-sharing joint extraction framework for harvesting aspect sentiment triplets. *Knowl Based Syst* 242:108366. <https://doi.org/10.1016/j.knsys.2022.108366>
- Liang B, Yin R, Du J, Gui L, He Y, Yang M, Xu R (2023) Embedding refinement framework for targeted aspect-based sentiment analysis. *IEEE Trans Affect Comput* 14(1):279–293. <https://doi.org/10.1109/TAFFC.2021.3071388>
- Li D, Li J, Li H, Nibbles JC, Hoi SCH (2021) Align and prompt: Video-and-language pre-training with entity prompts. IEEE/CVF conference on computer vision and pattern recognition (CVPR) 2022:4943–4953. <https://doi.org/10.1109/CVPR52688.2022.00490>
- Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International conference on machine learning. PMLR <https://doi.org/10.48550/arXiv.2201.12086>
- Ling Y, Yu J, Xia R (2022) 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), pp. 2149–2159. Association for Computational Linguistics, Dublin, Ireland . <https://doi.org/10.18653/v1/2022.acl-long.152>
- Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized BERT pretraining approach. *CoRR abs/1907.11692*
- Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F (2020) Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, pp. 121–137 . https://doi.org/10.1007/978-3-030-58577-8_8
- Luo Z, Huang S, Zhu KQ (2019) Knowledge empowered prominent aspect extraction from product reviews. *Inform Process Manag* 56(3):408–423. <https://doi.org/10.1016/j.ipm.2018.11.006>
- Mokady R, Hertz A, Bermano AH (2021) Clipcap: CLIP prefix for image captioning. *CoRR abs/2111.09734*
- Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748–8763 . <https://doi.org/10.48550/arXiv.2103.00020>
- Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. <https://api.semanticscholar.org/CorpusID:160025533>
- Sun L, Wang J, Zhang K, Su Y, Weng F (2021) Rpbert: A text-image relation propagation-based bert model for multimodal ner. *ArXiv abs/2102.02967* <https://doi.org/10.1609/aaai.v35i15.17633>
- Sun K, Zhang R, Mensah S, Mao Y, Liu X (2019) Aspect-level sentiment analysis via convolution over dependency tree. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 5679–5688. Association for Computational Linguistics, Hong Kong, China . <https://doi.org/10.18653/v1/D19-1569>
- Tang D, Qin B, Liu T (2016) Aspect level sentiment classification with deep memory network. In: Smith J (ed) Proceedings of the 2016 conference on empirical methods in natural language processing. Association for Computational Linguistics, Austin, pp 214–224

- Tian K, Jiang Y, Diao Q, Lin C, Wang L, Yuan Z (2023) Designing BERT for convolutional networks: sparse and hierarchical masked modeling. In: The Eleventh international conference on learning representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023
- Tu Y, Zhou C, Guo J, Gao S, Yu Z (2021) Enhancing the alignment between target words and corresponding frames for video captioning. *Pattern Recognit* 111:107702
- Tu Y, Li L, Su L, Gao S, Yan CC, Zha Z, Yu Z, Huang Q (2022) I²transformer: intra- and inter-relation embedding transformer for TV show captioning. *IEEE Trans Image Process* 31:3565–3577
- Wang W, Pan SJ (2020) Syntactically meaningful and transferable recursive neural networks for aspect and opinion extraction. *Comput Linguist* 45(4):705–736. https://doi.org/10.1162/coli_a_00362
- Wang W, Bao H, Dong L, Bjorck J, Peng Z, Liu Q, Aggarwal K, Mohammed OK, Singhal S, Som S, Wei F (2022) Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *ArXiv abs/2208.10442*<https://doi.org/10.48550/arXiv.2208.10442>
- Wang X, Gui M, Jiang Y, Jia Z, Bach N, Wang T, Huang Z, Tu K (2022) ITA: Image-text alignments for multi-modal named entity recognition. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 3176–3189. Association for Computational Linguistics, Seattle, United States . <https://doi.org/10.18653/v1/2022.naacl-main.232>
- Wu H, Cheng S, Wang J, Li S, Chi L (2020) Multimodal aspect extraction with region-aware alignment network. In: Zhu X, Zhang M, Hong Y, He R (eds) natural language processing and Chinese computing - 9th CCF international conference. Springer, pp 145–156
- Xue W, Li T (2018) Aspect based sentiment analysis with gated convolutional networks. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp. 2514–2523. Association for Computational Linguistics, Melbourne, Australia . <https://doi.org/10.18653/v1/P18-1234>
- Xu N, Mao W, Chen G (2019) Multi-interactive memory network for aspect based multimodal sentiment analysis. The thirty-third AAAI conference on artificial intelligence, AAAI 2019:371–378. <https://doi.org/10.1609/aaai.v33i01.3301371>
- Yan H, Dai J, Ji T, Qiu X, Zhang Z (2021) 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). Association for Computational Linguistics, Online . <https://doi.org/10.18653/v1/2021.acl-long.188>
- Yang L, Na JC, Yu J (2022) Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Inform Process Manag* 59(5):103038. <https://doi.org/10.1016/j.ipm.2022.103038>
- Yang X, Feng S, Wang D, Sun Q, Wu W, Zhang Y, Hong P, Poria S (2023) Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt. Findings of the association for computational linguistics: ACL 2023. Association for Computational Linguistics, Toronto
- Yang H, Zhao Y, Qin B (2022) Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In: Proceedings of the 2022 conference on empirical methods in natural language processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates . <https://doi.org/10.18653/v1/2022.emnlp-main.219>
- Yao F, Sun X, Yu H, Zhang W, Liang W, Fu K (2023) Mimicking the brain's cognition of sarcasm from multidisciplinary for twitter sarcasm detection. *IEEE Trans Neural Netw Learn Syst* 34(1):228–242. <https://doi.org/10.1109/TNNLS.2021.3093416>
- Yu W, Xu H, Meng F, Zhu Y, Ma Y, Wu J, Zou J, Yang K (2020) CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 3718–3727
- Yu J, Jiang J (2019) Adapting bert for target-oriented multimodal sentiment classification. In: International joint conference on artificial intelligence . <https://doi.org/10.24963/ijcai.2019/751>
- Yu J, Jiang J, Yang L, Xia R (2020) Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In: Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, Online, July 5-10, 2020, pp. 3342–3352. <https://doi.org/10.18653/v1/2020.acl-main.306>
- Yu J, Wang J, Xia R, Li J (2022) Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In: International joint conference on artificial intelligence . <https://doi.org/10.24963/ijcai.2022/622>
- Zadeh A, Chen M, Poria S, Cambria E, Morency L-P (2017) Tensor fusion network for multimodal sentiment analysis. Proceedings of the 2017 conference on empirical methods in natural language processing. Association for Computational Linguistics, Copenhagen, pp 1103–1114
- Zhang D, Wei S, Li S, Wu H, Zhu Q, Zhou G (2021) Multi-modal graph fusion for named entity recognition with targeted visual guidance. *Proc AAAI Conf Artif Intell* 35:14347–14355

- Zhao Q, Gao T, Guo N (2023) Tsvfn: two-stage visual fusion network for multimodal relation extraction. *Inform Process Manag* 60(3):103264. <https://doi.org/10.1016/j.ipm.2023.103264>
- Zhou J, Zhao J, Huang X, Hu Q, He L (2021) Masad: a large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing* 455:47–58
- Zhou R, Guo W, Liu X, Yu S, Zhang Y, Yuan X (2023) AoM: detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In: Findings of the association for computational linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada . <https://doi.org/10.18653/v1/2023.findings-acl.519>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.