

# Vanessa : Visual Connotation and Aesthetic Attributes Understanding Network for Multimodal Aspect-based Sentiment Analysis

Luwei Xiao<sup>♣</sup>, Rui Mao<sup>♣\*</sup>, Xulang Zhang<sup>♠</sup>, Liang He<sup>♠</sup> and Erik Cambria<sup>♠</sup>

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

<sup>♠</sup>College of Computing and Data Science, Nanyang Technological University, Singapore

louisshaw@stu.ecnu.edu.cn, lhe@cs.ecnu.edu.cn, {rui.mao, xulang.zhang, cambria}@ntu.edu.sg

## Abstract

Prevailing research concentrates on superficial features or descriptions of images, revealing a significant gap in the systematic exploration of their connotative and aesthetic attributes. Furthermore, the use of cross-modal relation detection modules to eliminate noise from comprehensive image representations leads to the omission of subtle contextual information. We present Vanessa, a visual connotation and aesthetic Attributes understanding network for multimodal aspect-based sentiment analysis. It incorporates a multi-aesthetic attributes aggregation (MA<sup>3</sup>) module that models intra- and inter-dependencies among bi-modal representations as well as emotion-laden aesthetic attributes. Moreover, we devise a self-supervised contrastive learning framework to explore the pairwise relevance between images and text via the Gaussian distribution of their CLIP scores. By dynamically clustering and merging multimodal tokens, Vanessa effectively captures both implicit and explicit sentimental cues. Extensive experiments on two widely adopted benchmarks verify Vanessa’s effectiveness.

## 1 Introduction

Multimodal aspect-based sentiment analysis (MABSA) marks a pivotal advancement in sentiment analysis by enhancing the machine’s ability to interpret human emotions, thus attracting growing scholarly interest (Susanto et al., 2020; Cambria et al., 2013). MABSA aims to identify aspect-sentiment pairs within sentences given image-text pairs. Examples of MABSA are shown in Fig. 1. The primary challenge of MABSA lies in leveraging image data to enrich textual sentiment analysis. Existing approaches typically fall into two major categories: (i) segmenting the image into multiple visual regions or extracting prominent visual objects to facilitate inter-dynamic modeling with textual sequences through tailored

\* Corresponding author.

Image		
Text	(a) What a wonderful [weather] <sub>loc</sub> !	(b) [Neymar] <sub>per</sub> has come to [saudi arabia] <sub>loc</sub> .
Aes/CLIP	0.35/0.52	0.57/0.65
Impr.	I feel a sense of awe and fearful when I look at this image. The dark clouds and the cityscape create a sense of awe and fearful.	I believe this image conveys a sense of excitement. It makes me think about the power of teamwork.
Aes-Cap	The lighting of the image is very powerful and the city is very powerful.	The man’s smile and the hands of his head convey a sense of happiness. The lighting of the room creates a sense of warmth.

Figure 1: Examples for MABSA, with aspect-sentiment pairs highlighted in the text. "Aes" and "CLIP" represent the aesthetic and CLIP scores (ranging from 0 to 1). "Impr" and "Aes-Cap" denote the impression and aesthetic caption generated by our fine-tuned BLIP.

fusion mechanisms (Xu et al., 2019; Yu and Jiang; Yu et al., 2019, 2020, 2022a,b; Zhang et al., 2021; Ling et al., 2022; Yang et al., 2022b; Zhou et al., 2023); (ii) translating the image into textual space and subsequently establishing linkages between primary text sequences and supplementary sentences (Khan and Fu, 2021; Yang et al., 2022a; Liu et al., 2022; Xiao et al., 2023; Wang et al., 2023).

Despite promising outcomes, the majority of studies confront two challenges. Firstly, they neglect the implicit emotions evoked by connotation and aesthetic elements of visual imagery. Psychologically, images serve as powerful stimuli that activate cognitive and perceptual pathways, eliciting affective responses through their portrayal of contextual, symbolic, and aesthetic elements (Lang and Bradley, 2007; Barrett and Bar, 2009). For example, a beautifully composed photograph with balanced colors and pleasing symmetry is likely to evoke positive emotions such as joy and admiration. Second, prevalent approaches utilize cross-modal relation detection modules to filter noise from holistic image representations, which can inadvertently eliminate subtle contextual cues.

To address the aforementioned issues, we introduce Vanessa, a model crafted to decipher the sentimental expressions conveyed through visual connotations and aesthetics. Additionally, Vanessa explores the semantic correlations between images and their associated textual content. The model comprises three primary components: the Multi-Aesthetic Attributes Aggregation (MA<sup>3</sup>) module, the Self-supervised Contrastive Learning for Image-Text Relevance (SSL-ITR), and the Dynamic Token Merge (DTM) module. Initially, the MA<sup>3</sup> generates emotionally rich multimodal representations and constructs a task-specific, aesthetic-aware multimodal dependency matrix. These are then processed through graph convolutional networks (GCNs) to adaptively model the intra- and inter-dynamics of aesthetic-aware emotions across modalities. Subsequently, SSL-ITR samples positive and negative image-text pairs based on the Gaussian distribution of their CLIP scores, thus enabling the model to selectively focus on both visual and textual information or primarily on textual content. Lastly, DTM dynamically models the aesthetic-aware multimodal features at both explicit and implicit levels. Experimental results indicate that Vanessa outperforms the state-of-the-art baseline by 1.2% and 0.9% in averaged F1 scores on two widely used Twitter datasets.

In a nutshell, we contribute the following: (1) To the best of our knowledge, this is the first study to explore the utilization of implicit emotions evoked by the connotation and aesthetic attributes of images to model complex intermodal relationships, while simultaneously learning sentimental cues at both explicit and implicit levels within MABSA; (2) We tailor a self-supervised contrastive learning framework to enable the model to grasp the semantic pairwise relevance of image-text pairs based on their CLIP score and Gaussian distribution; (3) We conducted comprehensive experiments and rigorous analyses on two widely recognized public datasets. The experimental results indicate that Vanessa achieves state-of-the-art performance.

## 2 Related Work

**Multimodal Aspect-based Sentiment Analysis.** Sentiment analysis is a widely studied field that aims to understand and quantify human emotions and opinions across various contexts (Zhang et al., 2023; Lu et al., 2023; Liu et al., 2023a; Mao et al., 2023; Cambria et al., 2024; Du et al., 2024).

With the exponential growth of multimodal content on social media (Zhang et al., 2024b), MABSA has gained significant attention (Liu et al., 2022; Mao and Li, 2021; Yue et al., 2023; Fan et al., 2024; Yang et al., 2024). The MABSA task consists of two sub-tasks: Multimodal Aspect Term Extraction (MATE) and Multimodal Aspect-based Sentiment Classification (MASC). MATE (Yang et al., 2023) aims at extracting all relevant aspect terms from the textual content given an image-text pair, while MASC (Zhou et al., 2021; Zhang et al., 2022) focuses on predicting the sentiment polarities associated with these extracted aspects. Recently, a group of studies successfully integrated these two sub-tasks into a unified framework, effectively streamlining the process of achieving MABSA (Ju et al., 2021; Yang et al., 2022b; Ling et al., 2022; Mu et al., 2023; Zhao et al., 2023a; Xiao et al., 2024; Cambria et al., 2023). However, most machine learning-based methods do not pay enough attention to the implicit emotions evoked by the connotation and aesthetic elements of visual imagery. Moreover, employing various cross-modal relation detection modules to filter noise from holistic image representations may inadvertently result in the loss of subtle contextual cues (Hu et al., 2022; Yan et al., 2023).

**Multimodal Representation Learning.** Multimodal representation learning has emerged as a critical research area (Liu et al., 2023b). Recent years have witnessed the development and widespread application of sophisticated multimodal learning techniques across multiple domains (Guo et al., 2023; Zhang et al., 2024a; Guo et al., 2024). A prominent example is CLIP (Radford et al., 2021), which is pre-trained on the WIT (WebImageText) dataset. Distinct from conventional vision models, CLIP concurrently trains an image encoder and a text encoder, thereby learning rich semantic relationships between linguistic and visual modalities. The CLIP score (Hessel et al., 2021) quantifies the semantic alignment between images and captions by computing the cosine similarity between the image embedding and the caption embedding using a pre-trained CLIP model. Similarly, BLIP (Li et al., 2022), a comprehensive vision-language framework, leverages knowledge distillation on captions to augment its performance. It achieves state-of-the-art results across various tasks and demonstrates exceptional zero-shot performance.

Building upon these advancements, BLIP-2 (Li et al., 2023), an enhanced vision-language model developed through an extensive pre-training strategy, exhibits a wide array of zero-shot image-to-text capabilities. In this study, we leverage the robust semantic alignment capabilities of CLIP to model the pairwise relationships between text and images. Furthermore, we employ BLIP for fine-tuning purposes to generate aesthetic captions imbued with rich emotional connotations.

### Visual Connotation & Image Aesthetic Analysis.

Visual connotation involves emotive and aesthetic meanings an image conveys beyond its explicit content, engaging viewers on deeper interpretative levels (Arnheim, 1954; Berger, 1972). The aesthetics of an image relate to its subjective evaluation or the admiration of its beauty (Ramachandran and Hirstein, 1999). Previous research has concentrated on the aesthetic score (see Fig. 1), a quantitative metric that evaluates the visual attractiveness of an image (Zeng et al., 2019). A higher aesthetic score is indicative of enhanced aesthetic quality. Recent scholarly efforts emphasized encouraging vision models to engage in generating visual metaphors and aesthetic-related captions (Akula et al., 2023; Chakrabarty et al., 2023; Ke et al., 2023). More recently, Kruk et al. (2023) presented a connotation-rich dataset termed Impressions, which enables the exploration of emotions, thoughts, and beliefs that images invoke, as well as an analysis of the aesthetic elements that trigger these responses. In this study, we employ visual connotation and aesthetic attributes to comprehensively capture the sentimental cues within visual content for MABSA. To the best of our knowledge, this is the inaugural effort to integrate visual connotation and aesthetic attributes into the MABSA framework.

## 3 Method

**Task Definition.** Given a image-text pair containing image  $V$  and sentence  $S = (w_1, w_2, \dots, w_n)$ , our objective is to predict the corresponding aspect-sentiment sequence  $Y = (y_1, y_2, \dots, y_n)$ . Here,  $y_i \in \{\text{B-POS}, \text{I-POS}, \text{B-NEG}, \text{I-NEG}, \text{B-NEU}, \text{I-NEU}\} \cup \{\text{O}\}$ . In this case, B denotes the beginning token of an aspect term; I refers to tokens that are part of the aspect term; O denotes tokens that are outside any specific aspect. POS, NEU, and NEG are the abbreviations of positive, neutral, and negative sentiment associated with aspect terms (Valdivia et al., 2018).

**Model Overview.** Fig. 2 illustrates the overall architecture of our proposed Vanessa, which comprises three main modules: the Multi-Aesthetic Attributes Aggregation module (MA<sup>3</sup>), the Self-Supervised Contrastive Learning for Image-Text Relevance module (SSL-ITR), and the Dynamic Token Merge module (DTM). Firstly, we fine-tune the BLIP on the Impression dataset to generate impression and aesthetic captions for the images. The image-text pairs and these auxiliary sentences are then fed into the MA<sup>3</sup> module, combined with aesthetic and CLIP scores, to construct an aesthetic-aware multimodal graph for modeling multimodal and textual features. Subsequently, multimodal features are passed into the SSL-ITR to learn the semantic pairwise image-text relationship, based on the CLIP score and its Gaussian distribution. Finally, the DTM module clusters and merges multimodal and textual features using a KNN-based algorithm and self-attention, capturing implicit and explicit sentimental cues for MABSA.

**Auxiliary Sentence Generation.** Initially, we fine-tuned a pre-trained BLIP (Li et al., 2022) using the Impression dataset (Kruk et al., 2023) to enable it to generate impression and aesthetic captions. For a given image  $V \in \mathbb{R}^{3 \times H \times W}$ , we then input it into the fine-tuned BLIP model to produce its corresponding impression and aesthetic captions, resulting in two auxiliary, emotion-rich sentences.

### 3.1 Multi-Aesthetic Attributes Aggregation module (MA<sup>3</sup>)

MA<sup>3</sup> is crucial for capturing complex sentimental relationships in multimodal data. Fig. 3 displays details of MA<sup>3</sup>. It unifies visual and textual features, impressions, and aesthetic attributes into a cohesive graph, allowing for precise modeling of sentimental expressions. The visual features of the image  $V_f \in \mathbb{R}^d$  are obtained using CLIP (Radford et al., 2021), and the hidden features of the input sentence  $H^s = (h_1^s, h_2^s, \dots, h_{N_s}^s) \in \mathbb{R}^{N_s \times d}$ , impression  $H^r = (h_1^r, h_2^r, \dots, h_{N_r}^r) \in \mathbb{R}^{N_r \times d}$ , and aesthetic caption  $H^a = (h_1^a, h_2^a, \dots, h_{N_a}^a) \in \mathbb{R}^{N_a \times d}$  are derived using RoBERTa (Liu et al., 2019). Graphs offer a unified and consistent framework for representing and integrating diverse data types. Then, we developed a task-specific aesthetic-aware multimodal graph (AMG) for each sample. The nodes  $H^g$  of the AMG comprise the concatenated hidden representations of the input sentence, visual

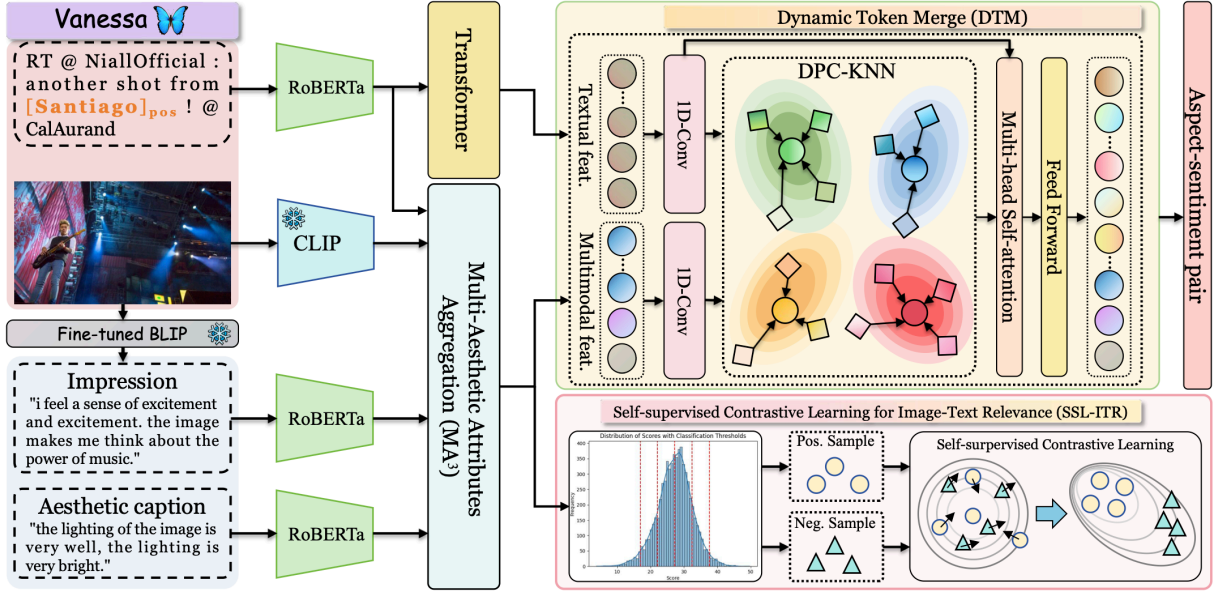


Figure 2: Overview of the Vanessa framework, covering a three-stage process: (1) multi-aesthetic attributes aggregation, (2) self-supervised contrastive learning for image-text relevance, and (3) dynamic token merge.

content, impression, and aesthetic caption:  $H^g = (h_1^g, h_2^g, \dots, h_{N_g}^g) = (h_1^s, \dots, h_{N_s}^s; V_f; h_1^r, \dots, h_{N_r}^r; h_1^a, \dots, h_{N_a}^a)$ .  $N_g = N_s + 1 + N_r + N_a$  denotes the length of the hidden representations. We define  $A \in \mathbb{R}^{N_g \times N_g}$  as the adjacency matrix of the AMG, with its elements initially set to zero. To clarify the construction of the AMG, we divide the procedure into two steps: 1) setting edges to model intra-dependency and 2) setting edges to model inter-dependency.

### 3.1.1 Model Intra-dependency

This sub-module improves the understanding of intra-dependencies within text, which is essential for accurately capturing the relationships among entities and their opinion words. Specifically, we employ the syntactic dependency tree<sup>1</sup> for text, combined with a self-attention mechanism, to assign weights to the edges between words tagged with specific part-of-speech (POS) (Xiao et al., 2022) for the sentence subgraph  $A^S \in \mathbb{R}^{N_s \times N_s}$  as:

$$A_{i,j}^S = \begin{cases} att(h_i^s, h_j^s), & \text{if } \mathcal{D}_{i,j}, (h_i^{s(p)}, h_j^{s(p)}) \in POS, \end{cases} \quad (1)$$

where  $att$  denotes the self-attention mechanism (Vaswani et al., 2017).  $\mathcal{D}_{i,j}$  indicates that there is a syntactic dependency between words  $h_i^s$  and  $h_j^s$ .  $h_i^{s(p)}, h_j^{s(p)}$  are the POS tags for the  $i$ -th and  $j$ -th words, respectively.  $POS = [nouns, adj, vb, cc, rb]$ .

<sup>1</sup>spaCy toolkit (<https://spacy.io>).

The subgraphs for impression  $A^R \in \mathbb{R}^{N_r \times N_r}$  and aesthetic caption  $A^C \in \mathbb{R}^{N_a \times N_a}$  are derived via the similar operation. Since the visual feature  $V_f$  is a feature vector, we set the intra-dependency as 1 to it.

### 3.1.2 Model Inter-dependency

To model the inter-dependency and capture explicit/implicit sentiment cues across different modalities, it is essential to: (1) track the semantic correlations between these modalities and (2) infer the sentiment expressions within the associated textual content. We define six inter-dependencies: visual-sentence, visual-impression, visual-aesthetic, sentence-impression, sentence-aesthetic, and impression-aesthetic.

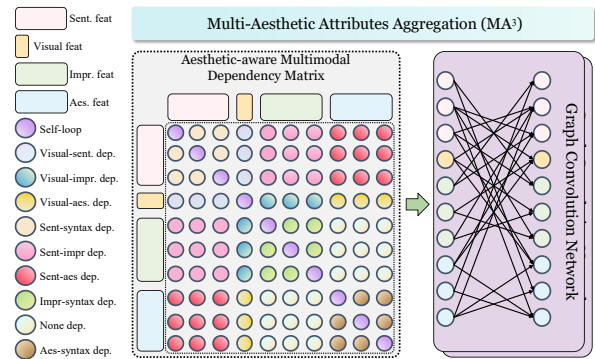


Figure 3: Details of the Multi-Aesthetic Attributes Aggregation (MA<sup>3</sup>). "Sent" is the input sentence, "Impr" refers to the impression, and "Aes" indicates the aesthetic caption.

We model cross-modal visual-sentence dependencies subgraph  $A^{VS} \in \mathbb{R}^{1 \times N_s}$  by first obtaining a skew-symmetric matrix  $S = S_0 - S_0^\top$  and map it to the special orthogonal group  $R^f = \sum_{n=0}^{\infty} \frac{S^n}{n!}$ , a Lie group (Humphreys, 2012).  $S_0$  is a randomly initialized matrix.  $R^f$  is the rotation matrix. Meanwhile, given a random matrix  $B \in \mathbb{R}^{d \times d}$ , the  $k$ -th column of the orthogonal matrix  $Q$ , denoted  $q_k$ , is obtained via the Gram-Schmidt process (Leon et al., 2013):

$$q_k = \frac{b_k - \sum_{j=1}^{k-1} \left( \frac{q_j \cdot b_k}{q_j \cdot q_j} \right) q_j}{\left\| b_k - \sum_{j=1}^{k-1} \left( \frac{q_j \cdot b_k}{q_j \cdot q_j} \right) q_j \right\|}, \quad (2)$$

where  $b_k$  is  $k$ -th column of the matrix  $B \in \mathbb{R}^{d \times d}$ . Then, we form the composite transformation matrix  $C = Q^{-1}R^fQ$  to rotate and align features from the input sentence  $H^s$  tagged with specific POS and visual  $V^f$  modalities while preserving their inherent data structure and characteristics:

$$V' = CV^f, H' = Ch_i^s, h_i^s \in POS^{vs}, \quad (3)$$

where  $V'$  and  $H'$  are transformed feature representations for vision and text, respectively. Then, we calculate the alignment loss  $\mathcal{L}_{align} = \|V' - H'\|_F^2$  between them.  $\|\cdot\|_F$  denotes the Frobenius norm. Finally, the value assigned to this edge is determined by the product of the CLIP score for the image-text pair and the Gaussian similarity between the transformed features as follows:

$$Gaussian = \exp \left( -\rho \sum_{i,j} (V'_{ij} - H'_{ij})^2 \right), \quad (4)$$

$$A_i^{VS} = \left\{ clip * Sim(V^f, h_i^s), \quad \text{if } h_i^s \in POS^{vs}, \quad (5) \right.$$

where  $clip$  denotes the corresponding CLIP score of the image-text pair.  $Gaussian$  is the calculation of Gaussian similarity and  $-\rho$  serves as the decay parameter within the Gaussian function.  $Sim$  indicates the whole calculation process of Gaussian similarity from skew-symmetric matrix to equation (4).  $POS^{vs} \in [nouns]$ . The visual-impression subgraph  $A^{VI} \in \mathbb{R}^{1 \times N_r}$  and the visual-aesthetic subgraph  $A^{VA} \in \mathbb{R}^{1 \times N_a}$  are constructed via similar process:

$$A_i^{VI} = \left\{ clip * Sim(V^f, h_i^r), \quad \text{if } h_i^r \in POS^{vi}, \quad (6) \right.$$

$$A_i^{VA} = \left\{ aes * Sim(V^f, h_i^a), \quad \text{if } h_i^a \in POS^{va}, \quad (7) \right.$$

where  $POS^{vi} \in [adj, rb, verbs]$  and  $POS^{va} \in [nouns, adj, verbs, rb]$ .  $aes$  is the aesthetic score of the image.

For the uni-modal inter-dependency sentence-impression subgraph  $A^{SI} \in \mathbb{R}^{N_r \times N_s}$  and the sentence-aesthetic subgraph  $A^{SA} \in \mathbb{R}^{N_a \times N_s}$ , we calculate the attention score between the corresponding textual representations, and multiply it by the CLIP score of the image-text pair  $A^{SI} = clip * att(H^s, H^r)$  and the aesthetic score of the image  $A^{SA} = aes * att(H^s, H^a)$ . The impression-aesthetic dependency is set to zero, as the correlation between these two auxiliary sentences provides limited information for this task. Finally, we establish a self-loop for each node,  $A_{i,i} = 1$ , in the AMG, resulting in the complete AMG  $A \in \mathbb{R}^{N_g \times N_g}$  as an undirected graph:

$$A = \begin{pmatrix} A^S & A^{VS} & A^{SI} & A^{SA} \\ (A^{VS})^T & 1 & A^{VI} & A^{VA} \\ (A^{SI})^T & (A^{VI})^T & A^R & 0 \\ (A^{SA})^T & (A^{VA})^T & 0 & A^C \end{pmatrix} \quad (8)$$

### 3.1.3 Multimodal Graph Convolution

This sub-module is vital for capturing and modeling the intra- and inter-dynamics of aesthetic-aware sentimental features across different modalities. We feed the task-specific AMG  $A \in \mathbb{R}^{N_g \times N_g}$  and the corresponding node representations  $H^g \in \mathbb{R}^{N_g \times d}$  into multi-layer GCNs to adaptively model the intra- and inter-dynamics of aesthetic-aware sentimental features across modalities:

$$G^l = \text{ReLU} \left( \hat{A} G^{l-1} W^l + b^l \right), \quad (9)$$

where  $\hat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ ,  $D$  denotes the degree matrix of  $A$  with  $D_{ii} = \sum_j A_{ij}$ .  $G^{l-1}$  represents the hidden features from the preceding GCN layer.  $W^l$  and  $b^l$  are the trainable parameters in the  $l$ -th GCN layer. The input for the first GCN layer is the concatenated multimodal hidden representations, denoted as  $G^0 = H^g$ . The multimodal feature  $G^L = \{g_i\}_{i=1}^{N_g}$  is derived from this module. Meanwhile, the hidden features of input sentence  $H^s$  are fed into the Transformer encoder to model the textual features  $H^t = \{h_i^t\}_{i=1}^{N_s}$ .

## 3.2 SSL for Image-text Relevance

We propose a Self-supervised Contrastive Learning for Image-Text Relevance (SSL-ITR) module, which models the semantic pairwise image-text relationship by utilizing the CLIP score and its Gaussian distribution. Conventional contrastive learning helps to distinguish the hidden states of positive and negative samples (Liang et al., 2024; Mao et al., 2024).

SSL-ITR dynamically prioritizes visual or textual modalities, improving the model’s ability to discern and utilize relevant multimodal features for better performance. The CLIP score (see examples in Fig. 1) is a quantitative metric that evaluates the semantic alignment between an image and its corresponding sentence. Initially, we use CLIP (Radford et al., 2021) to obtain CLIP scores for all image-text pairs in the dataset and calculate their mean and standard deviation. Based on the mean value, standard deviation, and twice the standard deviation, we categorize these CLIP scores into six relevance level labels  $\mathcal{R} \in \{r_0, r_1, r_2, r_3, r_4, r_5\}$ , tagging the image-text pairs with their corresponding relevance levels. For multimodal features  $\{g_i\}_{i=1}^{N_b}$  within each mini-batch  $\mathcal{B}$  ( $N_b$  being the size of the mini-batch), the *anchor*  $g_i$  is the sample with the highest CLIP score. If the relevance level  $\mathcal{R}_j$  of  $g_j$  exceeds a specified threshold (e.g.,  $\mathcal{R}_j \geq r_3$ ), then the sample is considered a *positive* pair; otherwise, it is a *negative* pair. The contrastive loss for all positive pairs is computed as follows:

$$\mathcal{L}_{\text{con}} = \frac{-1}{N_b} \sum_{g_i \in \mathcal{B}} \log \frac{\sum_{j \in \mathcal{B} \setminus i} \mathbb{I}_{[\mathcal{R}_j \geq \mathcal{R}]} \exp(f(g_i, g_j) / \tau)}{\sum_{j \in \mathcal{B} \setminus i} \exp(f(g_i, g_j) / \tau)} \quad (10)$$

where  $\mathbb{I}_{[\mathcal{R}_j \geq \mathcal{R}]} \in \{0, 1\}$  is an indicator that evaluates to 1, if  $\mathcal{R}_j$  is higher than the specified relevance level.  $f(g_i, g_j) = g_i^\top g_j / \|g_i\| \|g_j\|$  denotes the cosine similarity between  $g_i$  and  $g_j$ .  $\tau$  indicates the temperature parameter.

### 3.3 Dynamic Token Merge

This module is essential for adeptly selecting and merging aesthetic-aware and emotionally-rich features at both implicit and explicit levels. We employ DPC-KNN (Du et al., 2016; Jin et al., 2023), a KNN-based density peaks clustering algorithm, to dynamically select aesthetic-aware and emotionally-rich features by clustering the mixed representations  $M = (m_1, m_2, \dots, m_{N_m}) = (g_1, g_2, \dots, g_{N_g}, h_1^t, h_2^t, \dots, h_{N_s}^t)$  of multimodal and textual features ( $N_m = N_g + N_s$ ). We first pass the mixed representations to a one-dimensional convolutional layer, and compute the local density  $\psi_i = \exp(-\frac{1}{K} \sum_{m_k \in \text{KNN}(m_i)} \|m_k - m_i\|^2)$  of each token based on its  $K$ -nearest neighbors.  $\text{KNN}(m_i)$  indicates the  $K$ -nearest neighbors of  $m_i$ . Then the distance index  $\gamma_i$  of each token  $m_i$  is given by:

Dataset	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
# POS	928	303	317	1508	515	493
# NEU	1883	670	607	1638	517	573
# NEG	368	149	113	416	144	168
# Total	3179	1122	1037	3562	176	1234

Table 1: The statistics of two Twitter datasets. Pos: Positive, Neg: Negative, Neu: Neutral.

$$\gamma_i = \begin{cases} \min_{j: \psi_j > \psi_i} \|m_k - m_i\|^2, & \text{if } \exists j \text{ s.t. } \psi_j > \psi_i \\ \max_j \|m_k - m_i\|^2, & \text{otherwise,} \end{cases} \quad (11)$$

where  $\psi$  refers to the local density of tokens, and  $\gamma$  is the distance from other high-density tokens. Subsequently, tokens with relatively high  $\psi_i \times \gamma_i$  values are identified as cluster centers. The remaining tokens are assigned to the nearest cluster center according to Euclidean distances. We represent each cluster by the weighted average of its tokens. The textual features are then used as Q, and the weighted average tokens are used as K and V in a multi-head attention module to generate the final feature representation  $H^f \in \mathbb{R}^{N_s \times d}$ . Finally, the  $H^f$  is passed into a CRF layer to predict the aspect-sentiment sequence  $Y$ :

$$p(Y) = \frac{\exp(s(H^f, Y))}{\sum_{\hat{Y} \in Y_{H^f}} \exp(s(H^f, \hat{Y}))}, \quad (12)$$

$$s(H^f, Y) = \sum_{i=0}^{N_s} T_{y_i, y_{i+1}} + \sum_{i=1}^{N_s} h_i^f \cdot W^{y_i}, \quad (13)$$

where  $T$  is the transition matrix and  $Y_{H^f}$  denotes all possible label sequences for the input sample. The trainable matrix  $W^{y_i}$  is utilized to compute the emission score from the token  $h_i^f$  to the label  $y_i$ .

### 3.4 Model Training

The overall loss is the combination of task loss, alignment loss, and contrastive loss:

$$\mathcal{L}_{\text{total}} = -\log p(Y) + \alpha \mathcal{L}_{\text{align}} + \beta \mathcal{L}_{\text{con}}, \quad (14)$$

where  $\alpha$  and  $\beta$  are tradeoff hyper-parameters.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets and Evaluation Metrics.** We opt for two public multimodal datasets Twitter2015 and Twitter2017 (Yu et al., 2019) to evaluate the performance of our Vanessa. An overview of both

datasets is shown in Table 1. Moreover, We evaluate the performance of our proposed Vanessa on this task using three standard evaluation metrics: Micro-F1 score (F1), Precision (P), and Recall (R).

**Implementation Details.** We employ RoBERTa (Liu et al., 2019) to initialize the word representations and use CLIP (Radford et al., 2021) to extract visual features and generate the CLIP score. The model is trained for 40 epochs with a batch size of 16 on the MABSA dataset. Both learning rates are set to  $3 \times 10^{-5}$ , and the hidden sizes are set to 768. The hyper-parameters  $\alpha$  and  $\beta$  are set to 1 and 0.5, respectively. Additionally, we stack two layers in the GCNs. The aesthetic score of the image is generated using VILA (Ke et al., 2023).

**Compared Baselines.** (1) Text-based baselines: RoBERTa (Liu et al., 2019), BART (Yan et al., 2021), and D-GCN (Chen et al., 2020). (2) Multimodal baselines: UMT+TomBERT (Yu and Jiang; Yu et al., 2020), OSCGA+TomBERT (Yu and Jiang; Wu et al., 2020), OSCGA-collapse (Wu et al., 2020), RpBERT-collapse (Sun et al., 2021), UMT-collapse (Yu et al., 2020), JML (Ju et al., 2021), VLP-MABSA (Ling et al., 2022), CMMT (Yang et al., 2022b), MOCOLNet (Mu et al., 2023), VLP-MABSA-M2DF (Zhao et al., 2023a), Atlantis (Xiao et al., 2024), and AoM (Zhou et al., 2023).

## 4.2 Main Results

The main experimental results are presented in Table 2. Firstly, we observe that pre-trained language models RoBERTa and BART exhibit superior performance within the text-only baselines. Besides, the multimodal baselines generally outperform the text-based methods (Cambria, 2024). Secondly, among multimodal baselines, methods that integrate different pipelines into one framework lag significantly behind unified frameworks. Last but not least, Vanessa achieved state-of-the-art performance, surpassing all baselines. It improved the F1 score by 1.2% and 0.9%, and precision by 1.8% and 1.1%, compared to the second-best model, AoM, on two datasets. These results verify the effectiveness of incorporating visual connotations and aesthetic attributes, as well as learning semantic relevance between text and image via CLIP scores.

## 4.3 Ablation Study

Ablation study results are presented in Table 3.

**Aesthetic-aware Multimodal Graph.** We remove the AMG and the corresponding nodes for impression and aesthetic caption in  $H^g$ . The significant performance degradation across all evaluation metrics demonstrates that the intra- and inter-dependencies among multimodal features, visual connotations, and aesthetic attributes modeled by AMG are crucial for capturing complex sentimental relationships across modalities.

**Impression.** We remove the impression from the nodes  $H^g$  and its corresponding dependencies in AMG, resulting in their exclusion from Vanessa. The performance decline observed in Table 3 verifies the importance and effectiveness of incorporating visual connotations to extract implicit sentimental cues from images.

**Aesthetic Caption.** Similar to the removal of the impression, we discard the aesthetic caption from the nodes  $H^g$  and its corresponding dependencies in AMG. As can be seen from Table 3, this removal results in serious performance degradation, demonstrating that extracting explicit sentimental cues from the visual modality through aesthetic attributes enhances the understanding of visual elements, so as to improve MABSA performance.

**SSL-ITR.** Furthermore, Table 3 reveals that the ablation of the SSL-ITR module significantly degrades performance across all metrics. This finding verifies the importance of the proposed self-supervised contrastive learning strategy in comprehending the semantic relevance of image-text pairs.

**Dynamic Token Merge.** We substitute the DTM module with a simple concatenation of multimodal and textual features. As shown in Table 3, the removal of the DTM module results in performance degradation, which indicates that integrating aesthetic-aware multimodal features with textual features through clustering the most representative neighboring features is effective.

## 4.4 Analysis of Contrastive Learning

We investigate the impact of self-supervised contrastive learning for image-text relevance in Vanessa on representation quality. Specifically, we record training checkpoints from the “w/o SSL-ITR” variants and the complete Vanessa, and visualize the alignment and uniformity metrics of these checkpoints in Fig. 4. As demonstrated by Wang and Isola (2020), lower  $\mathcal{L}_{align}$  and  $\mathcal{L}_{uniform}$  lead to better performance.

Methods	Twitter2015			Twitter2017			
	P	R	F1	P	R	F1	
Text-based	RoBERTa (Liu et al., 2019)	61.8	65.3	63.5	65.5	66.9	66.2
	D-GCN* (Chen et al., 2020)	58.3	58.8	59.4	64.1	64.2	64.1
	BART* (Yan et al., 2021)	62.9	65.0	63.9	65.2	65.6	65.4
Multimodal	UMT+TomBERT* (Yu and Jiang; Yu et al., 2020)	58.4	61.3	59.8	62.3	62.4	62.4
	OSCGA+TomBERT* (Yu and Jiang; Wu et al., 2020)	61.7	63.4	62.5	63.4	64.0	63.7
	OSCGA-collapse* (Wu et al., 2020)	63.1	63.7	63.2	63.5	63.5	63.5
	RpBERT-collapse* (Sun et al., 2021)	49.3	46.9	48.0	57.0	55.4	56.2
	UMT-collapse* (Yu et al., 2020)	61.0	60.4	61.6	60.8	60.0	61.7
	JML* (Ju et al., 2021)	65.0	63.2	64.1	66.5	65.5	66.0
	VLP-MABSA* (Ling et al., 2022)	65.1	68.3	66.6	66.9	69.2	68.0
	CMMT* (Yang et al., 2022b)	64.6	68.7	66.5	67.6	69.4	68.5
	MOCOLNet (Mu et al., 2023)	66.3	67.8	67.1	67.2	68.7	67.9
	VLP-MABSA-M2DF (Zhao et al., 2023a)	66.8	68.0	67.3	67.8	68.4	68.1
	Atlantis (Xiao et al., 2024)	65.6	69.2	67.3	68.6	70.3	69.4
	AoM* (Zhou et al., 2023)	67.9	69.3	68.6	68.4	71.0	69.7
	<b>Vanessa (Ours)</b>	<b>68.6</b>	<b>71.1</b>	<b>69.8*</b>	<b>69.2</b>	<b>72.1</b>	<b>70.6*</b>

Table 2: MABSA evaluation results. \* denotes the results from (Zhou et al., 2023). \* denotes the improvement is statistically significant on a two-tailed t-test ( $p < 0.001$ ). We color each row as the **best** and **second best**.

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
<b>Vanessa</b>	<b>68.6</b>	<b>71.1</b>	<b>69.8</b>	<b>69.2</b>	<b>72.1</b>	<b>70.6</b>
w/o AMG	66.8	68.9	67.5	67.4	69.3	67.7
w/o Impr	67.7	70.0	68.8	68.3	70.6	69.5
w/o Aes-cap	67.1	69.4	68.1	68.0	70.2	69.2
w/o SSL-ITR	67.5	69.3	68.3	67.9	69.7	68.8
w/o DTM	67.8	70.2	68.5	68.4	70.7	69.4

Table 3: Ablation study results for the Vanessa. We color each row as the **best** and **second best**.

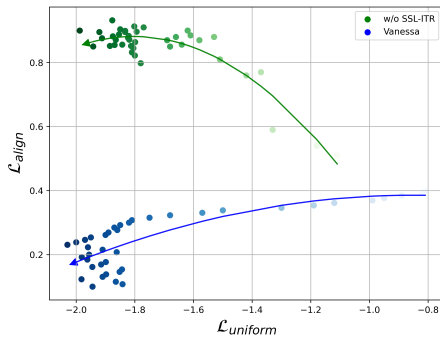


Figure 4: Visualization of contrastive representations for checkpoints at 40 training step intervals.

In Fig. 4, Vanessa consistently exhibits lower  $\mathcal{L}_{align}$  and  $\mathcal{L}_{uniform}$  values compared to the “w/o SSL-ITR” variant during training, which indicates that using the CLIP score improves Vanessa’s ability to learn sentimental clues for MABSA.

#### 4.5 Case Study

Fig. 5 presents two examples, accompanied by predictions from CMMT, AoM, and Vanessa. In example (a), CMMT incorrectly predicts the sentiment polarity for both “Cape Town” and “Regardt

Stander”, whereas AoM only misclassifies the sentiment for “Regardt Stander”. Vanessa accurately predicts the sentiments of both entities by effectively utilizing emotion-laden descriptions derived from impression and aesthetic attributes. This indicates Vanessa’s excellent capability in capturing and integrating implicit and explicit sentimental cues. In example (b), both CMMT and AoM incorrectly predict the sentiment of “LeBron James”. Due to the low semantic relevance between the image and text (CLIP score = 0.06), Vanessa primarily focuses on the text and accurately predicts the sentiment for both “LeBron James” and “NBA”. These observations highlight Vanessa’s adaptability in handling scenarios with varying levels of semantic relevance across modalities, ensuring robust sentiment predictions when the visual context provides minimal relevant information.

#### 4.6 Quantitative Analysis

We perform quantitative analysis to investigate the relationship between the impressions, aesthetic captions and our Vanessa across the test sets of two datasets. We input the hidden features of impressions and aesthetic captions produced by RoBERTa into a pre-trained TweetNLP (Loureiro et al., 2022) to obtain their sentiment distributions. Subsequently, we visualize the sentiment distributions of these auxiliary sentences alongside Vanessa’s predictions in the embedding space using the T-SNE (Van der Maaten and Hinton, 2008), as illustrated in Fig. 6. In Fig. 6 (a), impressions exhibit a bias toward positive samples, potentially introducing ambiguity in the training process. In contrast, Fig. 6 (b) shows that aesthetic captions





Image	(a) 	(b) 
Text	(a) The city of [Cape Town] <sub>Pos</sub> at night...photo by [Regardt Stander] <sub>Pos</sub> .	(b) [LeBron James] <sub>Neu</sub> to Produce [NBA] <sub>Neu</sub> Documentary -
	CLIP score: 0.61	CLIP score: 0.06
Impr.	I feel a sense of awe and wonder when I look at this image. The city lights and the dark background create a sense of tranquility and tranquility.	I feel a sense of curiosity and intrigue about the significance of the company's logo and its significance in the field of advertising.
Aes-Cap	The lighting of the city is very beautiful and the lighting of the city is very beautiful.	The visual elements that contribute to my impression include the people standing in front of the sign, which represents the company's commitment to promoting the brand.
CMMT	(Cape Town, Negative 😞) ✗ (Regardt Stander, Neutral 😐) ✗	(LeBron James, Positive 😄) ✗ (NBA, Neutral 😐) ✓
AoM	(Cape Town, Positive 😄) ✓ (Regardt Stander, Neutral 😐) ✗	(LeBron James, Positive 😄) ✗ (NBA, Neutral 😐) ✓
Vanessa	(Cape Town, Positive 😄) ✓ (Regardt Stander, Positive 😄) ✓	(LeBron James, Neutral 😐) ✓ (NBA, Neutral 😐) ✓

Figure 5: Two examples with predictions made by CMMT, AoM and Vanessa. The ground truth aspect-sentiment pair is annotated within the text.

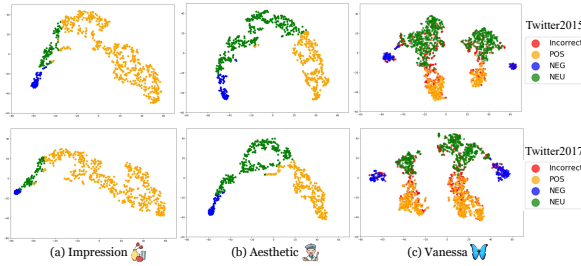


Figure 6: Visualization of sentiment distributions for auxiliary sentences and Vanessa's predictions.

present more distinct and separated sentiment clusters, likely providing clearer signals for model learning. From Fig. 6 (c), the majority of the Vanessa-predicted NEU samples coincide with the NEU samples in the aesthetic captions distribution. A subset of Vanessa-predicted NEG samples overlaps with NEG samples in both aesthetic captions and impressions. Despite noticeable differences in the distribution of POS samples between model predictions and both aesthetic captions and impressions, a degree of similarity is observed in the right half of the plots. In summary, given the intricate sentimental cues and alignment challenges in MABSA (Mao et al., 2025), we hypothesize that aesthetic captions offer more definitive sentimental cues compared to impressions on these test sets. Our ablation study supports this hypothesis, as the "w/o Aes-cap" variant results in greater performance degradation than the "w/o Impr" variant.

## 5 Conclusion

We proposed a novel **Visual Connotation** and **Aesthetic Attributes Understanding Network**

(Vanessa) for MABSA. Firstly, the MA<sup>3</sup> module adaptively modeled the intra- and inter-dynamics of aesthetic-aware emotions across modalities. Subsequently, the SSL-ITR module dynamically prioritized visual or textual modalities to improve the model's ability to discern and utilize relevant multimodal features. Finally, the DTM module adeptly selected and merged aesthetic-aware and emotionally rich features at both implicit and explicit levels. Experimental results on two widely used Twitter datasets verified the effectiveness of our Vanessa.

## Limitations

The proposed Vanessa has the following limitations. Firstly, the aesthetic-aware multimodal dependency matrix is a homogeneous graph, which limits its ability to represent diverse features. This constraint hinders the model's capacity to deeply explore the intra- and inter-dynamics between bi-modality, visual connotation, and aesthetic attributes. Future work will focus on constructing a heterogeneous graph to better model the diverse data, enhancing the model's ability to analyze complex multimodal relationships. Secondly, the generated impressions and aesthetic captions are not well-aligned with specific targets within the sentences, as the generated content predominantly pertains to the image and lacks sufficient relation to the specific targets. Thirdly, the reliability of results is paramount for applications ranging from market research to social media monitoring. Enhancing the robustness of models against abnormal or malicious inputs is essential to maintain this reliability (Zhao et al., 2023b, 2024).

## Ethics Statement

This article adheres to the ACL Code of Ethics. The datasets utilized do not contain sensitive private information and pose no harm to society. The proposed method is for multimodal sentiment analysis and enhancing machine understanding of human sentiment. To the best of our knowledge, there are no foreseeable risks associated with this technique.

## Acknowledgments

This research/project is funded by the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105901) and the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005).

## References

- Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. 2023. Metaclue: Towards comprehensive visual metaphors research. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Rudolf Arnheim. 1954. *Art and visual perception: A psychology of the creative eye*. Univ of California Press.
- Lisa Feldman Barrett and Moshe Bar. 2009. See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1325–1334.
- John Berger. 1972. *Ways of seeing*. BBC and Penguin.
- Erik Cambria. 2024. *Understanding Natural Language Understanding*. Springer, ISBN 978-3-031-73973-6.
- Erik Cambria, Newton Howard, Jane Hsu, and Amir Hussain. 2013. Sentic blending: Scalable multi-modal fusion for continuous interpretation of semantics and sentics. In *IEEE SSCI*, pages 108–117.
- Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. 2023. Seven pillars for the future of artificial intelligence. *IEEE Intelligent Systems*, 38(6):62–69.
- Erik Cambria, Xulang Zhang, Rui Mao, Melvin Chen, and Kenneth Kwok. 2024. SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing. In *Proceedings of International Conference on Human-Computer Interaction (HCI)*, Washington DC, USA.
- Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388. Association for Computational Linguistics.
- Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th international conference on computational linguistics*, pages 272–279.
- Kelvin Du, Rui Mao, Frank Xing, and Erik Cambria. 2024. Explainable stock price movement prediction using contrastive learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, Idaho, USA.
- Mingjing Du, Shifei Ding, and Hongjie Jia. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145.
- Chunxiao Fan, Jie Lin, Rui Mao, and Erik Cambria. 2024. Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion*, 106:102306.
- Xianjie Guo, Kui Yu, Lin Liu, and Jiuyong Li. 2024. Fedcsl: A scalable and accurate approach to federated causal structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12235–12243.
- Xianjie Guo, Kui Yu, Lin Liu, Peipei Li, and Jiuyong Li. 2023. Adaptive skeleton construction for accurate dag learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10526–10539.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.
- Linmei Hu, Ziwei Chen, Ziwang Zhao, Jianhua Yin, and Liqiang Nie. 2022. Causal inference for leveraging image-text matching bias in multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11141–11152.
- James E Humphreys. 2012. *Introduction to Lie algebras and representation theory*, volume 9. Springer Science & Business Media.
- Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. 2023. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2482.
- Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 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*, pages 4395–4405.
- Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. 2023. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10041–10051.
- Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3034–3042.
- Julia Kruk, Caleb Ziems, and Diyi Yang. 2023. Impressions: Visual semiotics and aesthetic impact understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12273–12291.

- Peter Lang and Margaret M Bradley. 2007. The international affective picture system (iaps) in the study of emotion and attention. *Handbook of emotion elicitation and assessment*, 29:70–73.
- Steven J Leon, Åke Björck, and Walter Gander. 2013. Gram-schmidt orthogonalization: 100 years and more. *Numerical Linear Algebra with Applications*, 20(3):492–532.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.
- Bin Liang, Lin Gui, Yulan He, Erik Cambria, and Ruifeng Xu. 2024. Fusion and discrimination: A multimodal graph contrastive learning framework for multimodal sarcasm detection. *IEEE Transactions on Affective Computing*, 15.
- Yan Ling, Jianfei Yu, and Rui Xia. 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)*, pages 2149–2159.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.
- Hui Liu, Wenya Wang, and Haoliang Li. 2022. 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*, pages 4995–5006.
- Hui Liu, Wenya Wang, and Haoliang Li. 2023a. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796.
- Hui Liu, Wenya Wang, Hao Sun, Anderson Rocha, and Haoliang Li. 2023b. Robust domain misinformation detection via multi-modal feature alignment. *IEEE Transactions on Information Forensics and Security*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.
- Qiang Lu, Xia Sun, Yunfei Long, Zhizezhang Gao, Jun Feng, and Tao Sun. 2023. Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. *IEEE Transactions on Neural Networks and Learning Systems*.
- Rui Mao, Kelvin Du, Yu Ma, Luyao Zhu, and Erik Cambria. 2023. Discovering the cognition behind language: Financial metaphor analysis with MetaPro. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 1211–1216. IEEE.
- Rui Mao, Mengshi Ge, Sooji Han, Wei Li, Kai He, Luyao Zhu, and Erik Cambria. 2025. A survey on pragmatic processing techniques. *Information Fusion*, 114:102712.
- Rui Mao, Kai He, Claudia Beth Ong, Qian Liu, and Erik Cambria. 2024. MetaPro 2.0: Computational metaphor processing on the effectiveness of anomalous language modeling. In *Findings of the Association for Computational Linguistics: ACL*, pages 9891–9908, Bangkok, Thailand. Association for Computational Linguistics.
- Rui Mao and Xiao Li. 2021. 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*, volume 35, pages 13534–13542.
- Jie Mu, Feiping Nie, Wei Wang, Jian Xu, Jing Zhang, and Han Liu. 2023. Mocolnet: A momentum contrastive learning network for multimodal aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Vilayanur S Ramachandran and William Hirstein. 1999. The science of art: A neurological theory of aesthetic experience. *Journal of consciousness Studies*, 6(6-7):15–51.
- Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868.
- Yosephine Susanto, Andrew Livingstone, Bee Chin Ng, and Erik Cambria. 2020. The Hourglass Model revisited. *IEEE Intelligent Systems*, 35(5):96–102.

- Ana Valdivia, Victoria Luzón, Erik Cambria, and Francisco Herrera. 2018. Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion*, 44:126–135.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Qianlong Wang, Hongling Xu, Zhiyuan Wen, Bin Liang, Min Yang, Bing Qin, and Ruifeng Xu. 2023. Image-to-text conversion and aspect-oriented filtration for multimodal aspect-based sentiment analysis. *IEEE Transactions on Affective Computing*.
- Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.
- Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International conference on multimedia*, pages 1038–1046.
- Luwei Xiao, Xingjiao Wu, Junjie Xu, Weijie Li, Cheng Jin, and Liang He. 2024. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Information Fusion*, page 102304.
- Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Jie Zhou, and Liang He. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6):103508.
- Luwei Xiao, Yun Xue, Hua Wang, Xiaohui Hu, Donghong Gu, and Yongsheng Zhu. 2022. Exploring fine-grained syntactic information for aspect-based sentiment classification with dual graph neural networks. *Neurocomputing*, 471:48–59.
- Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 371–378.
- Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 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)*, pages 2416–2429.
- Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. 2023. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*.
- Hao Yang, Yanyan Zhao, and Bing Qin. 2022a. 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*, pages 3324–3335.
- Li Yang, Jin-Cheon Na, and Jianfei Yu. 2022b. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing & Management*, 59(5):103038.
- Li Yang, Jieming Wang, Jin-Cheon Na, and Jianfei Yu. 2023. Generating paraphrase sentences for multimodal entity-category-sentiment triple extraction. *Knowledge-Based Systems*, 278:110823.
- Li Yang, Zengzhi Wang, Ziyang Li, Jin-Cheon Na, and Jianfei Yu. 2024. An empirical study of multimodal entity-based sentiment analysis with chatgpt: Improving in-context learning via entity-aware contrastive learning. *Information Processing & Management*, 61(4):103724.
- Jianfei Yu, Kai Chen, and Rui Xia. 2022a. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.
- Jianfei Yu and Jing Jiang. Adapting bert for target-oriented multimodal sentiment classification.(2019). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5408–5414.
- Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.
- Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 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*, pages 3342–3352.
- Jianfei Yu, Jieming Wang, Rui Xia, and Junjie Li. 2022b. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *IJCAI*, pages 4482–4488.
- Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. 2023. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion*, 100:101921.
- Hui Zeng, Zisheng Cao, Lei Zhang, and Alan C Bovik. 2019. A unified probabilistic formulation of image aesthetic assessment. *IEEE Transactions on Image Processing*, 29:1548–1561.

- Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14347–14355.
- Honglei Zhang, He Liu, Haoxuan Li, and Yidong Li. 2024a. Transfr: Transferable federated recommendation with pre-trained language models. *arXiv preprint arXiv:2402.01124*.
- Honglei Zhang, Xin Zhou, Zhiqi Shen, and Yidong Li. 2024b. Privfr: Privacy-enhanced federated recommendation with shared hash embedding. *IEEE Transactions on Neural Networks and Learning Systems*.
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.
- Xulang Zhang, Rui Mao, Kai He, and Erik Cambria. 2023. Neuro-symbolic sentiment analysis with dynamic word sense disambiguation. In *EMNLP*, pages 8772–8783.
- Fei Zhao, Chunhui Li, Zhen Wu, Yawen Ouyang, Jianbing Zhang, and Xinyu Dai. 2023a. 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*, pages 9057–9070.
- Shuai Zhao, Luu Anh Tuan, Jie Fu, Jinming Wen, and Weiqi Luo. 2024. Exploring clean label backdoor attacks and defense in language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Shuai Zhao, Jinming Wen, Anh Luu, Junbo Zhao, and Jie Fu. 2023b. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12303–12317.
- Jie Zhou, Jiabao Zhao, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. 2021. Masad: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, 455:47–58.
- Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. 2023. Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8184–8196.

## A Appendix

### A.1 Impact of different relation detection

To further assess the effectiveness of our substitution for noise reduction techniques, we conducted ablation studies by removing the SSL-ITR module from our framework. In this setup, we directly input the image-text pairs into the Relation Detection module as described in (Ju et al., 2021). Consistent with the methodology in (Ju et al., 2021), we calculated both soft and hard relation scores between the modalities. These relation scores were used to weight the image-related features—including those extracted by CLIP, impressions, and aesthetic captions. The weighted features were then fed into the MA<sup>3</sup> module to obtain the final MABSA results. The experimental outcomes are summarized in Table 4.

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
Vanessa(hard)	65.7	67.6	66.5	67.1	69.6	68.1
Vanessa(soft)	66.5	69.0	67.6	68.4	70.8	69.3
Vanessa(SSL-ITR)	68.6	71.1	69.8	69.2	72.1	70.6

Table 4: The results of different relation detection methods of Vanessa. Vanessa (hard) refers to the hard relation score, while "Vanessa (soft)" denotes the soft relation score. Vanessa(SSL-ITR) is our proposed method. We color each row as the **best**.

As illustrated in Table 4, the inclusion of the SSL-ITR module significantly enhances the model’s performance across both datasets. Specifically, our Vanessa with SSL-ITR achieves the highest F1 scores of 69.8% on Twitter2015 and 70.6% on Twitter2017, outperforming the versions without SSL-ITR by a considerable margin. The models utilizing hard and soft relation scores without SSL-ITR exhibit lower F1 scores, indicating that the absence of the SSL-ITR module impairs the model’s ability to effectively reduce noise and capture the nuanced interactions between modalities. These results demonstrate that the SSL-ITR module plays a crucial role in enhancing semantic alignment between images and text by effectively filtering out irrelevant or noisy information. By leveraging self-supervised contrastive learning under the guidance of CLIP scores, the SSL-ITR module improves the quality of the image-text representations, boosting the overall performance of the MABSA task.

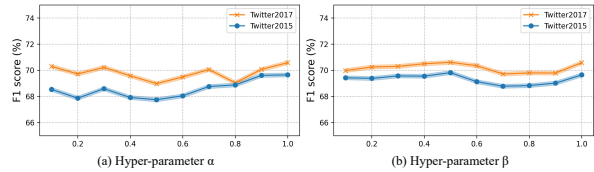


Figure 7: Effect of the hyper-parameters on two datasets.

### A.2 Impact of hyper-parameters

We conducted extra experiments to evaluate the impact of hyper-parameters on Vanessa’s performance. The  $\alpha$  modulates the influence of contrastive learning on image-text relevance and  $\beta$  controls the strength of the transformation matrix that aligns image and text feature spaces. From Fig. 7(a), we observed that as  $\alpha$  increases, the model’s performance improves, reaching its optimal value at  $\alpha = 1$ . This observation suggests that the contrastive loss is most effective when balanced appropriately, allowing the model to learn practical representations of image-text relevance. In Fig. 7(b), as the value of  $\beta$  varies, the performance of Vanessa exhibits relatively minor fluctuations, reaching its peak at  $\beta = 0.5$ . This optimal setting indicates that the transformation matrix effectively aligns the image and text representations when the contribution of  $\beta$  is neither too weak nor too strong. In conclusion, optimal tuning of both  $\alpha$  and  $\beta$  is essential for balancing semantic alignment with preserving the unique characteristics of each modality. Moderate values maximize performance by enhancing multimodal integration without overfitting or distorting feature spaces, as reflected in the results from both datasets.

### A.3 Comparison with large multimodal models

We further conducted a comparative evaluation against the open-source Multimodal Large Language Model (MLLM) LLaVA-1.5-7b (Liu et al., 2024) in a zero-shot setting on the test set. As illustrated in Table 5, the proposed Vanessa substantially outperforms LLaVA-1.5-7b across all evaluation metrics on both datasets. Vanessa achieves F1 scores that are more than twice those of LLaVA-1.5-7b (32.5% and 34.4%, respectively). This significant improvement is also reflected in the precision and recall, where Vanessa consistently demonstrates superior performance. These results underscore the efficacy of our task-specific approach in the domain of MABSA. While LLaVA-1.5-7b, as a

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
LLaVA-1.5-7b	30.8	33.7	32.5	33.3	35.6	35.2
Vanessa	68.6	71.1	69.8	69.2	72.1	70.6

Table 5: Main results compared with LLaVA-1.5-7b. We color each row as the **best**.

large-scale MLLM, offers generalizability and has shown impressive capabilities in zero-shot settings, it falls short in capturing the fine-grained sentiment cues present in multimodal social media data. In contrast, Vanessa is explicitly designed to model the intricate relationships between images and text. By incorporating specialized components such as the Multi-Aesthetic Attributes Aggregation (MA<sup>3</sup>) module and the Self-Supervised Image-Text Relevance (SSL-ITR) module, Vanessa effectively captures both explicit and implicit sentimental cues.