



# Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis

Luwei Xiao<sup>a,c</sup>, Xingjiao Wu<sup>b,c,\*</sup>, Junjie Xu<sup>a,c</sup>, Weijie Li<sup>a</sup>, Cheng Jin<sup>b</sup>, Liang He<sup>a,c</sup>

<sup>a</sup> School of Computer Science and Technology, East China Normal University, Shanghai, 200062, China

<sup>b</sup> School of Computer Science, Fudan University, Shanghai, 200433, China

<sup>c</sup> Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, 200062, China

## ARTICLE INFO

### Keywords:

Aesthetic-oriented  
Multiple Granularities Fusion  
Multi-modal Sentiment Analysis  
Image Aesthetic Assessment  
Aspect-based Sentiment Analysis

## ABSTRACT

Joint Multi-modal Aspect-based Sentiment Analysis (JMASA) is a challenging task that seeks to identify all aspect-sentiment pairs from multimodal data. Current JMASA studies are insufficient in bridging the representational gap between textual and visual modalities. Additionally, they largely emphasize image feature extraction, neglecting the exploration of image presentation forms, like aesthetic characteristics. In this paper, we propose an Aesthetic-oriented Multiple Granularities Fusion Network for JMASA, termed Atlantis. This trident-shaped framework comprises three branches: Textual-vision Alignment Aspect-sentiment Extraction, Sentiment-aware Image Aesthetic Assessment, and Aesthetic-aware JMASA. Notably, the first two branches function as auxiliary learning tasks, with Textual-vision Alignment Aspect-sentiment Extraction aimed at bridging the representational gap between modalities, and Sentiment-aware Image Aesthetic Assessment dedicated to understanding the aesthetic attributes of images. Concurrently, the Aesthetic-aware JMASA dynamically integrates varied granular features from both branches to perform JMASA. To the best of our knowledge, this is the first aesthetic-oriented approach in the present field. Experimental results on two public datasets verify that Atlantis outperforms a series of prior strong methodologies and achieves a new state-of-the-art (SOTA) performance. The enhancement highlights Atlantis's advanced capability in accurately identifying aspect-sentiment pairs with aesthetic features.

## 1. Introduction

Multimodal learning tasks integrate information from multiple sensory channels to perform multimodal reasoning by leveraging a harmonious combination of different modalities [1]. Such tasks call for a high degree of proficiency in information fusion, playing a crucial role in enabling both humans and intelligent robots to acquire the capability to understand and interpret the physical world [2]. In the last decade, the advancement of digital technologies, coupled with the widespread availability of advanced smartphones, has substantially improved the simplicity and quality of creating and sharing multimodal content on social media platforms [3,4]. Mining emotions conveyed by user-generated multimodal content offers invaluable insights into public opinions across a broad spectrum of fields, including business, political policy, healthcare, and others [5–7]. Joint Multimodal Aspect-based Sentiment Analysis (JMASA), a standard instance of multimodal learning tasks, aims to jointly extract all specific aspects within the sentence and predict their sentiment polarity (*Positive*, *Neutral* and

*Negative*) given a sentence-image pair. For instance, as depicted in Fig. 1(a), upon receiving a multimodal post as input, the JMASA framework is anticipated to identify two aspect-sentiment pairs, namely (Rob Manfred, *Positive*) and (MLB, *Neutral*). As a challenging task, JMASA has emerged as a hot topic within the realm of multimodal learning. It has drawn growing interest from scholarly circles in computer vision, natural language processing [8], and human–computer interaction, reflecting its interdisciplinary significance.

Numerous prior studies typically approach this task by devising visual-textual fusion-based strategies [9–11]. A pioneer work [9] first defines JMASA task and proposes a joint learning approach with image-text relation mechanism to assess the contribution of visual contents to extract aspect-sentiment pairs. Subsequently, Ling et al. [10] introduce a task-specific vision-language pre-training approach for JMASA. This approach involves aligning visual and textual representations through a set of collaborative vision-language pre-training strategies. More recently, Yang et al. [11] design a Cross-Modal Multitask Transformer

\* Corresponding author at: School of Computer Science, Fudan University, Shanghai, 200433, China.

E-mail addresses: [louisshaw@stu.ecnu.edu.cn](mailto:louisshaw@stu.ecnu.edu.cn) (L. Xiao), [xjwu\\_cs@fudan.edu.cn](mailto:xjwu_cs@fudan.edu.cn) (X. Wu), [jjxu\\_dr@stu.ecnu.edu.cn](mailto:jjxu_dr@stu.ecnu.edu.cn) (J. Xu), [Weij.Li@stu.ecnu.edu.cn](mailto:Weij.Li@stu.ecnu.edu.cn) (W. Li), [jc@fudan.edu.cn](mailto:jc@fudan.edu.cn) (C. Jin), [lhe@cs.ecnu.edu.cn](mailto:lhe@cs.ecnu.edu.cn) (L. He).

<https://doi.org/10.1016/j.inffus.2024.102304>

Received 7 January 2024; Received in revised form 2 February 2024; Accepted 12 February 2024

Available online 15 February 2024

1566-2535/© 2024 Elsevier B.V. All rights reserved.



Image		
Text	(a) Congrats to [Rob Manfred] <sub>Positive</sub> on being named the next Commissioner of [MLB] <sub>Neutral</sub> .	(b) RT @ NiallOfficial : another shot from [Santiago] <sub>Positive</sub> ! @ CalAurand
Aes-Cap	"Nice shot but i think the background is a little distracting."	"I like the lighting and the angle of the shot. the colors are nice."
Aes-Score	0.509	0.696

Fig. 1. Instances of JMASA. Aes-Cap and Aes-Score correspond to the aesthetic caption and aesthetic score generated by VILA, respectively.

(CMMT) to model inter-dynamics between visual and textual modalities via two auxiliary tasks and a dynamic multimodal gate mechanism. These cross-modal fusion-based methodologies have driven remarkable progress in the performance of JMASA.

However, they have been plagued with the following issues: (i) Despite various noise-filtering mechanisms being employed, these methods directly combine visual content with language, preventing the JMASA model from learning a more holistic grasp of multimodal data for aspect-sentiment pair extraction. (ii) A sizable body of studies has overlooked the fact that image aesthetic features typically embody richer expressive sentiments. “*Aesthetic experiences profoundly engage our emotions and feelings*” [12]. The aesthetic experience, inherently a visually expressed implicit emotion, offers an avenue to identify visual cues that correlate with this experience. Investigating these cues could lead to the detection of sentiment-informed visual features, advantageous for the JMASA framework. For example, as shown in Fig. 1(b), the sentence “RT @ NiallOfficial : another shot from Santiago! @ CalAurand” displays no sentimental clues for predicting the sentiment polarity of the aspect “Santiago”. Meanwhile, the image content is intricate and unrelated to the aspect within the sentence. In this scenario, regardless of the efficiency of the image noise filtering mechanism, it consistently introduces superfluous visual features that perturb the model. However, this issue can be effectively alleviated by incorporating Image Aesthetic Assessment (IAA). IAA seeks to assess the aesthetic qualities of images. This objective is achieved through a systematic evaluation, including techniques such as image aesthetic captioning and aesthetic scoring, to qualitatively and quantitatively analyze their visual appeal [13]. Psychological studies have empirically supported the notion that images possess the capacity to elicit a spectrum of emotional responses, a phenomenon influenced by both their aesthetic characteristics and semantic content [14]. For one thing, the aesthetic caption exhibits a seamless representational alignment with the sentence and is rich in emotional expression. For another, as a quantitative metric, the aesthetic score assesses the visual attractiveness of an image, enabling the quantification of the implicit emotions conveyed through an aesthetic form. Consequently, an optimally functioning JMASA framework is expected to be well-versed in cross-modal alignment, effectively leveraging the entirety of visual content, including aesthetic features.

In this paper, we propose an Aesthetic-oriented Multiple Granularities Fusion Network (Atlantis) for JMASA. This trident-shaped architecture mitigates the aforementioned challenges via three fundamental branches: Textual-vision Alignment Aspect-sentiment Extraction, Sentiment-aware Image Aesthetic Assessment, and Aesthetic-aware JMASA. Specifically, Textual-vision Alignment Aspect-sentiment

Extraction bridges the representational gap by converting the input image into a scene graph and a textual caption, both serving as textual supplements. It subsequently models the inter-dynamics by integrating them with the input sentence. Sentiment-aware Image Aesthetic Assessment guides the model to perceive image aesthetic features by incorporating image aesthetic attributes and textual sentiment features. Finally, the Aesthetic-aware JMASA adaptively merges the inter-dynamics with aesthetic-oriented features. This strategic combination enables the accurate identification of both explicit and implicit sentimental cues from language and vision modalities, facilitating the final prediction of aspect-sentiment pairs. Consequently, the Atlantis model accomplishes state-of-the-art performance across the majority of evaluative metrics on two extensively utilized, challenging Twitter datasets, verifying its marked effectiveness.

Briefly, our contributions can be summarized as follows:

- We introduce Atlantis, an innovative aesthetic-oriented methodology for Joint Multimodal Aspect-based Sentiment Analysis. This framework is designed to align images and text within the language modality across multiple granularities.
- We devise a Sentiment-aware Image Aesthetic Assessment to engage the model with the implicit emotional content conveyed by the image’s aesthetic attributes. To the best of our knowledge, this is the first attempt to explore the interrelationship between multimodal aspect-based sentiment analysis and image aesthetics.
- We conduct extensive experiments and meticulous analysis on two widely used public datasets. Experimental results show that Atlantis surpasses the performance of current state-of-the-art baselines, demonstrating its efficacy and superior capabilities.

The remainder of this paper is structured as follows: Section 2 discusses the related works in the fields of JMASA and IAA. Section 3 presents the Atlantis framework. Section 4 details the experimental results and corresponding analyses of Atlantis. The paper concludes in Section 5, summarizing the key findings and contributions of this study.

## 2. Related work

### 2.1. Multimodal aspect-based sentiment analysis

Multi-modal Aspect-based Sentiment Analysis (MABSA) is a challenging and fine-grained task that demands a model’s capability in both comprehending multimodal data and mining sentimental clues [15,16]. An increasing number of researchers have recently focused on this task.

Various approaches have been developed over this time span [17–20]. Generally speaking, the MABSA task can be grouped into three subtasks: Multimodal Aspect Term Extraction (MATE), Multimodal Aspect-based Sentiment Classification (MASC) and Joint Multimodal Aspect-based Sentiment Analysis (JMASA) [21]. Given a sentence-image pair, the goal of MATE is to extract all specific aspects/targets within the sentence. Concurrently, MASC aims to predict the sentiment polarity of each aspect term (target). More recently, a series of studies integrated these two subtasks into a singular end-to-end process, known as the JMASA. Approaches to MATE typically treat this task as a sequence-labeling problem. Various neural network-based methodologies have been developed, including those utilizing recurrent neural networks [22,23], Transformer [24,25], and graph neural networks [26,27]. In the context of MASC, the aspect terms or specific targets are generally provided. Numerous studies [28,29] throw emphasis on directly modeling the inter-dynamics between aspect, sentence, and image. Alternatively, some studies [30–32] endeavors involve translating the image into a textual space to facilitate alignment between visual and textual modalities. As for JMASA, Ju et al. [9] proposed the integration of MATE and MASC into an end-to-end task by exploiting a multimodal joint learning approach to capture the aspect-sentiment pairs. Ling et al. [10] developed a pre-training model for a unified multimodal encoder-decoder architecture tailored to MABSA, incorporating a range of vision-language pre-training techniques. More recently, Yang et al. [11] leveraged a Cross-Modal Multitask Transformer (CMMT) to derive the sentiment-aware features for each modality and dynamically controlled the impact of visual information on the textual content during the inter-modal interaction.

However, the innate semantic gap between visual and language modalities remains a huge challenge for the use of these methods. In addition, they lose sight of the aesthetic attributes of images, which potentially convey a more profound emotional expression than basic image features. To alleviate these issues, this study presents Atlantis, a framework tailored to align visual and textual modalities to the language space and extract image aesthetic attributes for the JMASA.

## 2.2. Image aesthetic assessment and emotion

The aesthetics of an image are defined as an evaluation or admiration of its beauty [33,34]. Image Aesthetic Assessment (IAA) aims to appraise the aesthetic quality of images by focusing on a systematic evaluation of their visual appeal. Psychological research substantiates the idea that images provoke a range of emotions, influenced by their aesthetic attributes and semantic content [14]. A wealth of studies have delved into exploring the interrelationship between aesthetic attributes and emotions. Datta et al. [35] introduced the concept of the “aesthetics gap” and investigated several fundamental aspects of algorithmic inference related to the emotions that image aesthetics arouse in humans. Joshi et al. [36] systematically analyzed the principal aspects of computational reasoning regarding aesthetics and emotional interpretation of images, grounded in the disciplines of philosophy, photography, painting, visual arts, and psychology. More recently, Yang et al. [37] developed a novel, attribute-rich, personalized image aesthetics database (PARA) and conducted extensive subjective studies on personalized image aesthetics. Their findings suggested that images with an aesthetic score below 2.0 are more inclined to evoke negative emotions, whereas those with higher aesthetic scores typically elicit positive emotional responses in subjects. Lan et al. [38] proposed a Hypernetwork of Emotion Fusion for (HNEF) image aesthetics assessment BY merging aesthetic and emotional features from the images. Ke et al. [13] exploited contrastive and generative objectives to train a Vision-Language Aesthetics (VILA) framework. This method focuses on modeling multimodal aesthetic representations from image-comment pairs and learning rich and comprehensive aesthetic semantics.

In conclusion, the relationship between image aesthetics and sentiment is inherently multidimensional, embracing a spectrum of individual psychological reactions, cultural backgrounds, and social contexts.

This relationship not only enriches our comprehension of the concept of beauty but also provides a unique perspective for exploring the deeper structures of human emotions. In this study, we integrated image aesthetic assessment into the JMASA to investigate the intricate relationships between visual elements and emotional impact.

## 3. Method

In this section, we discuss the task formulation for JMASA. Subsequently, we provide an overview of the proposed model, Atlantis, which consists of three primary branches: textual-vision alignment aspect sentiment extraction, sentiment-aware image aesthetic assessment, and aesthetic-aware JMASA. Finally, we provide a detailed introduction to each branch of Atlantis. The overall architecture of Atlantis is illustrated in Fig. 2.

### 3.1. Task formulation

Given a collection of multimodal sentence-image pairs, denoted as  $\mathcal{M}$ . Each pair  $m_i \in \mathcal{M}$  comprises a sentence  $S_i = (w_1, w_2, \dots, w_n)$  and a corresponding image  $V_i$ . The objective of JMASA is to predict the corresponding aspect-sentiment pair  $y = (y_1, y_2, \dots, y_n)$  for each sentence-image pair. Here,  $y_i \in \{B - POS, I - POS, B - NEG, I - NEG, B - NEU, I - NEU\} \cup \{O\}$ . In this case, B refers to the initial token of the aspect term, I indicates tokens within the specific aspect term and O indicates words “outside” the specific aspect. Moreover, POS, NEU, and NEG are abbreviations for positive, neutral, and negative sentiments associated with the specific aspect.

### 3.2. Model overview

The proposed Atlantis method comprises three branches: textual-vision alignment aspect sentiment extraction (TAAE), Sentiment-aware Image Aesthetic Assessment (SIAA), and Aesthetic-aware JMASA. The TAAE initially aligns textual and visual modalities, transforming an image into text-based representations through caption generation and construction of an unbiased scene graph. This generated textual content is then dynamically merged with the input sentence to predict auxiliary end-to-end textual aspect-sentiment pair extraction. SIAA focuses on learning aesthetic prediction principles based on identifiable themes and RGB color distributions. This process is further enriched by integrating aesthetic captions, aiding the framework in perceiving subtle emotions conveyed through aesthetic attributes. The aesthetic-aware JMASA converts an image into an emotionally rich aesthetic caption. Subsequently, it models the aesthetic-aware multimodal representation by adaptively fusing aligned cross-modal semantic features from the TAAE with an aesthetic caption to make JMASA predictions.

### 3.3. Textual-vision alignment aspect-sentiment extraction

Current approaches typically fuse the visual content with language either through intuitive, direct interaction or by implementing a variety of noise-filtering mechanisms. These mechanisms aim to eliminate image noise prior to modeling the interactive dynamics between modalities. Despite the effectiveness of these mechanisms, an inherent representational gap persists, which inadvertently introduces aspect-irrelevant visual noise. Therefore, we devised a Textual-vision Alignment Aspect-sentiment Extraction (TAAE) branch to align vision and language within a unified space.

**Caption and scene graph generation:** This component comprises two parts: a caption and a scene graph. For an input image  $V \in \mathbb{R}^{3 \times H \times W}$ , we feed it into a pre-trained CoCa [39] to generate its corresponding caption  $C = (w_1^c, w_2^c, \dots, w_{n_c}^c)$ . Although this caption offers a comprehensive representation of global image information, it fails in capturing detailed, fine-grained image-level knowledge, such as



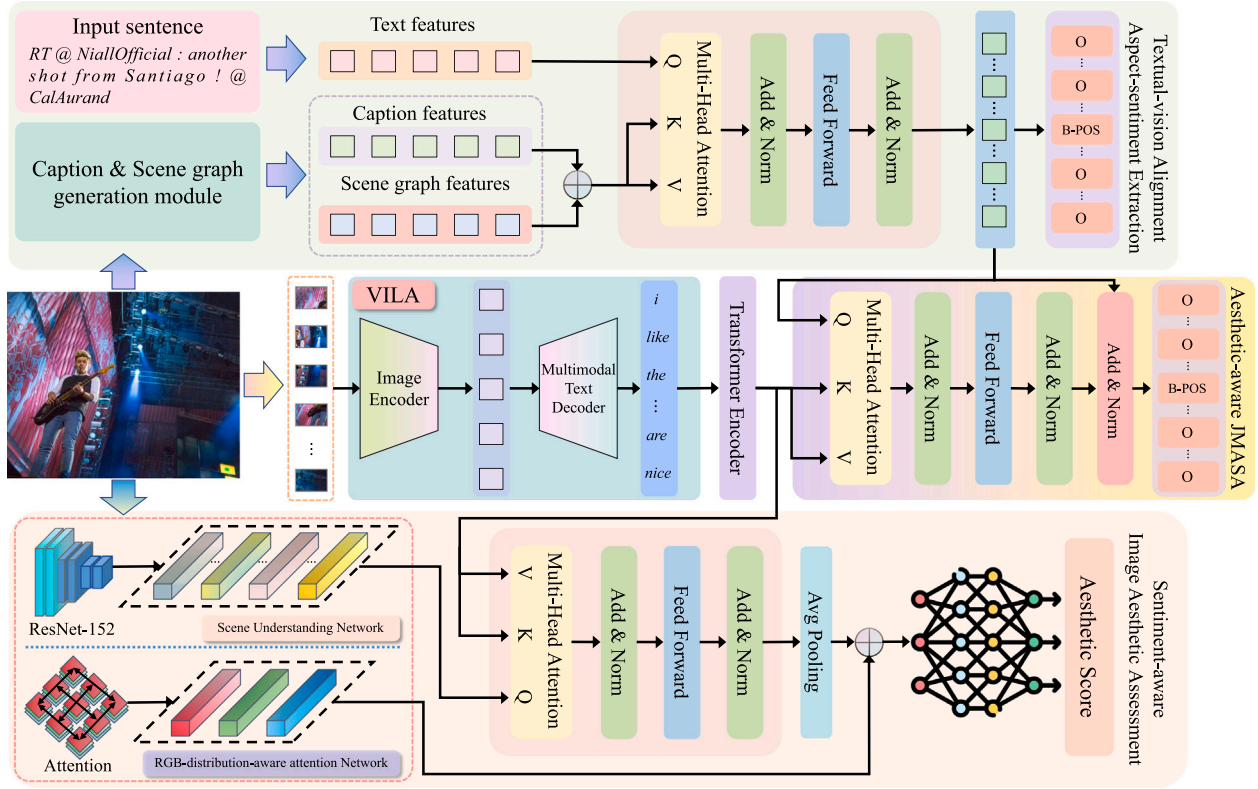


Fig. 2. The overview of Aesthetic-oriented Multiple Granularities Fusion Network (Atlantis).

the relationships among objects. Consequently, we adopt a scene graph generation framework [40] to derive a scene graph of the image. This scene graph is composed of recall@5 (subject, predicate, object) triples, such as *(Person, wearing, Jean)*, depicting the object-level relationship within the image. This granular, object-level information augments multimodal data integration, offering an enriched visual context. For the scene graph  $G = (g_1, g_2, \dots, g_{n_g})$ , we transform this graph into a linearized sentence by concatenating all triples:

$$G^s = (g_1; g_2; \dots; g_i \dots; g_{n_g}) \quad (1)$$

where  $g_i = (s_i, r_i, o_i)$ .  $s_i$  and  $o_i$  denote the subject and the object, respectively.  $r_i$  is a relationship predicate. “;” refers to concatenation. We utilize RoBERTa [41] to derive a contextualized feature representation of the input sentence, caption and serialized graph. These feature representations are denoted as  $H = (h_1, h_2, \dots, h_n)$ ,  $H^c = (h_1^c, h_2^c, \dots, h_{n_c}^c)$  and  $H^g = (h_1^g, h_2^g, \dots, h_{n_g}^g)$ , respectively. To comprehensively represent the visual content, we further concatenate the caption feature  $H^c = (h_1^c, h_2^c, \dots, h_{n_c}^c)$  with the serialized scene graph feature  $H^g = (h_1^g, h_2^g, \dots, h_{n_g}^g)$ , denoted as the visual content feature  $V^c = (H^c; H^g)$ . We consider the text feature  $H$  as *Query* and the visual content feature  $V^c$  as the *Key* and *Value*. These features are passed through a Transformer encoder [42] to model the cross-modal aligned features as follows:

$$\text{ATT}^i(H, V^c, V^c) = \text{softmax} \left( \frac{[W_q^i H]^\top [W_k^i V^c]}{\sqrt{d/h}} \right) [W_v^i V^c]^\top \quad (2)$$

$$\tilde{H} = W_h [\text{ATT}^1(H, V^c, V^c); \dots; \text{ATT}^h(H, V^c, V^c)]^\top \quad (3)$$

$$\bar{H} = \text{LayerNorm}(\tilde{H} + H) \quad (4)$$

$$H^{cm} = \text{LayerNorm}(\text{FFN}(\bar{H}) + \bar{H}) \quad (5)$$

where  $h$  refers to the number of head within the multi-head attention (MHA),  $\{W_q^i, W_k^i, W_v^i\} \in \mathbb{R}^{d/h \times d}$  indicate the  $i$ th learnable parameter matrices for *Query*, *Key*, and *Value*, respectively.  $W_h \in \mathbb{R}^{d \times d}$  is the parameter matrix of MHA. *LayerNorm* and *FFN* are the layer normalization [43] and feed-forward network, respectively. Moreover,  $H^{cm} = (h_1^{cm}, h_2^{cm}, \dots, h_n^{cm}) \in \mathbb{R}^{d \times n}$  is the cross-modal aligned features.

**Textual-vision Alignment Auxiliary Supervision:** The cross-modal aligned feature  $h_i^{cm}$  is exploited to predict the aspect-sentiment label for the  $i$ th word. To optimize the auxiliary End-to-End Textual aspect-sentiment pair extraction, we employed the standard cross-entropy loss as follows:

$$p_i^{cm} = \text{Softmax}(W^\top h_i^{cm} + b) \quad (6)$$

$$\mathcal{L}_{cm} = \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^{n_j} l(p_i^{cm}, z_i) \quad (7)$$

where  $n_j$  represents the count of words in the  $j$ th sample, while  $N$  signifies the total number of samples. The variable  $z_i$  corresponds to the ground-truth for the  $i$ th token. Furthermore,  $l$  signifies the cross-entropy loss function.

### 3.4. Sentiment-aware image aesthetic assessment

Aesthetic attributes such as composition, color, and texture convey subtle emotional cues, offering deeper insight into the sentiments conveyed by visual content. The incorporation of image aesthetic features into the JMASA framework has the potential to capture both explicit and implicit emotional expressions, thereby augmenting the overall effectiveness of JMASA. Consequently, we propose a Sentiment-aware Image Aesthetic Assessment (SIAA) branch to steer the model toward recognizing aesthetic features and extracting the underlying emotional cues.

**Composition Understanding Network:** The arrangement of elements within an image, such as the positioning of objects, symmetry, and balance, is crucial for aesthetic attributes. First, we extract the composition features of the image by employing a pre-trained Residual Network (ResNet) [44]. The input image  $V$  is initially subjected to pre-processing, resizing it to dimensions of  $224 \times 224$  pixels. This processed image is then input into a pre-trained 152-layer ResNet model. The features extracted from the final convolutional layer of this model are subsequently utilized as the composition features  $\hat{V} \in \mathbb{R}^{2048 \times 49}$ . Within the composition feature  $\hat{V}$ , it consists of  $7 \times 7$  visual regions. Each of these regions is represented by a feature vector with a dimensionality of 2048. Moreover, we implement a linear transformation function on  $\hat{V}$  to obtain the final composition feature  $\bar{V} = W^T \hat{V}$ , where  $W \in \mathbb{R}^{2048 \times d}$  is trainable parameter matrix and  $\bar{V} \in \mathbb{R}^{d \times 49}$ .

**High-level RGB-aware Attention Network:** High-level color features can reveal the harmony or discord among colors in an image, which is a crucial aspect of aesthetic appeal [45]. Additionally, colors play a significant role in conveying emotions. For instance, warm colors may evoke feelings of warmth and comfort, whereas cool colors may create a sense of calm or sadness. Analyzing high-level color features helps understanding of the emotional tone of an image. Motivated by He et al. [46], we propose to leverage a color-distribution-aware self-attention mechanism to model high-level color features. Specifically, the input image is partitioned into distinct, non-overlapping patches. Each patch is characterized by a central point, defined as the mean of the raw pixel RGB values within that patch. For each image  $V$ , irrespective of its input size, the patch space comprises  $k \times k$  central points. This configuration yields a linear relationship with the input size, ensuring low computational complexity. In this context,  $k$  represents a hyper-parameter, which is assigned a value of 12. Subsequently, it focuses on the extraction of inter-patch relationships, effectively bypassing the need for multiplication by the input. Assuming  $V_{fa}$  and  $V_{fb}$  as two central points, the color features  $V_{rgb}$  is calculated as follows:

$$V_{rgb} = \left\|_{ch=1}^{N_{ch}} \left( \text{Softmax} \left( W_{rgb} \left( \frac{[W_q^{ch} V_{fa}]^T [W_k^{ch} V_{fb}]}{\sqrt{d_r}} \right) \right) \right) \right\| \quad (8)$$

where  $\|_{ch=1}^{N_{ch}}$  is the concatenation of RGB channels.  $W_q^{ch}$  and  $W_k^{ch}$  are weight matrices for *Query* and *Key*, respectively.  $d_r$  is the dimension.  $W_{rgb}$  refers to a color-specific weight matrix, responsible for converting RGB-aware features into refined color features. These features are subsequently fed into a *softmax* layer, obtaining the ultimate high-level color features  $V_{rgb}$ .

**Cross-modal aesthetic attribute fusion:** Aesthetic comments provide a textual interpretation of the aesthetic elements within an image, offering a linguistic perspective that complements the visual data. The integration of textual and visual aesthetic descriptions enriches our understanding of an image's aesthetic attributes. To this end, we exploit a pretrained framework for learning image aesthetics (VILA) to generate an aesthetic caption for each image. Subsequently, we extract textual aesthetic features  $H^{fa} = (h_1^{fa}, h_2^{fa}, \dots, h_{n_a}^{fa})$  from the aesthetic caption using a transformer encoder, the details of which are elaborated in Section 3.5. We treat the composition feature  $\bar{V}$  as *Query*, and the textual aesthetic features  $H^a$  are regarded as *Key* and *Value*. These feature representations are then collectively processed through a transformer encoder, following a computational procedure akin to that described in Eqs. (2)–(5). Consequently, the cross-modal aesthetic features  $V^{ac} \in \mathbb{R}^{d \times 49}$  is obtained. An average pooling operation is then applied to  $V^{ac}$ , leading to the derivation of the refined cross-modal aesthetic features  $\hat{V}^{ac}$ .

**Aesthetic score generation:** “Aesthetic score” is generally indicative of a quantitative evaluation reflecting the aesthetic quality or visual appeal of an image [47]. This score is typically derived from evaluating various elements that contribute to the overall aesthetic

experience, such as composition, color harmony, symmetry, and emotional impact. In this study, the aesthetic score was assigned a value ranging from 0 to 1, where a higher score correlates with an improved aesthetic quality. Our research initially focused on harnessing aesthetic scores as a means of comprehending the comprehensive aesthetic attributes of the JMASA. Specifically, we adopt a pre-trained VILA [13] to generate the aesthetic score for each image. For example, as shown in Fig. 1, we present two images accompanied by their respective aesthetic scores, as produced by VILA. We then employ the aesthetic score to assist Atlantic in perceiving the aesthetic attributes of the visual content.

**Aesthetic Auxiliary Supervision:** We concatenate the high-level color features  $V_{rgb}$  and the cross-modal aesthetic features  $\hat{V}^{ac}$  to predict the aesthetic score  $\hat{s}_i$  of the image as follows:

$$\hat{s}_i = \sigma \left( W_s \left( \text{ReLU} \left( V_{rgb}; \hat{V}^{ac} \right) \right) \right) \quad (9)$$

where  $W_s$  is the learnable matrix,  $\sigma$  denotes the *sigmoid* function. For optimizing the auxiliary aesthetic score prediction, we employ the standard MSE loss as follows:

$$\mathcal{L}_{score} = \frac{1}{N} \sum_{i=1}^N (s_i - \hat{s}_i)^2 \quad (10)$$

where  $N$  is the number of samples.  $s_i$  indicates the aesthetic score generated by VILA for the corresponding image.

### 3.5. Aesthetic-aware JMASA

Given that the majority of aesthetic attributes are conveyed through visual content, an approach that directly models these features alongside textual-based, cross-modal aligned features is prone to encountering a relatively significant semantic gap. Therefore, this branch aims to generate the textual aesthetic features and model the inter-modal dynamics between these features and the cross-modal aligned features. Additionally, the textual aesthetic features act as an intermediate bridge, linking the learning of textual sentiment with the understanding of image aesthetic attributes, making them go hand in hand.

**Aesthetic caption generation:** An aesthetic caption refers to textual descriptions of image aesthetics. For instance, as shown in Fig. 1, “nice shot but i think the background is a little distracting”. and “i like the lighting and the angle of the shot. the colors are nice”. reflects different aesthetic opinions, respectively. We utilize pre-trained VILA [13] to generate the aesthetic caption for each image. For each aesthetic caption, we utilize RoBERTa [41] to derive its contextualized feature representation  $H^a = (h_1^a, h_2^a, \dots, h_{n_a}^a)$ . We further feed it into a transformer encoder as follows:

$$\text{ATT}^i(H^a, H^a, H^a) = \text{softmax} \left( \frac{[W_q^i H^a]^T [W_k^i H^a]}{\sqrt{d/h}} \right) [W_v^i H^a]^T \quad (11)$$

$$\tilde{H}^a = W_h [\text{ATT}^1(H^a, H^a, H^a); \dots; \text{ATT}^h(H^a, H^a, H^a)]^T \quad (12)$$

$$\bar{H}^a = \text{LayerNorm}(\tilde{H}^a + H^a) \quad (13)$$

$$H^{fa} = \text{LayerNorm}(\text{FFN}(\bar{H}^a) + \bar{H}^a) \quad (14)$$

where  $h$  denotes the number of heads in the multi-head attention (MHA). The set  $\{W_q^i, W_k^i, W_v^i\} \in \mathbb{R}^{d/h \times d}$  represents the  $i$ th learnable weight matrices for *Query*, *Key*, and *Value* respectively. The parameter matrix for MHA is denoted as  $W_h \in \mathbb{R}^{d \times d}$ . *LayerNorm* and *FFN* refer to layer normalization [43], and the feed-forward network, respectively. Additionally,  $H^{fa} = (h_1^{fa}, h_2^{fa}, \dots, h_{n_a}^{fa}) \in \mathbb{R}^{d \times n_a}$  is the textual aesthetic features.

**Aesthetic and cross-modal feature fusion:** We merge the cross-modal aligned features  $H^{cm}$  and the textual aesthetic features  $H^{fa}$ , attaining fused aesthetic-aware textual representation as follows:

$$\text{ATT}^i(H^{cm}, H^{fa}, H^{fa}) = \text{softmax} \left( \frac{[W_q^i H^{cm}]^T [W_k^i H^{fa}]}{\sqrt{d/h}} \right) [W_v^i H^{fa}]^T \quad (15)$$

$$\tilde{H}^{cm} = W_h [\text{ATT}^1(H^{cm}, H^{fa}, H^{fa}); \dots; \text{ATT}^h(H^{cm}, H^{fa}, H^{fa})]^T \quad (16)$$

$$\bar{H}^{cm} = \text{LayerNorm}(\tilde{H}^{cm} + H^{cm}) \quad (17)$$

$$\hat{H}^{cm} = \text{LayerNorm}(FFN(\bar{H}^{cm}) + \bar{H}^{cm}) \quad (18)$$

$$H^{ac} = \text{LayerNorm}(\hat{H}^{cm} + H^{cm}) \quad (19)$$

where  $H^{ac} \in \mathbb{R}^{d \times n}$  is the final fused aesthetic-aware textual representation.

**Output layer:** The final fused aesthetic-aware textual representation  $H^{ac} = (h_1^{ac}, h_2^{ac}, \dots, h_n^{ac}) \in \mathbb{R}^{d \times n}$  is subsequently passed into a conventional Conditional Random Field (CRF) layer to predict the label sequence  $y$ :

$$P(y) = \frac{\exp(\text{score}(H^{ac}, y))}{\sum_{y' \in Y_{H^{ac}}} \exp(\text{score}(H^{ac}, y'))} \quad (20)$$

$$\text{score}(H^{ac}, y) = \sum_{i=0}^n T_{y_i, y_{i+1}} + \sum_{i=1}^n E_{h_i^{ac}, y_i} \quad (21)$$

$$E_{h_i^{ac}, y_i} = w^{y_i} \cdot h_i^{ac} \quad (22)$$

where  $T_{i,j}$  represents the transition score from label  $i$  to label  $j$ ,  $E_{h_i^{ac}, y_i}$  denotes the emission score associated with label  $y_i$ , and  $w^{y_i}$  refers to the corresponding weight vector for  $y_i$ .

**JMASA supervision:** We minimize the negative log-probability of the accurately labeled sequence as follows:

$$\mathcal{L}_{\text{JMASA}} = -\frac{1}{N} \sum_{j=1}^N \left( \text{score}(H_j^{ac}, y_j) - \log \sum_{y'_j \in Y_{H_j^{ac}}} \exp(\text{score}(H_j^{ac}, y'_j)) \right) \quad (23)$$

where  $N$  indicates the total count of samples,  $y_j$  represents the correct label sequence for the  $j$ th sample.  $Y_{H_j^{ac}}$  embodies the complete set of potential label sequences applicable to the input tokens.

### 3.6. Model training

To achieve optimal parameter tuning in Atlantis, we optimize the principal task and two auxiliary tasks as follows:

$$\mathcal{L} = \mathcal{L}_{\text{JMASA}} + \alpha \mathcal{L}_{cm} + \beta \mathcal{L}_{scores} \quad (24)$$

where  $\alpha$  and  $\beta$  serve as tradeoff hyper-parameters, functioning to regulate the respective contributions of each auxiliary task.

## 4. Experiments

In this section, a series of comprehensive experiments are conducted on two JMASA datasets to verify the efficacy of the proposed Atlantis framework.

### 4.1. Experimental settings

**Datasets:** Akin to Yang et al. [11], this study exploits two twitter-based datasets developed by Yu et al. [25]. Comprehensive statistics of these datasets are presented in Table 1. The Twitter-2015 dataset consists of 3179 sentence-image pairs for training, 1122 for development, and 1037 for testing. It includes 17,180 words and 5338 associated images, with an average tweet length of approximately 16.8 words. Similarly, the Twitter-2017 dataset contains 3562 sentence-image pairs for training, 1176 for development, and 1234 for testing. This dataset includes 11,962 words and 5972 images, with an average tweet length of about 16.2 words. Both datasets categorize specific aspects within tweets into Positive, Negative, and Neutral sentiment polarities.

**Implementation details:** We employ For initializing word representations in tweets, captions, serialized scene graphs, and aesthetic captions, we utilize RoBERTa<sub>base</sub>. The learning rate for both datasets is configured at 4e-5. Hyper-parameters  $\alpha$  and  $\beta$  are fixed at 0.1 and 0.7, respectively, across the datasets. Other hyper-parameters, such as hidden dimension  $d$ , warm-up proportion, and number of attention heads  $h$ , follow the default settings by Liu et al. [41], with values 768, 0.1, and 12, respectively. Furthermore, the maximum sentence length is fixed at 128, and the training is conducted over 40 epochs. The effectiveness of the Atlantis model was evaluated using the precision (P), recall (R), and Micro-F1 score (F1) as key performance metrics. All the models were constructed using the PyTorch framework.

### 4.2. Compared baselines

We performed a comparative evaluation of Atlantis by contrasting it with two methodological classes. The first comprises text-based methods:

- **SPAN** [48] is a span-based approach for the End-to-End Textual Aspect-Based Sentiment Analysis (ABSA) task. It initially employs an LSTM-based multi-span decoding algorithm to extract aspects, subsequently predicting the sentiment polarities grounded in the identified span representations.
- **D-GCN** [49], a BERT-based Directional Graph Convolutional Network, approaches the End-to-End ABSA task through the paradigm of sequence labeling. It integrates syntactic dependency to facilitate the simultaneous detection of aspects and their sentiment polarity.
- **RoBERTa** [41] serves as a textual baseline, which directly feeds the textual representation into a Transformer encoder, followed by a CRF layer for identifying aspect-sentiment pairs.
- **BART** [50] modifies the textual End-to-End ABSA task for implementation within the BART architecture, formulating it as a problem of index generation.

Moreover, for a comprehensive comparison, this study includes the following strong methodologies based on multimodal frameworks:

- **UMT+TomBERT** is a hybrid method consists of Unified Multimodal Transformer (UMT) [51] for the fusion of text and image representations to identify aspects. Concurrently, it employs the Target-oriented Multimodal BERT (TomBERT) [28] to predict the sentiment polarity of each detected aspect.
- **RpBERT-collapse** [24], **UMT-collapse** [25] and **OSCGA-collapse** [22] are approaches developed for Multimodal Aspect-based Target Extraction (MATE). These approaches have been adapted for the Joint Multimodal Aspect-based Sentiment Analysis (JMASA) by Ju et al. [9]. RpBERT-collapse represents a relation propagation-based model that investigates methodologies for disseminating text-image relations throughout the model during the training process. UMT-collapse primarily concentrates on the interaction between text and image through the utilization of multiple Cross-Modal Transformer layers. In contrast, OSCGA-collapse tends to integrate object-level visual features and character-level textual features into neural networks.

**Table 1**

The statistics of Twitter-2015 and Twitter-2017. Pos: Positive, Neg: Negative, Neu: Neutral, AL: Average length.

	Twitter-2015							Twitter-2017						
	Pos	Neg	Neu	Total	Image	Word	AL	Pos	Neg	Neu	Total	Image	Word	AL
Train	928	368	1883	3179	3179	9023	16.7	1508	416	1638	3562	3562	6027	16.2
Dev	303	149	670	1122	1122	4238	16.7	515	144	517	1176	1176	2922	16.3
Test	317	113	607	1037	1037	3919	17.0	493	168	573	1234	1234	3013	16.3

**Table 2**Main results (%) on two Twitter datasets. The best results are highlighted in **atlantisblue**, while suboptimal results are marked in **atlantisgreen**.

Modality	Model	Venue	Twitter-2015			Twitter-2017		
			Precision	Recall	F1	Precision	Recall	F1
Text only	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	64.1	64.2	64.1
	RoBERTa <sup>a</sup>	–	61.8	65.3	63.5	65.5	66.9	66.2
	BART	ACL 2021	62.9	65.0	63.9	65.2	65.6	65.4
Multimodal	TomBERT+UMT	IJCAI 2019; ACL 2020	58.4	61.3	59.8	62.3	62.4	62.4
	RpBERT-collapse	AAAI 2021	49.3	46.9	48.0	57.0	55.4	56.2
	UMT-collapse <sup>a</sup>	ACL 2020	60.4	61.6	61.0	60.0	61.7	60.8
	OSCGA-collapse	ACM MM 2020	63.1	63.7	63.2	63.5	63.5	63.5
	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
	MOCOLNet	TKDE 2023	66.3	67.8	67.1	67.2	68.7	67.9
	VLP-MABSA-M2DF	EMNLP 2023	66.8	68.0	67.3	67.8	68.4	68.1
	UMT-RoBERTa <sup>a</sup>	ACL 2020	61.6	66.4	63.9	65.3	68.2	66.7
	JML-RoBERTa <sup>a</sup>	EMNLP 2021	65.4	64.0	64.7	65.3	66.2	65.8
	CapTrRoBERTa <sup>a</sup>	ACM MM 2021	60.6	66.1	63.2	67.1	67.4	67.3
	CMMT <sup>a</sup>	IPM 2022	64.6	68.7	66.5	67.6	69.4	68.5
Ours	Atlantis	–	65.6	69.2	67.3	68.6	70.3	69.4

<sup>a</sup> Denotes that the results are derived from Yang et al. [11].

- **JML** [9] incorporates a cross-modal relation detection strategy, which is instrumental in initiating the visual gate. This approach subsequently adopts a hierarchical structure, integrating the visual gate with transformer layers to enhance its multimodal processing efficacy.
- **CapTrRoBERTa** [30] employs an image captioning model DETR [52] to convert the image into a textual caption. This is followed by the concatenation of the auxiliary sentence with the original text, which is then input into a RoBERTa model and a CRF layer for JMASA.
- **VLP-MABSA** [10] is a pre-trained vision-language model dedicated to MABSA. It employs five distinct task-specific pre-training strategies designed to effectively model aspects, opinions, and their alignments within the multimodal context.
- **CMMT** [11] is a multi-task learning framework for JMASA. Within this framework, CMMT establishes two auxiliary tasks aimed at guiding the formulation of intra-modal representations. Additionally, it introduces a multimodal gate mechanism to dynamically regulate the influence of visual content in the modeling of inter-modal interactions.
- **MOCOLNet** [53] is a multimodal contrastive learning approach for JMASA, augmented with an auxiliary momentum strategy, to align representations of multimodal data prior to fusion. Additionally, a cross-modal matching technique is employed to extract interactive semantic information between textual content and associated image data.
- **VLP-MABSA-M2DF** [54] is an advancement over VLP-MABSA, incorporates the Multi-Grained Multi-Curriculum Denoising Framework (M2DF). This framework introduces both coarse-grained and fine-grained noise metrics for quantifying noise levels in training images. It then integrates a dual-pronged approach, combining a single and a multiple denoising curriculum, to effectively diminish the adverse effects of image noise.

#### 4.3. Experimental results

The experimental results for JMASA, including precision, recall, and Macro-F1 scores, are shown in Table 2. From Table 2, it can be observed that Atlantis attained commendable performance across both Twitter datasets compared with the text-only and multimodal baselines.

A thorough analysis of the results in Table 2, reveal several discoveries. Firstly, RoBERTa and BART exhibit superior performance within the text-only baseline category and notably exceed the capabilities of several multimodal methods. This finding corroborates the advantages of leveraging a large-scale pre-trained language model. It is noteworthy our proposed Atlantis model significantly surpasses all text-based models by a considerable margin. This outcome suggests that the conversion of visual content into textual space as supplementary textual information within our model is beneficial in our model.

Secondly, combined approaches like TomBERT+UMT and adaptive methods (notated with a “collapse” subscript) typically result in suboptimal outcomes when contrasted with methods that approach MABSA as an end-to-end task. This discrepancy may stem from the division of MABSA into two subtasks, which tends to amplify error propagation between the MATE and the MASC tasks. Moreover, CMMT outperforms all other baselines, highlighting the critical role of employing auxiliary tasks in modeling aspect/sentiment-aware intra-modal features.

Thirdly, our proposed Atlantis shows notable advancements over the majority of existing multimodal methodologies, exceeding the current state-of-the-art (SOTA) CMMT approach in terms of precision, recall, and Macro-F1 on both Twitter-2015 and Twitter-2017. Specifically, Atlantis achieves commendable improvements of 1.0%, 0.5%, and 0.8% on Twitter-2015, and more substantial enhancements of 1.0%, 0.9%, and 0.9% on Twitter-2017, in these respective metrics. Furthermore, it also outperforms the powerful task-specific pre-trained framework VLP-MABSA in all evaluated metrics across both Twitter datasets. In the assessment of the advanced VLP-MABSA-M2DF and the robust baseline MOCOLNet, it is noted that although the Atlantis



**Table 3**

Results (%) of the ablation study on the Atlantis framework. Superior outcomes are emphasized in [atlantisblue](#).

Model	Twitter-2015			Twitter-2017		
	Precision	Recall	F1	Precision	Recall	F1
Atlantis	<b>65.6</b>	<b>69.2</b>	<b>67.3</b>	<b>68.6</b>	<b>70.3</b>	<b>69.4</b>
w/o TAAE	63.1	65.8	64.4	66.1	68.5	67.3
w/o SIAA	63.5	67.5	65.4	64.7	67.7	66.2
w/o cap & scene	62.5	65.9	64.2	66.4	67.9	67.1
w/o RGB-aware att.	61.8	65.2	63.5	64.6	67.3	65.9
rep. aes-cap with cap	62.8	66.8	64.7	66.9	68.3	67.5

model falls short of achieving peak precision on the Twitter-2015 dataset, it surpasses the performance of both VLP-MABSA-M2DF and MOCOLNet in multiple metrics across the two datasets. The superior performance of Atlantis over comparative approaches can be attributed to the following factors: (1) Unlike CMMT, which directly fuses visuals with textual features and employs a multimodal gate for dynamic control of text and image representation inflow, Atlantis converts the input image into textual format through captions and serialized scene graphs. This approach effectively mitigates the representational gap between visual and language modalities and offers a more interpretable alternative to multimodal gates. (2) VLP-MABSA adopts visual aspect-opinion generation (AOG) pre-training task to identify sentimental cues in images, largely dependent on the adjective-noun pair detector DeepSentiBank [55]. However, this approach falls short of capturing implicit emotions conveyed through aesthetic attributes. Conversely, Atlantis utilized a sentiment-aware image aesthetic assessment branch that better guides the framework in recognizing image aesthetic features, thereby having a leg up on capturing implicit sentimental cues from images. (3) In comparison to CapTrRoBERTa, which converts images into textual captions to bridge the modality gap between vision and language, the generated captions are predominantly neutral and lack useful sentiment indicators. By contrast, Atlantis employs emotionally rich aesthetic captions to model inter-dynamics, infusing aesthetic-related sentiment clues. These enhancements are instrumental in augmenting the model's efficacy in sentiment prediction. In conclusion, the gathered observations verify that Atlantis, by bridging the representational gap between different modalities and fusing multiple granular features like captions, scene graphs, and aesthetic attributes, proficiently excels in the extraction of aspect-sentiment pairs.

#### 4.4. Ablation studies

We conducted multiple ablation studies to evaluate the effectiveness of the individual components within the Atlantis framework. Specifically, we investigated the influence of the following components: (1) "w/o TAAE" indicates the removal of the Textual-Vision Alignment Aspect-Sentiment Extraction loss, while retaining the remainder of the branch. (2) "w/o SIAA" signifies the total exclusion of the Sentiment-aware Image Aesthetic Assessment branch. (3) "w/o cap & scene" denotes the elimination of the Caption & Semantic Graph Generation modules. (4) "w/o RGB-aware att." refers to the removal of the High-level RGB-aware Attention Network. (5) "rep. aes-cap with cap" involves substituting the aesthetic caption with a caption generated by the pre-trained CoCa model. The results are shown in [Table 3](#).

**Effects of TAAE:** Auxiliary loss in the Textual-Vision Alignment Aspect-Sentiment Extraction branch was excluded. As indicated in [Table 3](#), the removal of the auxiliary textual-vision alignment supervision task resulted in an approximate decline of 3% across all three evaluation metrics on the Twitter-2015 dataset and around 2% on the Twitter-2017 dataset. These findings suggest that the auxiliary textual-vision alignment supervision significantly boosts the learning of cross-modal representations.

**Effects of SIAA:** [Table 3](#) reveals a decline in the Atlantis model's performance on both datasets following the exclusion of the SIAA branch, with the Twitter-2017 dataset showing a more apparent decrease compared to Twitter-2015. This discrepancy can be attributed to the variation in sample distribution between the datasets. As detailed in [Table 1](#), Twitter-2017 had a higher count of both positive and negative samples than Twitter-2015, with the number of positive samples being approximately 1.5 times greater. Because the SIAA branch is tailored to detect implicit emotional cues in images, its absence results in a greater performance decline in datasets with a higher concentration of sentiment-rich samples.

**Effects of caption & scene graph:** The removal of the caption and semantic graph generation module, while treating the input sentence as *Query*, *Key*, and *Value* simultaneously, significantly impacts the Atlantis model. As shown in [Table 3](#), this change resulted in substantial reductions in precision, recall, and F1 scores: 3.1%, 3.3%, and 3.1% respectively on Twitter-2015, and 2.2%, 2.4%, and 2.3%, respectively, on Twitter-2017. These results highlight the critical role of this module in aiding the Atlantis to mitigate the representational gap for JMASA.

**Effects of RGB-aware attention:** As can be seen from [Table 3](#), omitting the High-level RGB-aware Attention Network results in the most inferior performance across both datasets, surpassing even the negative impact of removing the entire Sentiment-aware Image Aesthetic Assessment (SIAA) branch. As mentioned in [Section 3.4](#), color inherently has the capacity to elicit specific emotional states, with varying hues and saturations that are frequently linked to unique emotional reactions. For example, warm colors may evoke sensations of warmth or excitement, whereas cool tones often imply tranquility or sadness. Without high-level color features, the SIAA branch is less effective in guiding the model's learning of aesthetic features, potentially incorporating unexpected image noise. This even leads to inferior outcomes compared to the complete exclusion of the SIAA branch, highlighting the essential role of high-level color features in image aesthetics assessment and our Atlantis.

**Effects of aesthetic caption:** We further substituted the aesthetic caption with the global caption generated by CoCa. In [Table 3](#), there is a notable decrease in Atlantis's performance post-replacement, with reductions of 2.8%, 2.4%, and 2.6% in precision, recall, and F1 score on Twitter-2015, and 2.7%, 2.0%, and 1.9% in these respective metrics on Twitter-2017. This trend suggests that textual aesthetic captions play a vital role in connecting cross-modal aligned features with aesthetic attributes, which is crucial for enhancing a framework's aesthetic sensitivity.

#### 4.5. In-depth performance analysis

In this subsection, we present a comparative analysis of the Atlantis model using two baseline methods. This includes a detailed breakdown of the F1 scores for each sentiment class and an evaluation of the performance of the model in Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect sentiment classification (MASC).

**Performance breakdown by Sentiment Classes:** [Table 4](#) reveals that Atlantis achieved the highest overall performance on both datasets. Specifically, in the positive sentiment class, Atlantis shows comparable results on both datasets. In the negative class, Atlantis exceeded the two baselines by a large margin of 6.3% and 4.2% for Twitter-2015, and 1.9% and 0.1% for Twitter-2017. In the neutral class, Atlantis demonstrated notable improvements over the baselines by 2.3% and 0.8% on Twitter-2015 and exhibited similar performance on the Twitter-2017 dataset. Drawing from these observations, we discover that Atlantis exhibits superior performance in identifying aspect-sentiment pairs with a negative sentiment tendency compared to the other two baselines. This indicates Atlantis' enhanced capability in extracting richer sentimental cues from multimodal data. These findings and enhancements further corroborate the efficacy of our Atlantis framework.









Image				
Text	(a) RT @ NiallOfficial : another shot from [Santiago] <sub>Positive</sub> ! @ CalAurand	(b) As [Lionel Messi] <sub>positive</sub> turns 30, #HappyBirthday to one of the best footballers in history.	(c) Why rape works as oppression amp why the [UN] <sub>Negative</sub> does not : interesting views from Madeleine . . .	
Attention Visualization				
Aesthetic Caption	I like the lighting and the angle of the shot. the colors are nice.	Great capture great expression great colors great focus great dof great timing	I like the idea but the background is a little distracting.	
Aesthetic Score	0.696	0.654	0.243	
Output	JML-RoBERTa	(Santiago, Neutral 😐) ❌	(Lionel Messi, Neutral 😐) ❌	(UN, Negative 😞) ✅
	CMMT	(Santiago, Neutral 😐) ❌	(Lionel Messi, Positive 😊) ✅	(Madeleine , Positive 😊) ❌
	Atlantis	(Santiago, Positive 😊) ✅	(Lionel Messi, Positive 😊) ✅	(UN, Negative 😞) ✅

Fig. 3. Three instances illustrating predictions made by JML-RoBERTa, CMMT, and Atlantis are presented. The ground truth aspect-sentiment pair is annotated within the text.

Table 4

Performance breakdown (%) for Atlantis by sentiment polarity: It is important to note that the F1 score is reported. Here, POS, NEG, and NEU represent positive, negative, and neutral sentiments, respectively. The best results are distinguished in [atlantisblue](#).

Approaches	Twitter-2015				Twitter-2017			
	POS	NEG	NEU	Overall	POS	NEG	NEU	Overall
UMT-RoBERTa	58.2	53.7	68.1	63.9	66.8	61.1	68.0	66.7
CMMT	<b>63.9</b>	55.2	69.6	66.5	<b>70.4</b>	62.9	68.5	68.5
Atlantis	63.3	<b>60.0</b>	<b>70.4</b>	<b>67.3</b>	70.0	<b>63.0</b>	<b>70.7</b>	<b>69.4</b>

**Performance breakdown by two subtasks:** Additional experiments focusing on MATE and MASC were conducted to further investigate the benefits offered by the Atlantis framework. As shown in [Table 5](#), it can be observed that for the MATE subtask, Atlantis yielded a competitive performance across both datasets. Notably, Atlantis outperformed the current SOTA baseline, CMMT, in terms of Precision and Macro-F1, with improvements of 0.3% and 0.2%, respectively, on the Twitter-2015 dataset. Additionally, Atlantis's performance on the Twitter-2017 dataset was on par with existing benchmarks. In the MASC subtask, Atlantis outperformed all counterparts on both datasets, achieving an accuracy that was 1.4% and 0.4% higher than that of the current SOTA approach CMMT, respectively. The experimental results indicated that Atlantis generally performed better on MASC than MATE. We conjecture that this is due to Atlantis being an aesthetically oriented framework, primarily geared towards extracting visual sentiment cues from aesthetic features. Consequently, it is more adept at predicting sentiment polarity rather than identifying named entities.

#### 4.6. Case study

To delve deeper into the efficacy of Atlantis, we present three illustrative examples accompanied by their aesthetic visual attention visualizations. These serve to compare the predictive performance of two baseline approaches, JML and CMMT, with the outputs from Atlantis.

It can be noted that in [Fig. 3\(a\)](#), while both JML and CMMT accurately identify the aspect “Santiago”, they erroneously predict its associated sentiment. We hypothesize that this is primarily attributed to the scarcity of sentimental cues within the sentence and the weak correlation between the visual content and the text contributing to noise in sentiment prediction. Nevertheless, Atlantis, by converting the image into a textual representation and leveraging the emotionally charged aesthetic caption, “I like the lighting and the angle of the shot. The colors are nice.”, successfully identifies the accurate aspect-sentiment pair.

In example (b), JML does not successfully capture the sentimental clues associated with the identified aspect term “Lionel Messi”. A plausible explanation for this shortcoming is the predominantly blurred background of the image, which likely introduces extraneous visual noise, leading to confusion within the model. Conversely, CMMT, which utilizes adjective-noun pairs (ANP) for visual sentiment understanding, is particularly advantageous in this scenario, given the image's singular subject focus.

In another example (c), the CMMT baseline model inaccurately identified the aspect term. We conjecture that the primary subject of the image, a woman, unintentionally misguides the model's focus toward irrelevant entities. This error arises because the CMMT bases its predictions on the detection of ANP. The aesthetic visual attention for this example is predominantly concentrated in the left-bottom corner, leading to a relatively low aesthetic score of “2.43”. Given this lower aesthetic rating, Atlantis tends to leave a “negative” impression of this

**Table 5**

Performance evaluation (%) of the Atlantis in relation to two subtasks: MATE and MASC. Superior results are highlighted in **atlantisblue**.

Methods	Twitter-2015				Twitter-2017			
	MATE			MASC	MATE			MASC
	Precision	Recall	F1	Acc	Precision	Recall	F1	Acc
UMT-RoBERTa	83.6	<b>88.4</b>	85.9	75.6	91.1	93.2	92.1	73.2
CMMT	83.9	88.1	85.9	77.9	<b>92.2</b>	<b>93.9</b>	<b>93.1</b>	73.8
Atlantis	<b>84.2</b>	87.7	<b>86.1</b>	<b>79.3</b>	91.8	93.2	92.7	<b>74.2</b>

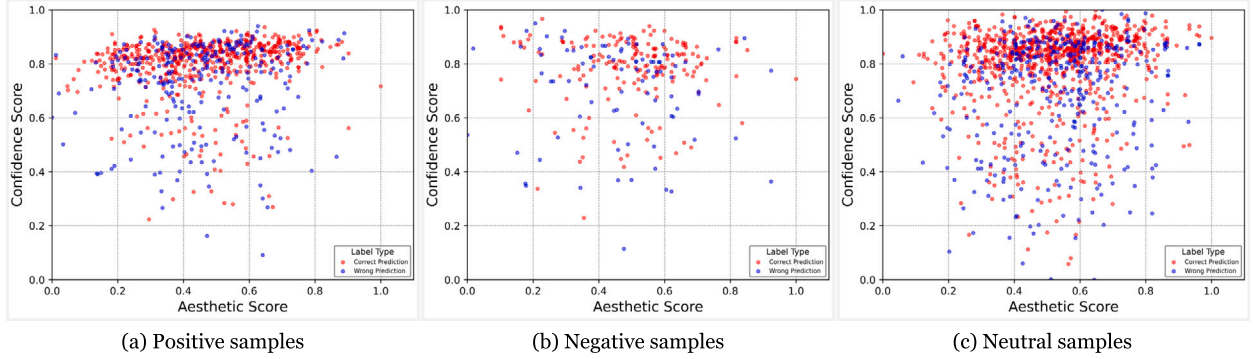
**Fig. 4.** Scatter visualization of confidence scores versus aesthetic scores for different sentimental samples.

photo and shifts its emphasis toward the textual content, thus achieving an accurate prediction.

The case study suggests that in particular scenarios, such as when a sentence-image pair comprises a single one aspect-sentiment pair, images scoring higher in aesthetic quality are more likely to express positive sentiments, and vice versa. This discovery is consistent with the findings presented by Yang et al. [37].

These cases demonstrate that our well-performing approach, Atlantis, effectively captures all correct aspect-sentiment pairs. This is accomplished by merging multiple granular features to align the visual and textual modalities, coupled with the perception of aesthetic attributes.

#### 4.7. Quantitative analysis

We engage in an additional quantitative analysis to probe the intricate interrelationship between the sentiment polarity and the aesthetic attributes of images. In particular, we visualize the sentimental confidence scores against the aesthetic scores for each respective aspect in the test set. The scatter visualization of the sentimental confidence score and the aesthetic score for positive, negative and neutral samples are shown in Fig. 4, respectively. It can be observed from Fig. 4(a) that the red points, which denote correct predictions, are mostly concentrated in the upper half of the plot. Specifically, the distribution of red points indicates that samples with higher aesthetic scores tend to have marginally higher sentimental confidence scores. This trend is particularly discernible in the upper regions of the plot, where there is a noticeable density of red points with elevated confidence scores corresponding to higher aesthetic scores. Despite the presence of several outliers, the general trajectory of the red points hints at a slight positive trendline, suggesting that as aesthetic scores increase, there is a tendency, however slight, for confidence scores to rise in conjunction. In Fig. 4(b), there appears to be a marginal correlation between the sentimental confidence scores and aesthetic scores for negative samples. A trend can be inferred wherein a lower aesthetic score is generally associated with a relatively higher confidence score for a negative sample. Conversely, a higher aesthetic score tends to coincide with a lower confidence score in the negative prediction.

This pattern suggests a slight, yet noticeable, inverse relationship between the perceived aesthetic attributes of an image and the model's confidence in predicting a negative sentiment. As shown in Fig. 4(c), there does not appear to be an apparent, direct correlation between the sentiment confidence score and the aesthetic score, as the red dots are spread throughout the plot without a clear pattern of convergence. This distribution suggests that the model's confidence in assigning a neutral sentiment does not consistently increase or decrease in alignment with the aesthetic valuation of the image. The above observations imply that while aesthetic attributes partially influence sentiment prediction, they are not universally definitive in determining sentiment polarity. In JMASA, it is crucial to consider additional factors, including the semantic content of language and vision, as well as the interactions between different modalities. Therefore, aesthetic attributes, although informative, should be integrated as a component of a comprehensive set of multimodal features, rather than being regarded as an isolated predictor.

#### 4.8. Error analysis

To further assess the limitations of the Atlantis model, we have selected a series of representative error cases for detailed analysis. Fig. 5 exhibits three unsuccessful instances, which have been categorized into three distinct types: (1) Scenarios where aesthetic attributes do not adequately represent the entirety of visual sentiment. (2) Cases involving the extraction of multiple aspect-sentiment pairs with varying sentiments. (3) Cases where aspect term is within a long-span boundaries. In Fig. 5, example (a) illustrates a failed instance arising from limited aesthetic attributes in the image. The aesthetic score and caption offer useless information in discovering sentiment cues within text-based visual content. Furthermore, this example illustrates the unique scenario where the same word conveys different emotional expressions in varied contexts, thereby more challenging for Atlantis.

From Fig. 5(b), it can be seen that the composition of the image, characterized by fluctuating light and shadow and an absence of a definitive subject, hinders a logical and precise aesthetic assessment. Moreover, the aesthetic caption intentionally focuses on "people", prompting the Atlantis model to identify entities pertinent to "individuals", which consequently obstructs the accurate identification

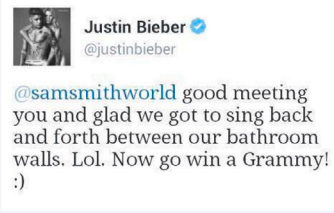

Image			
Text	(a)RT @ TIMELESSBIEBER : when justin bieber says go win a [grammy] <sub>Neutral</sub> , you go win a damn [grammy] <sub>Negative</sub> .	(b)RT @ YahooMusic : Prince is in the house to present Record of the Year to . . . [Gotye] <sub>Positive</sub> ! - Laura F [Grammys] <sub>Neutral</sub> .	(c)RT @ whufc official : PIC : [Academy of Football] <sub>Neutral</sub> graduate [Joe Cole] <sub>Positive</sub> is a [West Ham United] <sub>Positive</sub> player once more !
Attention Visualization			
Aesthetic Caption	I'm not sure what the text is but it's a good idea.	I like the way the lights are lit. the people are a bit distracting.	I'm not sure what the subject is but it's a nice shot. i'm not sure if the background is a bit distracting.
Aesthetic Score	0.361	0.365	0.498
Atlantis Output	(justin bieber, Neutral 😊)	(Prince, Neutral 😊) (Laura, Neutral 😊)	(of, Neutral 😊) (Joe Cole, Positive 😊) (West Ham United, Positive 😊)

Fig. 5. Error instances in the predictions made by Atlantis are presented. The accurate aspect-sentiment pairs are annotated within the text for reference.

of related aspect-sentiment pairs. Consequently, an over-reliance on aesthetic attributes can, in some instances, result in suboptimal performance when identifying multiple aspect-sentiment pairs with varying sentiment polarities.

In another error instance presented in Fig. 5(c), Atlantis is unable to accurately extract the complete aspect term “Academy of Football”. However, it does correctly predict the sentiment associated with “of” and accurately identifies the remaining aspect-sentiment pairs. This case indicates that while Atlantis demonstrates notable ability in sentiment prediction, its effectiveness in span-based aspect term mining is found to be less than satisfactory.

In light of the error analysis, future improvements for our proposed model are anticipated in the following perspectives: (1) Integration of fine-grained linguistic knowledge (e.g., metaphoric expressions [56,57]) into the model; (2) Inclusion of more delicate visual sentiment beyond mere aesthetic attributes; (3) Development of a more sophisticated and robust mechanism for aspect-term extraction.

## 5. Conclusions

This paper presents an Aesthetic-oriented Multiple Granularities Fusion Network (**Atlantis**) for Joint Multimodal Aspect-based Sentiment Analysis. Atlantis is a trident-like framework comprising three branches: textual-vision alignment aspect-sentiment extraction, sentiment-aware image aesthetic assessment, and aesthetic-aware JMASA. Firstly, Textual-vision Alignment Aspect-sentiment Extraction alleviates the representational gap by converting the image into a textual space and fusing it with the input sentence to derive the cross-modal aligned features. Meanwhile, Sentiment-aware Image Aesthetic Assessment extracts the composition and high-level color features and combines them with textual aesthetic features for understanding the aesthetic attributes of images. Finally, the Aesthetic-aware JMASA branch dynamically merges the features of the other two branches to

jointly extract the aspect-sentiment pairs. To the best of our knowledge, we are the first to introduce image aesthetic assessment to this domain, with the aim of exploring the complex relationships between visual aesthetic attributes and multimodal sentiment analysis. The experimental results on two Twitter datasets verify that our proposed Atlantis significantly outperforms numerous state-of-the-art baselines.

## CRedit authorship contribution statement

**Luwei Xiao:** Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Xingjiao Wu:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Junjie Xu:** Writing – review & editing, Software, Data curation. **Weijie Li:** Visualization, Validation, Software, Data curation. **Cheng Jin:** Supervision, Resources, Formal analysis. **Liang He:** Supervision, Project administration, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

This work was supported by the National Key R and D Program of China (2022ZD0161800), the Science and Technology Commission of Shanghai Municipality, China (22DZ2229004), the Fundamental Research Funds for the Central Universities, China, the 2022 East China



Normal University Outstanding Doctoral Students Academic Innovation Ability Improvement Project (YBNLTS2022-005), and the computation is performed in ECNU Multifunctional Platform for Innovation (001). All authors approved the version of the manuscript to be published.

## References

- [1] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep boltzmann machines, *Adv. Neural Inf. Process. Syst.* 25 (2012).
- [2] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, *IEEE Trans. Pattern Anal. Mach. Intell.* 41 (2) (2018) 423–443.
- [3] R. Mao, X. Li, Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2021, pp. 13534–13542.
- [4] H. Liu, W. Wang, H. Sun, A. Rocha, H. Li, Robust domain misinformation detection via multi-modal feature alignment, *IEEE Trans. Inf. Forensics Secur.* (2023).
- [5] L. Yang, J. Wang, J.-C. Na, J. Yu, Generating paraphrase sentences for multi-modal entity-category-sentiment triple extraction, *Knowl.-Based Syst.* 278 (2023) 110823.
- [6] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, *Inf. Fusion* 91 (2023) 424–444.
- [7] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, *Inf. Fusion* 95 (2023) 306–325.
- [8] R. Mao, Q. Liu, K. He, W. Li, E. Cambria, The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection, *IEEE Trans. Affect. Comput.* (2022).
- [9] X. Ju, D. Zhang, R. Xiao, J. Li, S. Li, M. Zhang, G. 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*, 2021, pp. 4395–4405.
- [10] Y. Ling, J. Yu, R. Xia, Vision-language pre-training for multimodal aspect-based sentiment analysis, in: *Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 2149–2159.
- [11] L. Yang, J.-C. Na, J. Yu, Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis, *Inf. Process. Manage.* 59 (5) (2022) 103038.
- [12] A.P. Shimamura, S.E. Palmer, *Aesthetic Science: Connecting Minds, Brains, and Experience*, OUP USA, 2012.
- [13] J. Ke, K. Ye, J. Yu, Y. Wu, P. Milanfar, F. Yang, VILA: Learning image aesthetics from user comments with vision-language pretraining, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10041–10051.
- [14] W. Köhler, Gestalt psychology, *Psychol. Forsch.* 31 (1) (1967) XVIII–XXX.
- [15] J. Ye, J. Zhou, J. Tian, R. Wang, J. Zhou, T. Gui, Q. Zhang, X. Huang, Sentiment-aware multimodal pre-training for multimodal sentiment analysis, *Knowl.-Based Syst.* 258 (2022) 110021.
- [16] X. Zhang, R. Mao, K. He, E. Cambria, Neuro-symbolic sentiment analysis with dynamic word sense disambiguation, in: *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 8772–8783.
- [17] T. Yue, R. Mao, H. Wang, Z. Hu, E. Cambria, KnowNet: Knowledge fusion network for multimodal sarcasm detection, *Inf. Fusion* 100 (2023) 101921.
- [18] L. Xiao, X. Wu, S. Yang, J. Xu, J. Zhou, L. He, Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis, *Inf. Process. Manage.* 60 (6) (2023) 103508.
- [19] J. Zhou, J. Zhao, J.X. Huang, Q.V. Hu, L. He, MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis, *Neurocomputing* 455 (2021) 47–58.
- [20] H. Liu, W. Wang, H. Li, Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement, in: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 4995–5006.
- [21] J. Ye, J. Zhou, J. Tian, R. Wang, Q. Zhang, T. Gui, X.-J. Huang, RethinkingTMSA: An empirical study for target-oriented multimodal sentiment classification, in: *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 270–277.
- [22] Z. Wu, C. Zheng, Y. Cai, J. Chen, H.-f. Leung, Q. 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*, 2020, pp. 1038–1046.
- [23] Q. Zhang, J. Fu, X. Liu, X. Huang, Adaptive co-attention network for named entity recognition in tweets, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, 2018.
- [24] L. Sun, J. Wang, K. Zhang, Y. Su, F. Weng, RpBERT: a text-image relation propagation-based BERT model for multimodal NER, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2021, pp. 13860–13868.
- [25] J. Yu, J. Jiang, R. Xia, Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification, *IEEE/ACM Trans. Audio Speech Lang. Process.* 28 (2019) 429–439.
- [26] D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, G. Zhou, Multi-modal graph fusion for named entity recognition with targeted visual guidance, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2021, pp. 14347–14355.
- [27] L. Yuan, Y. Cai, J. Wang, Q. Li, Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, 2023, pp. 11051–11059.
- [28] J. Yu, J. Jiang, Adapting BERT for target-oriented multimodal sentiment classification, (2019), in: *International Joint Conference on Artificial Intelligence, IJCAI*, 2019, pp. 5408–5414.
- [29] J. Yu, K. Chen, R. Xia, Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis, *IEEE Trans. Affect. Comput.* (2022).
- [30] Z. Khan, Y. Fu, Exploiting BERT for multimodal target sentiment classification through input space translation, in: *ACM International Conference on Multimedia*, ACMMM, 2021, pp. 3034–3042.
- [31] L. Xiao, E. Zhou, X. Wu, S. Yang, T. Ma, L. He, Adaptive multi-feature extraction graph convolutional networks for multimodal target sentiment analysis, in: *2022 IEEE International Conference on Multimedia and Expo, ICME*, IEEE, 2022, pp. 1–6.
- [32] Y. Huang, Z. Chen, J. Chen, J.Z. Pan, Z. Yao, W. Zhang, Target-oriented sentiment classification with sequential cross-modal semantic graph, in: *International Conference on Artificial Neural Networks*, Springer, 2023, pp. 587–599.
- [33] L. Celona, M. Leonardi, P. Napolitano, A. Rozza, Composition and style attributes guided image aesthetic assessment, *IEEE Trans. Image Process.* 31 (2022) 5009–5024.
- [34] W. Li, Y. Wan, X. Wu, J. Xu, L. He, UMAAF: Unveiling aesthetics via multifarious attributes of images, 2023, arXiv preprint arXiv:2311.11306.
- [35] R. Datta, J. Li, J.Z. Wang, Algorithmic inferring of aesthetics and emotion in natural images: An exposition, in: *2008 15th IEEE International Conference on Image Processing*, IEEE, 2008, pp. 105–108.
- [36] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J.Z. Wang, J. Li, J. Luo, Aesthetics and emotions in images, *IEEE Signal Process. Mag.* 28 (5) (2011) 94–115.
- [37] Y. Yang, L. Xu, L. Li, N. Qie, Y. Li, P. Zhang, Y. Guo, Personalized image aesthetics assessment with rich attributes, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19861–19869.
- [38] G. Lan, S. Xiao, J. Yang, Y. Zhou, J. Wen, W. Lu, X. Gao, Image aesthetics assessment based on hypernetwork of emotion fusion, *IEEE Trans. Multimed.* (2023).
- [39] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, Y. Wu, Coca: Contrastive captioners are image-text foundation models, 2022, arXiv preprint arXiv:2205.01917.
- [40] K. Tang, Y. Niu, J. Huang, J. Shi, H. Zhang, Unbiased scene graph generation from biased training, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3716–3725.
- [41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
- [42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems*, NIPS, 2017, pp. 5998–6008.
- [43] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, *Statistics* 1050 (2016) 21.
- [44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Conference on Computer Vision and Pattern Recognition, CVPR*, 2016, pp. 770–778.
- [45] P. O'Donovan, A. Agarwala, A. Hertzmann, Color compatibility from large datasets, in: *ACM SIGGRAPH 2011 Papers*, 2011, pp. 1–12.
- [46] S. He, Y. Zhang, R. Xie, D. Jiang, A. Ming, Rethinking image aesthetics assessment: Models, datasets and benchmarks, in: *IJCAI*, 2022.
- [47] J. Ren, X. Shen, Z. Lin, R. Mech, D.J. Foran, Personalized image aesthetics, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 638–647.
- [48] M. Hu, Y. Peng, Z. Huang, D. Li, Y. 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*, 2019, pp. 537–546.
- [49] G. Chen, Y. Tian, Y. Song, Joint aspect extraction and sentiment analysis with directional graph convolutional networks, in: *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 272–279.
- [50] H. Yan, J. Dai, T. Ji, X. Qiu, Z. 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)*, 2021, pp. 2416–2429.
- [51] J. Yu, J. Jiang, L. Yang, R. Xia, Improving Multimodal Named Entity Recognition Via Entity Span Detection with Unified Multimodal Transformer, *Association for Computational Linguistics*, 2020.



- [52] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: *European Conference on Computer Vision, ECCV*, Springer, 2020, pp. 213–229.
- [53] J. Mu, F. Nie, W. Wang, J. Xu, J. Zhang, H. Liu, MOCOLNet: A momentum contrastive learning network for multimodal aspect-level sentiment analysis, *IEEE Trans. Knowl. Data Eng.* (2023).
- [54] F. Zhao, C. Li, Z. Wu, Y. Ouyang, J. Zhang, X. Dai, M2DF: Multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis, in: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9057–9070.
- [55] T. Chen, D. Borth, T. Darrell, S.-F. Chang, Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks, 2014, arXiv preprint [arXiv:1410.8586](https://arxiv.org/abs/1410.8586).
- [56] R. Mao, X. Li, M. Ge, E. Cambria, MetaPro: A computational metaphor processing model for text pre-processing, *Inf. Fusion* 86 (2022) 30–43.
- [57] R. Mao, X. Li, K. He, M. Ge, E. Cambria, MetaPro online: A computational metaphor processing online system, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2023, pp. 127–135.