Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081626

ARTICLE

Multi-Branch Cross-Modal Cross-Attention for Image–Text Multimodal Sentiment Classification

Xinshan Huang[1], **Zirui Pei**[1], **Chaohong Tan**[2] **and Zuqiang Meng**[1,*]

[1]College of Computer, Electronics and Information, Guangxi University, Nanning, China
[2]Guangxi Key Laboratory of Digital Infrastructure, Guangxi Zhuang Autonomous Region Information Center, Nanning, China
*Corresponding Author: Zuqiang Meng. Email: zqmeng@126.com

Received: 05 March 2026; Accepted: 13 May 2026; Published: 15 June 2026

ABSTRACT: Multimodal Sentiment Analysis (MSA) plays an important role in understanding social media content; however, existing methods often struggle with the heterogeneity and complex interactions between images and text. These challenges include inter-modal information asymmetry, insufficient feature fusion, and noise interference, which collectively limit robustness and accuracy. To address these issues, we propose a multimodal sentiment classification model termed Multi-Branch Cross-Modal Cross-Attention Gating (MB-CMCAG). The model first incorporates a Transformer-based image caption generation module to convert raw images into semantically rich auxiliary textual descriptions, which complement the original text and form paired textual inputs with enhanced visual semantics. To capture multi-source features, MB-CMCAG adopts a dual-branch feature extraction architecture: the visual branch encodes images using a Vision Transformer (ViT), while the textual branch encodes text with Bidirectional Encoder Representations from Transformers (BERT); a Contrastive Language-Image Pre-training (CLIP) model is also introduced for joint image–text feature extraction. To exploit cross-modal correlations and enable hierarchical fusion, we construct a cross-modal attention module that supports bidirectional information flow from image to text and from text to image. Building on this, a cross-modal gated mechanism is introduced to selectively regulate the transmission and aggregation of features from different sources, thereby improving noise suppression and sentiment sensitivity. Experimental results on the public MVSA-Single and MVSA-Multiple datasets show that MB-CMCAG achieves accuracies of 76.38% and 73.87%, respectively, outperforming existing baselines by a clear margin in image–text multimodal sentiment classification.

KEYWORDS: Multimodal sentiment analysis; cross-modal cross-attention; cross-modal gated fusion

1 Introduction

Sentiment analysis, as an important research direction in natural language processing, aims to identify, interpret, and process the emotions and affective states expressed by humans. With the rapid development of Internet technologies and social platforms, an increasing number of people, both domestically and internationally, use social media such as Weibo, Twitter, and Facebook to express their thoughts and feelings. At the same time, the carriers of expressed content have evolved from the traditional text modality to multiple modalities, including images, video, and audio. Emotion recognition is a crucial area of artificial intelligence, and multimodal emotion recognition has become a research hotspot in recent years [1]. Multimodal Sentiment Analysis (MSA) has broad applications across many domains, including human–computer interaction [2] and advertising and commerce [3].

Existing MSA methods can be broadly categorized by their feature-fusion strategies: early fusion, intermediate fusion, and late fusion [4]. Early fusion combines modalities at the feature-extraction stage but struggles with fine-grained cross-modal information and suffers from feature redundancy [5]. Late fusion trains separate classifiers for each modality but fails to model inter-modal correlations effectively [6]. Intermediate fusion, which facilitates feature interaction within neural network layers, has gained attention for its ability to capture cross-modal relationships [7,8].

With the development of deep learning, MSA has shifted from traditional fusion methods to more sophisticated fusion mechanisms. A growing number of researchers have begun to leverage vision–language models such as CLIP [9] for MSA. For example, reference [10] proposed a CLIP-based sentiment analyzer that employs a high-level feature extraction module and integrates features using attention mechanisms and custom layers, thereby addressing the limitations of traditional methods in effectively combining multimodal data. In addition, reference [8] introduced a multimodal sentiment classification model based on a Gated Attention mechanism, which uses a pretrained convolutional neural network to extract fine-grained visual features and fuses them with textual features through gated attention, in order to reduce noise interference and highlight the key textual components that influence sentiment polarity.

Although substantial research efforts have been devoted to advancing MSA, existing methods still suffer from three major limitations: (1) reliance on a single feature extraction pathway, which hinders the full extraction of discriminative representations from images and text across multiple perspectives and scales; (2) relatively shallow cross-modal interaction mechanisms, which typically rely on simple concatenation or summation and fail to precisely capture the complementary relationships and accurate alignment between visual and textual modalities in fine-grained semantic space; and (3) static and rigid modality fusion strategies that lack dynamic screening and adaptive re-weighting of cross-modal information. In particular, many CLIP-based or interaction-based methods emphasize either global alignment or shallow cross-modal fusion, but they do not simultaneously model fine-grained local semantics, coarse-grained global consistency, and adaptive modality translation in a unified framework. Consequently, these methods are susceptible to redundant noise interference, thereby limiting sentiment discrimination performance in complex real-world scenarios.

To address these challenges, this paper proposes a multimodal sentiment analysis model based on a multi-branch cross-modal cross-attention gated network (MB-CMCAG). To overcome the limitations of single-path feature extraction, we introduce a dual-branch complementary encoding mechanism. The fine-grained branch captures local semantics by employing the Vision Transformer (ViT) [11] model to encode images and the Bidirectional Encoder Representations from Transformers (BERT) [12] model to encode text. The coarse-grained branch utilizes CLIP to capture global alignment. These two branches form mutually complementary representations at different semantic levels, providing a richer foundation for subsequent fusion.

Furthermore, to bridge the semantic gap between visual and textual modalities and enable effective dynamic fusion of semantic information, we adopt a cross-modal cross-attention module and a cross-modal gated mechanism. In addition, to enhance the model's understanding of visual emotional semantics, we employ a transformer-based caption generation module that converts images into captions. This transforms visual semantic information into textual form, thereby constructing an auxiliary sentence that enriches the original text. In summary, the main contributions of our work are as follows:

- **Multi-Perspective Hierarchical Feature Extraction Mechanism**—We propose a dual-branch encoding architecture that integrates fine-grained modeling (ViT+BERT) with coarse-grained modeling (CLIP). Unlike previous single-path or dual-stream methods that operate at a single semantic scale,

our framework is the first to explicitly construct complementary multi-scale representations in parallel, simultaneously capturing local details and global semantics within each modality.

- **Hierarchical Cross-Modal Interaction and Fusion Framework**—Bidirectional cross-attention is adopted within each branch to explicitly model fine-grained semantic alignment, while a gated mechanism is introduced between branches to enable dynamic re-weighting of cross-scale information. This design goes beyond shallow and unidirectional interactions in existing methods by enabling bidirectional, multi-level semantic alignment in a true cross-modal space.
- **Modality Translation Enhancement Strategy**—We introduce an image captioning module to explicitly convert visual semantics into textual representations and incorporate them into the fusion process. By reinforcing visual emotional signals in the textual dimension, this approach fundamentally bridges the modality gap and improves the comparability and fusibility of visual information in semantic space.
- **Comprehensive Evaluation and Component Analysis**—Extensive experiments, ablation studies, and efficiency analysis demonstrate the effectiveness and practicality of the proposed framework.

2 Related Work

2.1 Text Sentiment Analysis

Text sentiment analysis is one of the earliest and most mature research directions in affective computing and natural language processing. Early studies mainly relied on sentiment lexicons and traditional machine learning methods, whereas the recent rise of deep learning models has significantly advanced the performance of this field.

First, lexicon-based methods [13] primarily rely on precompiled sentiment lexicons to determine overall sentiment by identifying sentiment-bearing words and their intensities in text. For example, Alwan et al. [14] proposed a hybrid model based on rough set theory and sentiment lexicons for polarity analysis of Arabic political articles. To address low classification accuracy on Arabic political texts, they constructed a dedicated corpus of 206 annotated documents and performed classification via three core steps—preprocessing, feature extraction, and hybrid-model construction—achieving an accuracy of 85.483%. In a comment-classification task, Hamouda et al. employed the SentiWordNet lexicon [15], which assigns positive, neutral, and negative sentiment scores to lexical items [16]. Although lexicon-based methods offer interpretability, require no training data, and are grounded in manual rules, they struggle to capture contextual shifts, sarcasm or metaphor, and domain-specific expressions. Traditional machine-learning approaches commonly use algorithms such as k-nearest neighbors (k-NN), support vector machines (SVM) [17], and naive Bayes (NB) [18]. Goel et al. [19] presented a real-time Twitter sentiment-analysis system that combines naive Bayes with SentiWordNet, improving classification accuracy by incorporating SentiWordNet's scoring scheme. Rathor et al. [20] compared SVM, NB, and maximum entropy (ME) for sentiment analysis of Amazon product reviews and found SVM achieved the highest accuracy. More recently, the rise of deep learning has substantially advanced text sentiment analysis. Kim [21] applied convolutional neural networks (CNNs) to sentence-level sentiment classification with strong results. Zhou et al. [22] proposed a bilingual LSTM model with an attention mechanism for cross-lingual sentiment classification. Wang et al. [23] introduced a tree-structured region-dividing CNN–LSTM hybrid model for dimensional sentiment analysis; by partitioning text into multiple semantic regions, extracting local features via CNNs, and capturing long-range cross-region dependencies with LSTMs, the model produces continuous-valued predictions of sentiment intensity.

2.2 Image Sentiment Analysis

Image sentiment analysis aims to identify and interpret the emotions conveyed in images by analyzing their visual content. Similar to text sentiment analysis, image sentiment analysis has also evolved from traditional approaches to deep learning–based methods.

Early approaches typically relied on hand-crafted visual features, such as low-level cues including color, texture, shape, and illumination, combined with machine learning classifiers for sentiment prediction. For example, Siersdorfer et al. [24] investigated the relationship between the affective content of images in social media and their metadata and visual content. By analyzing more than 586,000 images from Flickr, they explored the feasibility of predicting image sentiment polarity by combining visual features (e.g., color distributions and SIFT descriptors) with textual sentiment labels extracted using the SentiWordNet lexicon. In [25], a large-scale Visual Sentiment Ontology (VSO) containing more than 3000 adjective–noun pairs (ANPs) was constructed, and a SentiBank library with 1200 ANP detectors was developed on this basis, which significantly improved the accuracy of sentiment prediction for visual content by detecting affective concepts in images. Yang et al. [26] proposed a graph-based sentiment analysis method that jointly models user-posted images and comments from friends to infer users' affective states. However, these methods often struggle to capture complex high-level semantic and affective information in images. With the advent of deep learning, image sentiment analysis has achieved substantial progress. Song et al. [27] proposed a visual-attention-based image sentiment analysis method (SentiNet-A), which integrates convolutional neural networks (CNNs) with a visual attention mechanism to improve image sentiment classification performance. Islam and Zhang [28] adopted transfer learning by initializing the network with parameters from a pretrained GoogLeNet [29] and, combined with data augmentation, achieved an accuracy of 86.1% on a Twitter image dataset.

2.3 Multimodal Sentiment Analysis

Although previous studies have demonstrated the effectiveness of unimodal approaches for emotion recognition, comparative analyses in the literature consistently show that multimodal strategies achieve superior performance [30].

MSA seeks to infer sentiment by integrating heterogeneous modalities—such as vision, text, and audio—to provide a more comprehensive understanding of human affect. The recent surge in multimodal content has driven renewed interest in MSA. You et al. [31] proposed a Cross-modal Consistency Regression (CCR) model that uses CNNs to extract visual features and a paragraph-vector model to obtain textual features; consistency across modalities is then enforced via Kullback–Leibler (KL) divergence to enable joint optimization. Zhao et al. [32] introduced an image–text consistency–driven approach that first uses an SVM classifier to assess whether image and text semantics align, and then combines intermediate visual features from SentiBank [25] with textual and social features for adaptive MSA. To strengthen inter-modal interaction, several works employ attention mechanisms. Zhang et al. [33] proposed an attention-based model with a symmetric architecture that processes text and image data in parallel: a denoising autoencoder extracts robust textual features, while an improved variational autoencoder with attention extracts salient visual features. They further design a cross-feature fusion module that uses attention to learn complementary information between text and image features rather than relying on simple concatenation or one-way influence. Zhou et al. [34] introduced CAHFW-Net, a method for video-based emotion recognition that combines cross-attention with a hybrid feature-weighted neural network. Its hierarchical attention encoding network—particularly the cross-attention module—effectively captures complementary cues between facial expressions and scene context, mitigating emotion confusion and misinterpretation. Khan and Fu [35] developed a dual-stream model based on input-space translation (EF-CapTrBERT): an object-aware Transformer

first translates images into natural-language descriptions, which are then formed into auxiliary sentences to inject multimodal information into BERT. This approach addresses image noise and modality fusion without modifying BERT's internal architecture. With the rapid progress of vision–language models (VLMs) [36], their strong multimodal capabilities have seen wide adoption. For example, Huang et al. [37] proposed CLIP-MSA, a CLIP-based MSA model that introduces CLIP's cross-modal encoder as the second branch of a dual-branch feature extractor. By leveraging CLIP's pretrained semantic alignment and external knowledge, CLIP-MSA substantially improves the quality of unimodal representations.

In summary, although existing methods have achieved notable progress on MSA, they still exhibit clear limitations in hierarchical modeling of visual–textual features and in the fine-grained characterization of dynamic inter-modal interactions. Most current approaches follow a single-path feature extraction paradigm and perform multimodal fusion via simple concatenation or summation. Such coarse-grained modeling is insufficient to fully explore discriminative intra-modal representations from multiple perspectives, and it fails to effectively capture the complementarity and alignment between images and text in a fine-grained semantic space, which often leads to suboptimal sentiment recognition in complex real-world scenarios. Recent CLIP-based methods (e.g., FAMEAC [10], SentiCLIP [38]) and large vision-language models have made important advances; however, they have not yet simultaneously achieved multi-scale hierarchical representation, bidirectional fine-grained cross-modal alignment, and adaptive dynamic fusion in a unified framework. To bridge these remaining gaps, the proposed MB-CMCAG network performs multi-path collaborative modeling to learn hierarchical visual–textual representations at both coarse and fine granularities, employs bidirectional cross-modal cross-attention for deep semantic interaction, and introduces a novel cross-modal interactive gating mechanism that adaptively selects and reweights cross-modal information, thereby effectively suppressing redundant noise and highlighting key cues critical for sentiment discrimination.

3 Methodology

3.1 Task Definition

Given a multimodal sentiment analysis dataset $\mathcal{D} = \{(I_i, S_i, y_i)\}_{i=1}^{N}$, each sample consists of an image I_i, an original text S_i, and a sentiment label $y_i \in \{\text{positive}, \text{neutral}, \text{negative}\}$. To better exploit the semantic and emotional cues embedded in the visual modality, an image caption generation model $g(\cdot)$ is employed to generate a descriptive caption $C_i = g(I_i)$ for each image. The generated caption is concatenated with the original text to form the final textual input $T_i = \text{Concat}(S_i, C_i)$. The goal of this task is to learn a multimodal sentiment classification mapping function $f : (I, T) \mapsto y$ based on the image I_i and the augmented text T_i. By incorporating image-generated captions, the model can jointly learn complementary emotional representations from both visual and linguistic modalities, thereby achieving more accurate multimodal sentiment understanding.

3.2 Model Overview

The proposed MB-CMCAG model consists of five main components: an image caption generation module, a multi-branch feature-extraction layer, a cross-modal cross-attention interaction layer, a cross-modal cross-gating fusion layer, and a sentiment classification layer. As illustrated in Fig. 1, the image captioning module converts an image into a textual description, which is then concatenated with the original text to form the final composite text. The multi-branch feature-extraction module leverages different pretrained models to obtain multi-granularity feature representations from both images and text. The cross-modal cross-attention interaction module enhances information exchange between modalities through bidirectional attention, i.e., image-to-text and text-to-image. The cross-modal cross-gating fusion module adaptively

controls the fusion strength of features from different modalities and integrates contextual information. Finally, the fused multimodal representation is fed into a multi-layer classifier for sentiment prediction.

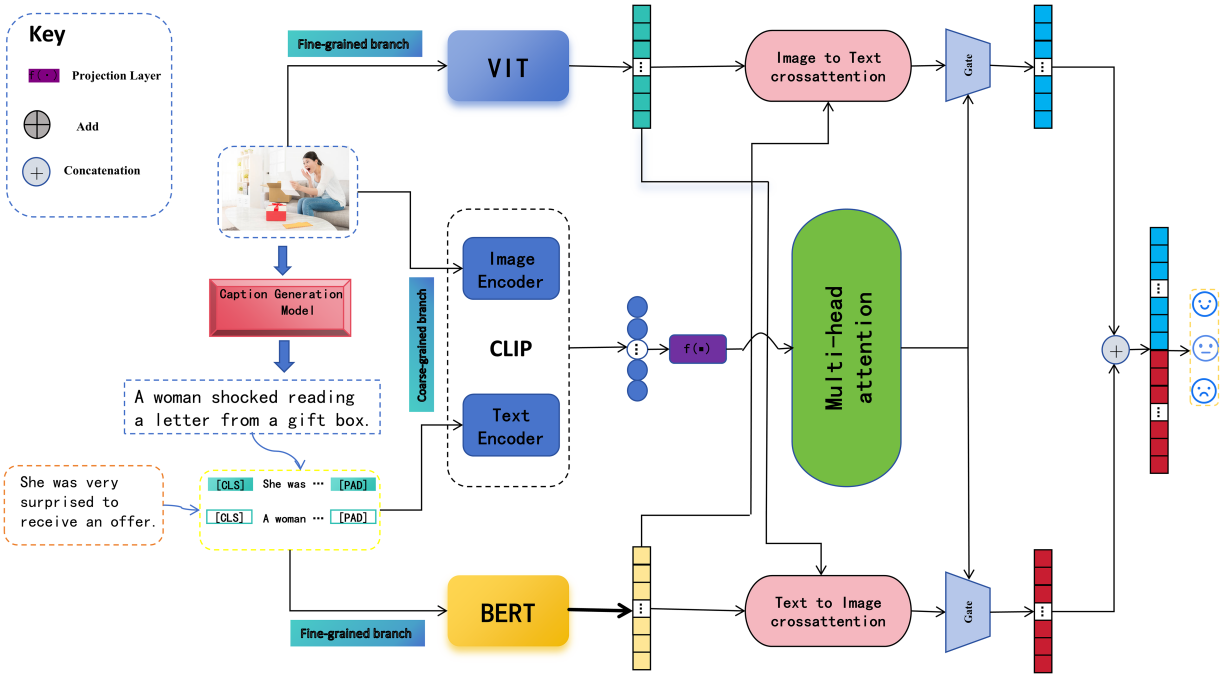

Figure 1: Overall architecture of the MB-CMCAG model.

3.2.1 Multi-Branch Feature Extraction

In multimodal sentiment analysis, our inputs consist of two primary modalities: an image I_i and a text T_i, where T_i is an augmented text formed by concatenating the original text S_i with the image-generated caption $C_i = g(I_i)$. To comprehensively capture the characteristics of both modalities, we design a multi-branch feature extraction module.

Unlike single-branch encoding strategies that rely on a unified representation space, the proposed dual-branch design aims to explicitly model complementary semantic information at different granularity levels, thereby mitigating the limitations of either overly coarse or overly local feature representations.

Specifically, the fine-grained branch employs ViT to extract patch-level local features from the image and BERT to perform token-level semantic analysis on the enhanced text, thereby preserving rich spatial structures and sequential details for fine-grained semantic alignment between modalities. The coarse-grained branch leverages a pre-trained CLIP model to extract global semantic embeddings for both image and text, providing high-level cross-modal consistency alignment. However, CLIP's highly compressed global vectors inevitably discard the original spatial layout and sequential structure. Therefore, the proposed multi-branch architecture is not a simple stacking of pre-trained models; instead, it deliberately constructs two complementary representation perspectives—the fine-grained branch captures local discriminative information while the coarse-grained branch supplies robust global semantic context—forming a richer, more hierarchical feature space that provides a sufficient and complementary foundation for the subsequent cross-modal cross-attention module and dynamic gating mechanism.

Visual Feature Extraction Using ViT

For the input image I_i, we employ a ViT to extract visual features. The ViT divides the image into fixed-size patches, linearly embeds each patch, and adds positional encodings to form a sequence that is fed into the Transformer encoder. Concretely, the image I_i is split into P equal-sized patches; each patch is mapped via a linear projection into a d_v-dimensional feature space, producing the image feature sequence V_i:

$$V_i = \mathrm{ViT}(I_i) = [vi, 1, vi, 2, \ldots, v_{i,P}] \in \mathbb{R}^{P \times d_v} \tag{1}$$

where $v_{i,j} \in \mathbb{R}^{d_v}$ denotes the feature vector of the j-th image patch, P is the total number of patches, and d_v is the feature dimensionality of the ViT output.

Text Feature Extraction Using BERT

For the textual input $T_i = \mathrm{Concat}(S_i, C_i)$, the text is first tokenized and then encoded using the BERT model. Owing to its bidirectional Transformer architecture, BERT effectively captures contextual information in the text, thereby producing richer textual representations. The text sequence T_i is encoded by BERT as follows:

$$H_i = \mathrm{BERT}(T_i) = [hi, 1, hi, 2, \ldots, h_{i,L}] \in \mathbb{R}^{L \times d_h} \tag{2}$$

where $h_{i,j} \in \mathbb{R}^{d_h}$ denotes the contextual semantic feature of the j-th token, L is the length of the text sequence, and d_h is the feature dimensionality of the BERT output.

Multimodal Feature Extraction Using CLIP

In order to enhance the model's understanding of cross-modal semantic relations, we incorporate the Contrastive Language–Image Pretraining (CLIP) model as an additional pathway to obtain joint image–text representations. CLIP uses contrastive learning to project images and text into a shared embedding space, where corresponding image–text pairs are pulled closer, thereby providing globally aligned semantic representations. Images and text are processed by CLIP's image encoder and text encoder, respectively. For a given image I_i and text T_i, we compute the outer product ($\otimes$) between their CLIP features to capture cross-modal interactions beyond simple concatenation, followed by a diagonalization operation to reduce redundancy and retain the most informative aligned components:

$$V_i^c = \mathrm{CLIP}_{\mathrm{image}}(I_i) \in \mathbb{R}^{d_c} \tag{3}$$

$$H_i^c = \mathrm{CLIP}_{\mathrm{text}}(T_i) \in \mathbb{R}^{d_c} \tag{4}$$

$$U_i = \mathrm{Diag}\left(V_i^c \otimes H_i^c\right) \in \mathbb{R}^{d_c \times d_c} \tag{5}$$

where d_c denotes the output dimensionality of the CLIP model, $\otimes$ denotes the outer product, and $\mathrm{Diag}(\cdot)$ is applied to extract the diagonal values.

3.2.2 Multimodal Feature Fusion Module

In the image–text feature fusion module, feature fusion remains one of the core bottlenecks that constrain further performance improvement. Because the visual modality (e.g., facial expressions and scene composition) and the textual modality (e.g., dialogue semantics and affective words) exhibit substantial heterogeneity in representation form, information density, and spatiotemporal characteristics, conventional fusion strategies—such as concatenation, weighted averaging, or simple attention mechanisms—are insufficient to achieve deep semantic interaction. To address this limitation, we propose a bidirectional cross-modal attention mechanism to overcome the technical bottleneck, as illustrated in Fig. 2. By constructing dual

information flow paths from image to text and from text to image, the model enables deep and collaborative modeling of cross-modal features.

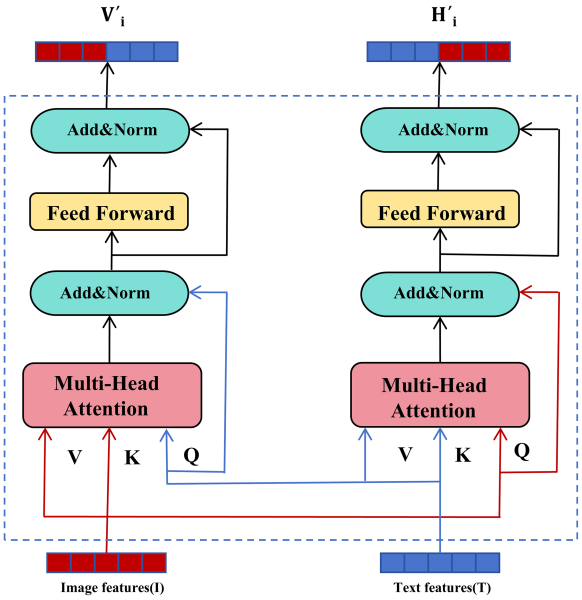

Figure 2: The cross-modal attention mechanism.

Image-to-Text Cross-Modal Attention

In the Image-to-Text cross-modal attention module, image features are used as queries, while textual features serve as keys and values. This design enables the model to attend to text regions that are relevant to the visual content. The computation of the image-to-text cross-modal multi-head attention follows the scaled dot-product attention formulation proposed by Vaswani et al. [39] and is formulated as follows:

$$head_h^{I2T} = \text{Softmax}\left(\frac{(V_i W_{I2T}^Q)(H_i W_{I2T}^K)^T}{\sqrt{d_k}}\right)(H_i W_{I2T}^V) \tag{6}$$

$$V_i' = \text{MultiHead}_{I2T}(V_i, H_i, H_i) = \text{Concat}(head_1^{I2T}, \ldots, head_{n_{heads}}^{I2T}) W_{I2T}^O \tag{7}$$

where $W_{I2T}^Q \in \mathbb{R}^{d_v \times d_k}$, $W_{I2T}^K \in \mathbb{R}^{d_h \times d_k}$, and $W_{I2T}^V \in \mathbb{R}^{d_h \times d_v}$ are learnable projection matrices, and $W_{I2T}^O \in \mathbb{R}^{n_{\text{heads}} \times d_v \times d_v}$ is the output projection matrix. Here, d_k denotes the dimensionality of the attention mechanism.

Text-to-Image Cross-Modal Attention

In the Text-to-Image cross-modal attention module, textual features are used as queries, while visual features serve as keys and values. This design enables the model to attend to image regions that are relevant to the textual content. The computation of the text-to-image multi-head attention mechanism is formulated as follows:

$$head_h^{T2I} = \text{Softmax}\left(\frac{(H_i W_{T2I}^Q)(V_i W_{T2I}^K)^T}{\sqrt{d_k}}\right)(V_i W_{T2I}^V) \tag{8}$$

$$H_i' = \text{MultiHead}_{T2I}(H_i, V_i, V_i) = \text{Concat}(head_1^{T2I}, \ldots, head_{n_{heads}}^{T2I}) W_{T2I}^O \tag{9}$$

where $W_{T2I}^Q \in \mathbb{R}^{d_h \times d_k}$, $W_{T2I}^K \in \mathbb{R}^{d_h \times d_k}$, and $W_{T2I}^V \in \mathbb{R}^{d_v \times d_v}$ are learnable projection matrices, and $W_{T2I}^O \in \mathbb{R}^{n_{\text{heads}} \times d_v \times d_h}$ is the output projection matrix.

Cross-Attention over CLIP Features

Building upon CLIP-based feature extraction, this module introduces a cross-attention mechanism to enable deep semantic interaction between the visual and textual modalities. Previously, we captured second-order interactions between image and text features via the tensor outer product and extracted key auto-correlation terms through a diagonalization operation, thereby constructing a compact bilinear representation. On this basis, cross-attention further aggregates and recalibrates cross-modal information, producing a fused representation with higher information density and stronger discriminative power.

$$z_i = W_p U_i + b_p, \quad z_i \in \mathbb{R}^d, \tag{10}$$

$$F_{CLIP} = \text{MultiHeadAttn}(z_i). \tag{11}$$

where $W_p \in \mathbb{R}^{d \times d_c}$ and $b_p \in \mathbb{R}^d$ denote the learnable projection matrix and bias term, respectively, and d is the hidden dimension of the attention module.

3.2.3 Cross-Modal Interactive Gating Mechanism

During multimodal feature fusion—particularly between image and text modalities—the contribution of each modality to final sentiment polarity typically exhibits marked imbalance. This imbalance is compounded by inter-modal redundancy, noise, and cross-modal feature misalignment, which undermine conventional fusion strategies and can lead to loss of complementary information or suboptimal representations. To address this issue, we propose a cross-modal gated fusion mechanism that dynamically adjusts the contribution of each modality according to the current context, as illustrated in Fig. 3. The mechanism introduces learnable attention-gating units that adaptively adjust fusion weights during training, selectively amplifying sentiment-relevant shared signals while suppressing redundant or noisy components.

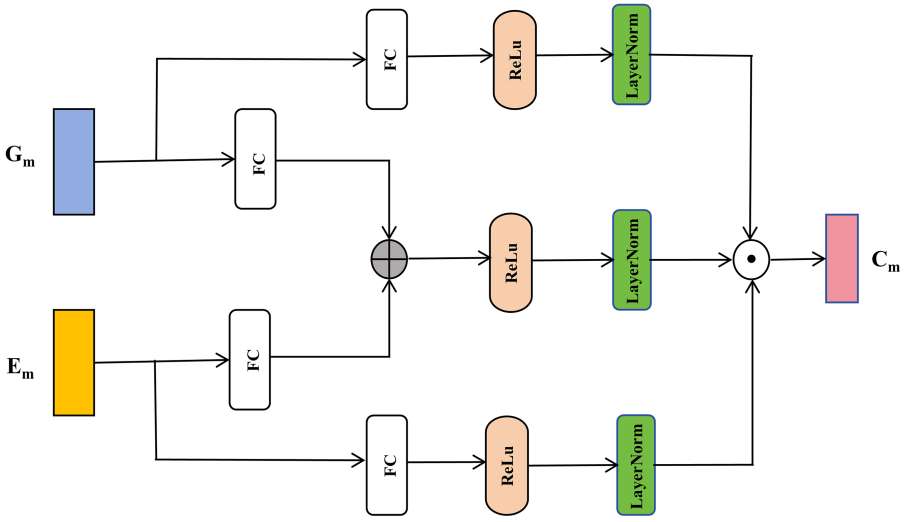

Figure 3: The cross-modal gated mechanism.

Concretely, the gating mechanism first computes a gating signal from the image–text features that reflects the relative importance of each modality in the current context. The computed gate is then used to weight the features, and the gated outputs are applied via element-wise modulation to the target modality feature maps. Finally, the gated features are concatenated and integrated to form a more discriminative multimodal joint representation. This adaptive gating strategy increases fusion flexibility and partially

mitigates modality misalignment, yielding more robust and information-rich features for subsequent sentiment classification.

To adaptively fuse the global semantic representation with bidirectional cross-modal interaction features, we design a gated fusion module. Given F_{CLIP} and a cross-modal feature $X \in \{V_i', H_i'\}$, the gated fusion is defined as follows.

First, three transformation branches are constructed:

$$T_G = \text{LN}\big(\text{ReLU}(W_G F_{\text{CLIP}} + b_G)\big), \tag{12}$$

$$T_X = \text{LN}\big(\text{ReLU}(W_X X + b_X)\big), \tag{13}$$

$$T_C = \text{LN}\big(\text{ReLU}(W_C^{(G)} F_{\text{CLIP}} + W_C^{(X)} X + b_C)\big), \tag{14}$$

where $W_G, W_X, W_C^{(G)}, W_C^{(X)} \in \mathbb{R}^{d\times d}$ and $b_G, b_X, b_C \in \mathbb{R}^d$ are learnable parameters.

Then, the joint gated interaction feature is obtained via element-wise multiplication:

$$Z = T_G \odot T_X \odot T_C, \tag{15}$$

where $\odot$ denotes the Hadamard product.

To further enhance representation capability and stabilize optimization, a residual refinement mapping is applied:

$$C = \text{LN}\big(Z + \text{ReLU}(W_\Phi Z + b_\Phi)\big), \tag{16}$$

where $W_\Phi \in \mathbb{R}^{d\times d}$ and $b_\Phi \in \mathbb{R}^d$ are learnable parameters.

Accordingly, the gated fusion outputs for the two cross-modal directions are computed as:

$$C_i^{(V)} = \mathcal{G}(F_{\text{CLIP}}, V_i'), \qquad C_i^{(H)} = \mathcal{G}(F_{\text{CLIP}}, H_i'), \tag{17}$$

where $\mathcal{G}(F_{\text{CLIP}}, X)$ denotes the gated fusion function defined by Eqs. (12)–(16).

Finally, the two gated features are concatenated and projected to form the final fused representation:

$$\widetilde{F}_i = \text{LN}\big(W_f[C_i^{(V)} \| C_i^{(H)}] + b_f\big), \tag{18}$$

where $[\cdot\|\cdot]$ denotes the concatenation operation, $W_f \in \mathbb{R}^{d\times 2d}$ and $b_f \in \mathbb{R}^d$ are learnable parameters.

3.2.4 Sentiment Classification and Objective

Finally, given the fused multimodal representation **F**, we feed it into a classifier to predict the sentiment label $\hat{y}$. The model adopts a fully connected layer followed by a Softmax function for multi-class classification, and is trained with the cross-entropy loss $\mathcal{L}_{ce}$. The entire network is optimized in an end-to-end manner by minimizing this loss.

$$\hat{y} = \text{softmax}(\mathbf{W}\widetilde{\mathbf{F}}_i + b) \tag{19}$$

$$\mathcal{L}_s = \text{CrossEntropyLoss}\big(\hat{y}, y\big) \tag{20}$$

where **W** and b are learnable parameters.

3.2.5 Image Caption Generation Module

To more comprehensively extract visual information, we employ an image captioning module adapted from the Caption Transformer (CaTr) architecture proposed by Khan and Fu [35]. This module converts the raw image into a natural language description, enabling seamless integration with text-based language models without modifying the core architecture.

Specifically, the CaTr module employs ResNet-101 [40] as the convolutional backbone to extract visual features from the input image $I_i \in \mathbb{R}^{3\times H\times W}$. These features are subsequently projected to a lower-dimensional embedding space of $d = 256$ via a 1×1 convolutional layer. Fixed positional encodings are then added to the resulting feature map, which is passed through a stack of Transformer encoder layers adapted from the DETR architecture [41]. The encoder leverages multi-head self-attention to capture object-level dependencies across the image. Finally, the decoder produces a natural language description of the image through non-autoregressive generation.

4 Experiments

4.1 Datasets

To evaluate the performance of the proposed MB-CMCAG model, this study utilizes two widely adopted Multi-View Sentiment Analysis (MVSA) datasets: MVSA-Single and MVSA-Multiple [42]. Both datasets are sourced from Twitter and consist of image-text pairs, each annotated with sentiment labels: positive, neutral, and negative. The MVSA-Single dataset contains 5129 image-text pairs, each labeled by a single annotator. After preprocessing to remove samples with inconsistent sentiment labels across modalities, 4511 valid samples remain, including 2683 positive samples, 470 neutral samples, and 1358 negative samples. The MVSA-Multiple dataset comprises 19,600 image-text pairs, labeled by three annotators. The final sentiment label is determined through majority voting. After preprocessing, 17,024 valid samples are obtained, including 11,318 positive samples, 4408 neutral samples, and 1298 negative samples. To facilitate comparisons with other models, both datasets are randomly split into training, validation, and test sets in an 8:1:1 ratio. Table 1 provides detailed information on the dataset distribution.

Table 1: Detailed statistics of the MVSA-single and MVSA-multiple datasets.

Dataset	Split	Positive	Neutral	Negative	Total
MVSA-Single					
	Train	2146	376	1086	3608
	Valid	268	47	135	450
	Test	269	47	137	453
	Total	2683	470	1358	4511
MVSA-Multiple					
	Train	9054	3526	1038	13,618
	Valid	1131	440	129	1700
	Test	1133	442	131	1706
	Total	11,318	4408	1298	17,024

4.2 Experimental Settings and Hyperparameters

All experiments were conducted on a server equipped with an NVIDIA RTX 4090 GPU (24 GB VRAM). The model was implemented in PyTorch. All experimental results are averaged over three runs with different random seeds. For feature extraction of the original text and image captions, we used a pretrained BERT tokenizer with a maximum sequence length of 128; sequences longer than this were truncated and shorter ones were padded. For visual feature extraction, we employed an ImageNet-pretrained ViT-B/16 model with a patch size of 16×16. For the additional branch, we used the CLIP-ViT-B/32 pretrained model to extract complementary semantic features. Models were trained with the Adam optimizer using a learning rate of 5×10^{-5} and a weight decay of 1×10^{-4}. Owing to differences in dataset sizes, the batch size was set to 32 for MVSA-Single and 128 for MVSA-Multiple. Training proceeded for up to 30 epochs with early stopping based on validation performance and a patience of 5 epochs. Detailed hyperparameter settings are reported in Table 2.

Table 2: Hyperparameter settings.

Parameter	Value
Learning_Rate	5e−5
Optimizer	Adam
Batch size	32/128
Epochs	30
Dropout	0.3
Weight decay	1e−4

In terms of evaluation metrics, we employ Accuracy and Macro-F1 as the primary measures, which are standard metrics for multi-class sentiment analysis tasks. Accuracy measures the overall proportion of correctly classified samples, while Macro-F1 provides a balanced evaluation of class-wise precision and recall, making it more suitable for datasets with class imbalance. The calculation formulas are as follows:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i) \tag{21}$$

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \tag{22}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{23}$$

$$F1_c = 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \tag{24}$$

$$Macro\text{-}F1 = \frac{1}{|C|}\sum_{c \in C} F1_c \tag{25}$$

where N denotes the total number of test samples, $\hat{y}_i$ and y_i denote the predicted label and the ground-truth label of the i-th sample, respectively, $\mathbb{I}(\cdot)$ is the indicator function, and C denotes the set of sentiment classes. For each class c, TP_c, FP_c, and FN_c are computed in a one-vs.-rest manner.

4.3 Model Comparisons and Baselines

To demonstrate the effectiveness of the proposed MB-CMCAG model, we conduct comparative experiments against representative baseline methods from the literature. The performance differences are quantitatively evaluated from both unimodal and multimodal perspectives using two core metrics, namely Accuracy and Macro-F1. The comparative results are reported in Table 3. The baseline methods included in the comparison are described as follows:

Table 3: Comparison with other models.

Modality	Method	MVSA-Single		MVSA-Multiple	
		ACC	F1	ACC	F1
Text	BERT	71.31	69.71	67.57	66.27
	BiLSTM	70.14	65.11	67.52	66.82
Image	ViT	63.95	62.44	62.12	61.24
	ResNet-50	64.72	61.64	61.93	61.05
Text and Image	MultiSentiNet	69.84	69.63	68.86	68.11
	Co-Memory	70.16	70.43	69.78	69.94
	DMAF	71.87	71.59	70.47	70.38
	MVAN	73.08	73.12	72.42	72.35
	ITIN	75.14	74.77	73.38	**73.26**
	CLMLF	75.33	73.45	72.00	69.83
	CLIP-CA-CG	75.36	75.18	73.56	73.79
	SentiCLIP	76.05	75.27	73.32	70.45
	VILA	76.21	75.39	73.26	72.91
	LLaVAC	75.82	74.91	73.15	72.68
	MB-CMCAG	**76.38**	**75.46**	**73.87**	72.64

Note: *indicates statistically significant improvement over the strongest baseline in the corresponding column based on three independent runs (paired t-test, $p < 0.05$). Bold denotes the best result. All results are averaged over three independent runs with different random seeds.

BERT and BiLSTM [22]: BERT and BiLSTM represent classical methods for modeling textual modalities, with extensive applications in text classification tasks.

ViT and ResNet-50 [40]: ResNet-50 and ViT represent the most prevalent benchmark models in image classification.

MultiSentiNet [43]: This model introduces visual semantic features (objects and scenes) as auxiliary information and employs a visually guided attention LSTM to mitigate semantic loss and weak image-text correlation caused by naive feature fusion.

Co-Memory [44]: This work addresses insufficient image-text interaction in multimodal sentiment analysis by proposing a co-memory network, where text guides visual feature selection and images attend to key words, with stacked layers for iterative refinement.

DMAF [45]: This model uses unimodal attention to highlight sentiment regions and words, and integrates intermediate and late fusion for robust multimodal sentiment prediction.

MVAN [46]: This paper proposes a Multi-View Attention Network that jointly models object-level and scene-level views of images and employs an interactive learning mechanism to improve multimodal sentiment analysis performance.

CLMLF [47]: This method employs a Transformer-based multi-layer fusion module to align and fuse textual and visual features at the token level, and incorporates label- and data-driven contrastive learning to enhance the learning of sentiment-discriminative shared representations.

CLIP-CA-CG [48]: This author builds a CLIP-based cross-modal sentiment model that extracts visual and textual features with ResNet50 and RoBERTa, enhances cross-modal interaction via multi-head attention, and fuses multi-level features using a cross-modal gating module.

SentiCLIP [38]: This multimodal sentiment analysis model leverages the CLIP encoder to construct a unified semantic space and employs a feature interaction module to enhance cross-modal emotion understanding.

ITIN [49]: The paper proposes an image-text interaction network that enhances multimodal sentiment analysis accuracy through fine-grained alignment of image regions and text words, along with adaptive feature fusion.

VILA [50]: The paper significantly improves visual language model performance and surpasses LLaVA-1.5 by unfreezing the LLM, adopting interleaved data, and using mixed text instruction fine-tuning.

LLaVAC [51]: The paper designs a structured prompt incorporating independent image and text sentiment labels as well as a joint multimodal label, and performs single-round LoRA fine-tuning on LLaVA to directly use it as a multimodal sentiment classifier.

The results in Table 3 show that MB-CMCAG achieves statistically significant improvements over the strongest baselines on the corresponding metrics, further confirming the robustness of the proposed method. Although MB-CMCAG achieves the best accuracy on MVSA-Multiple, its F1-score is slightly lower than that of some comparison methods. This is mainly because MVSA-Multiple is a more challenging dataset, where image-text semantic inconsistencies, ambiguous sentiment expressions, and class imbalance are more pronounced. Since F1-score is more sensitive to minority-class recognition than Accuracy, a small fluctuation in difficult categories may lead to a lower F1 even when overall accuracy remains strong. Therefore, this result does not contradict the effectiveness of MB-CMCAG, but rather reflects the difficulty of the dataset and the remaining room for improvement in minority-class discrimination.

Furthermore, to more intuitively illustrate the discriminative capability of the MB-CMCAG model across different categories, we plot the corresponding confusion matrices on the MVSA-Single and MVSA-Multiple datasets, respectively. The confusion matrices indicate that the model achieves higher classification accuracy on positive and negative samples, whereas its performance on the neutral category is relatively weaker, as shown in Fig. 4. This observation is consistent with the slightly lower F1-score on MVSA-Multiple, since errors in minority classes have a larger impact on Macro-F1 than on overall Accuracy. Future work may explore targeted data augmentation and cost-sensitive learning to further improve neutral-class recognition and alleviate class imbalance.

4.4 Ablation Study

To systematically evaluate the contribution of each core component in the proposed MB-CMCAG, we conducted a seven-group ablation study on both the MVSA-Single and MVSA-Multiple datasets. In each variant, one key module is removed or replaced while all other components remain unchanged. The seven settings are: (1) Text Only, (2) Image Only, (3) MB-CMCAG without CLIP (w/o CLIP), (4) MB-CMCAG

without BERT-ViT (w/o BV), (5) MB-CMCAG without Cross-Modal Cross-Attention (w/o CMCA), (6) MB-CMCAG without Gated Cross-Attention (w/o GCA), and (7) MB-CMCAG without Caption Generation Module (w/o CGM). Performance is evaluated using Accuracy and Macro-F1. The detailed experimental settings and complete results are reported in Table 4.

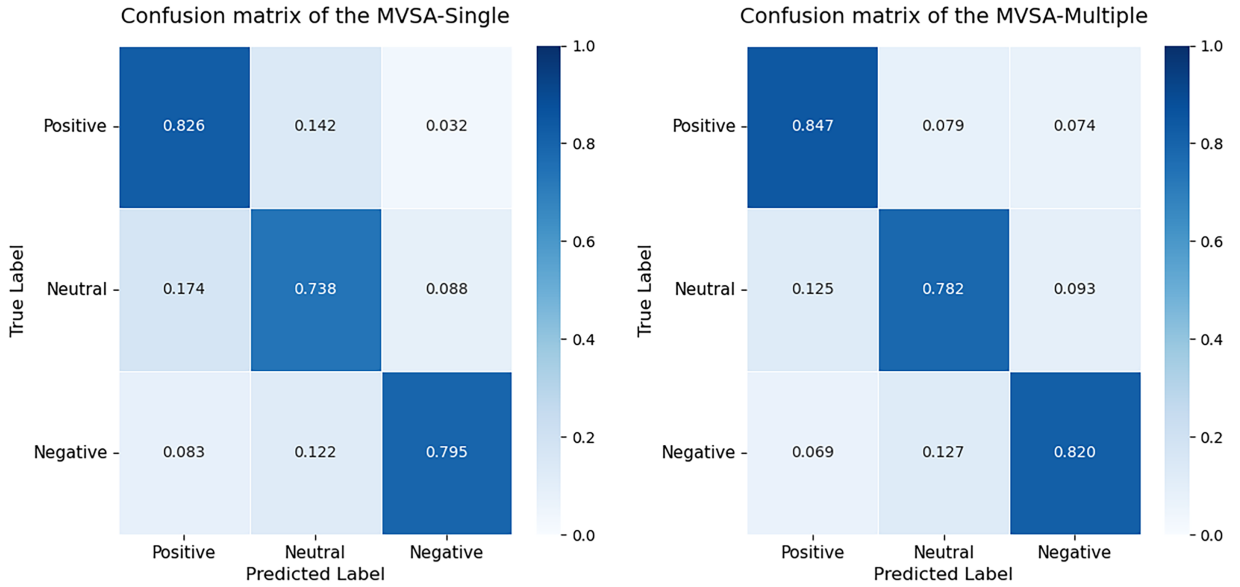

Figure 4: Confusion matrices of the MB-CMCAG model on the MVSA-Single and MVSA-Multiple datasets.

Table 4: Ablation studies of the MB-CMCAG model.

Model	MVSA-Single		MVSA-Multiple	
	Accuracy	F1	Accuracy	F1
Text Only	71.68	70.56	69.12	68.56
Image Only	69.97	69.05	68.64	68.17
MB-CMCAG w/o CLIP	73.85	72.46	71.22	69.59
MB-CMCAG w/o BV	72.77	71.98	70.29	69.32
MB-CMCAG w/o CMCA	72.48	70.53	70.74	69.18
MB-CMCAG w/o GCA	74.81	72.03	72.58	70.39
MB-CMCAG w/o CGM	75.79	74.82	73.06	71.96
MB-CMCAG	76.38	75.46	73.87	72.64

The ablation results clearly show that every module contributes positively to the final performance. Removing any single component leads to a consistent decrease in both Accuracy and Macro-F1 on the two datasets, confirming the necessity and complementarity of the proposed design. Specifically, the unimodal variants (Text Only and Image Only) suffer the largest performance drops, with average accuracy decreases of 4.73% and 5.82%, respectively, highlighting the importance of multimodal fusion. Removing the CLIP branch (w/o CLIP) or the fine-grained BV branch (w/o BV) reduces average accuracy by 2.59% and 3.59%, respectively, indicating that coarse-grained global alignment and fine-grained local modeling are both essential. Eliminating the cross-modal cross-attention module (w/o CMCA) causes a 3.52% drop, showing

that bidirectional fine-grained interaction is critical for semantic alignment. Replacing the gated mechanism with simple concatenation (w/o GCA) leads to a 1.43% decline, demonstrating the effectiveness of adaptive noise suppression and feature re-weighting. Finally, removing the image caption generation module (w/o CGM) results in a 0.70% decrease, confirming that visual-to-textual semantic enrichment provides useful complementary information. Overall, these results verify that the performance gains of MB-CMCAG arise from the coordinated interaction of its modules rather than from any single component alone, which further supports the effectiveness of the proposed hierarchical design.

4.5 Computational Efficiency Analysis

To further evaluate the computational efficiency of the proposed method, we report the model size, computational cost in terms of Floating Point Operations per Second (FLOPs), inference time, and GPU memory usage in Table 5. All measurements are conducted on the same server under the same settings as Section 4.2, with a batch size of 1 to ensure fair inference time comparison.

Table 5: Computational complexity and efficiency analysis.

Model Variant	Parameters (M)	FLOPs (G)	Inference Time (ms)	GPU Memory (GB)
BERT only	110	11.2	18.4	2.1
ViT-B/16 only	86	15.8	22.6	3.4
CLIP-ViT-B/32 only	151	12.5	19.7	2.8
MB-CMCAG (full model)	347	39.6	45.3	6.9
LLaVAC	7000	520	248	14.2
VILA	7500	580	265	15.1

Although MB-CMCAG integrates three pre-trained backbones, its total parameter count (347M) and inference time (45.3 ms) remain significantly lower than recent large vision-language models such as LLaVAC and VILA (both over 7B parameters). The cross-modal gated mechanism effectively suppresses redundant computation, resulting in a practical trade-off between performance and efficiency. Compared with these large-scale VLMs, our model achieves competitive accuracy while requiring only about 5% of the parameters and less than 20% of the inference time, demonstrating its suitability for real-world deployment on standard GPUs.

4.6 Case Study

To further illustrate the effectiveness of MB-CMCAG in multimodal text-image classification tasks, we selected four representative cases from the MVSA-Single test set. These examples are used to qualitatively analyze the model's ability to capture subtle emotional cues in text-image pairs. As shown in Fig. 5, we compare our proposed model with CLIP-CA-CG and SentiCLIP on these four cases. In addition, we input image captions as auxiliary sentences to further examine their effect on sentiment prediction.

First, all three models correctly classify the first and second examples as positive and negative, respectively, since the image-text pairs in these cases convey clear and consistent emotional signals. In the third example, where both modalities express no obvious sentiment tendency, CLIP-CA-CG predicts negative and SentiCLIP predicts positive, whereas our model correctly identifies the sample as neutral. This result suggests that the proposed multi-branch feature extraction strategy can better preserve complementary semantic information from both modalities, thereby improving the recognition of neutral samples with ambiguous emotional cues. In the fourth example, both CLIP-CA-CG and SentiCLIP classify the sample

as positive, mainly because the image itself does not explicitly express a negative tendency while the text suggests positive sentiment. In contrast, our model correctly predicts negative, which indicates that the image captioning module can provide additional visual semantic cues and help the model handle image-text inconsistency more effectively.

Label	Positive	Negative	Neutral	Negative
CLIP-CA-CG	**Positive**	**Negative**	Negative	Positive
SentiCLIP	**Positive**	**Negative**	Positive	Positive
MB-CMCAG	**Positive**	**Negative**	**Neutral**	**Negative**

Figure 5: Classification case studies of CLIP-CA-CG, SentiCLIP, and MB-CMCAG on the MVSA dataset.

In summary, these representative cases show that MB-CMCAG can better capture complementary sentiment cues from images and texts, especially in challenging samples with weak or inconsistent modality alignment. The results also indicate that image captions as auxiliary textual descriptions can further enhance cross-modal understanding and improve sentiment classification in complex scenarios.

5 Conclusion

In this study, we propose a multimodal sentiment analysis framework, MB-CMCAG, based on a multi-branch cross-modal cross-attention gated network. The framework addresses several key challenges in feature extraction, cross-modal alignment, and fusion. It first converts visual information into supplementary text via an image captioning model, thereby enriching the semantic representation of the original textual input. Next, it employs a dual-path parallel feature extraction architecture: one path leverages ViT and BERT for fine-grained visual and textual modeling, while the other exploits a pre-trained CLIP model to extract cross-modal coarse-grained features and adopts its diagonal values as an efficient semantic representation. Cross-modal cross-attention and multi-head attention modules are then specifically designed for each path to strengthen inter-modal semantic alignment. Finally, a cross-modal interaction gated fusion mechanism integrates the features from both paths, which are subsequently fed into a classification layer for sentiment prediction.

Extensive experiments on the MVSA benchmark datasets show that MB-CMCAG achieves competitive performance compared to existing methods in terms of accuracy and F1 score. Although the framework demonstrates promising results, future work may extend it to additional modalities such as video and audio, enabling the capture of richer multidimensional emotional expressions.

Acknowledgement: This work was supported by the National Natural Science Foundation of China and the Open Fund of the Guangxi Key Laboratory of Digital Infrastructure.

Funding Statement: This research project was supported by the National Natural Science Foundation of China (62266004) and the Open Fund of the Key Laboratory of Digital Infrastructure in Guangxi (GXDINBC202401).

Author Contributions: Xinshan Huang conceived the idea, designed the proposed model, implemented the framework, conducted the experiments, and drafted the manuscript. Zirui Pei and Chaohong Tan assisted with experimental validation and manuscript revision. Zuqiang Meng provided conceptual guidance and funding support. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The MVSA dataset used in this study is publicly available and can be accessed through [42].

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang S, Yang Y, Chen C, Zhang X, Leng Q, Zhao X. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. Expert Syst Appl. 2024;237:121692.
2. Chowdary MK, Nguyen TN, Hemanth DJ. Deep learning-based facial emotion recognition for human-computer interaction applications. Neural Comput Appl. 2023;35(32):23311–28. doi:10.1007/s00521-021-06012-8.
3. Zhu L, Zhu Z, Zhang C, Xu Y, Kong X. Multimodal sentiment analysis based on fusion methods: A survey. Inform Fus. 2023;95:306–25. doi:10.1016/j.inffus.2023.02.028.
4. Pandey A, Vishwakarma DK. Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey. Appl Soft Comput. 2024;152:111206. doi:10.1016/j.asoc.2023.111206.
5. Chen B, Cao Q, Hou M, Zhang Z, Lu G, Zhang D. Multimodal emotion recognition with temporal and semantic consistency. IEEE ACM Trans Audio Speech Lang Process. 2021;29:3592–603. doi:10.1109/taslp.2021.3129331.
6. Dixit C, Satapathy SM. Deep CNN with late fusion for real time multimodal emotion recognition. Expert Syst Appl. 2024;240:122579. doi:10.1016/j.eswa.2023.122579.
7. Aslam A, Sargano AB, Habib Z. Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks. Appl Soft Comput. 2023;144:110494. doi:10.1016/j.asoc.2023.110494.
8. Du Y, Liu Y, Peng Z, Jin X. Gated attention fusion network for multimodal sentiment classification. Knowl Based Syst. 2022;240:108107. doi:10.1016/j.knosys.2021.108107.
9. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. Cambridge, MA, USA: PMLR; 2021. p. 8748–63.
10. Nie Z. Feature-attentive multimodal emotion analyzer with CLIP. In: 2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). Piscataway, NJ, USA: IEEE; 2023. p. 317–23.
11. Dosovitskiy A. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Kerrville, TX, USA: ACL; 2019. p. 4171–86.
13. Kanayama H, Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Kerrville, TX, USA: ACL; 2006. p. 355–63.

14. Alwan JK, Hussain AJ, Abd DH, Sadiq AT, Khalaf M, Liatsis P. Political Arabic articles orientation using rough set theory with sentiment lexicon. IEEE Access. 2021;9:24475–84. doi:10.1109/access.2021.3054919.
15. Esuli A, Sebastiani F. SentiWordNet: a publicly available lexical resource for opinion mining. In: International Conference on Language Resources and Evaluation (LREC 2006). Paris, France: ELRA; 2006. p. 417–22.
16. Hamouda A, Rohaim M. Reviews classification using sentiwordnet lexicon. In: Proceedings of the World Congress on Computer Science and Information Technology (WCSIT). Dubai, United Arab Emirates: WCSIT; 2011. p. 104–5.
17. Naz S, Sharan A, Malik N. Sentiment classification on twitter data using support vector machine. In: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). Piscataway, NJ, USA: IEEE; 2018. p. 676–9.
18. Xu S. Bayesian Naïve Bayes classifiers to text classification. J Inform Sci. 2018;44(1):48–59. doi:10.1177/0165551516677946.
19. Goel A, Gautam J, Kumar S. Real time sentiment analysis of tweets using Naive Bayes. In: 2016 2nd International Conference on Next Generation Computing Technologies (NGCT). Piscataway, NJ, USA: IEEE; 2016. p. 257–61.
20. Rathor AS, Agarwal A, Dimri P. Comparative study of machine learning approaches for Amazon reviews. Procedia Comput Sci. 2018;132:1552–61. doi:10.1016/j.procs.2018.05.119.
21. Kim Y. Convolutional neural networks for sentence classification. arXiv:1408.5882. 2014.
22. Zhou X, Wan X, Xiao J. Attention-based LSTM network for cross-lingual sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Kerrville, TX, USA: ACL; 2016. p. 247–56.
23. Wang J, Yu LC, Lai KR, Zhang X. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. IEEE ACM Trans Audio Speech Lang Process. 2019;28:581–91. doi:10.1109/taslp.2019.2959251.
24. Siersdorfer S, Minack E, Deng F, Hare J. Analyzing and predicting sentiment of images on the social web. In: Proceedings of the 18th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2010. p. 715–8.
25. Borth D, Ji R, Chen T, Breuel T, Chang SF. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM International Conference on Multimedia. New York, NY, USA: ACM; 2013. p. 223–32.
26. Yang Y, Jia J, Zhang S, Wu B, Chen Q, Li J, et al. How do your friends on social media disclose your emotions? Proc AAAI Conf Artif Intell. 2014;28(1):306–12.
27. Song K, Yao T, Ling Q, Mei T. Boosting image sentiment analysis with visual attention. Neurocomputing. 2018;312:218–28. doi:10.1016/j.neucom.2018.05.104.
28. Islam J, Zhang Y. Visual sentiment analysis for social images using transfer learning approach. In: 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom). Piscataway, NJ, USA: IEEE; 2016. p. 124–30.
29. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2015. p. 1–9.
30. Abdullah SMSA, Ameen SYA, Sadeeq MA, Zeebaree S. Multimodal emotion recognition using deep learning. J Appl Sci Technol Trends. 2021;2(1):73–9. doi:10.38094/jastt20291.
31. You Q, Luo J, Jin H, Yang J. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. New York, NY, USA: ACM; 2016. p. 13–22.
32. Zhao Z, Zhu H, Xue Z, Liu Z, Tian J, Chua MCH, et al. An image-text consistency driven multimodal sentiment analysis approach for social media. Inform Process Manage. 2019;56(6):102097. doi:10.1016/j.ipm.2019.102097.
33. Zhang K, Geng Y, Zhao J, Liu J, Li W. Sentiment analysis of social media via multimodal feature fusion. Symmetry. 2020;12(12):2010. doi:10.3390/sym12122010.
34. Zhou S, Wu X, Jiang F, Huang Q, Huang C. Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks. Int J Environ Res Public Health. 2023;20(2):1400. doi:10.3390/ijerph20021400.

35. Khan Z, Fu Y. Exploiting BERT for multimodal target sentiment classification through input space translation. In: Proceedings of the 29th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2021. p. 3034–42.
36. Li G, Duan N, Fang Y, Gong M, Jiang D. Unicoder-VL: a universal encoder for vision and language by cross-modal pre-training. Proc AAAI Conf Artif Intell. 2020;34(7):11336–44.
37. Huang Q, Cai P, Nie T, Zeng J. Clip-MSA: incorporating inter-modal dynamics and common knowledge to multimodal sentiment analysis with clip. In: ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE; 2024. p. 8145–9.
38. An J, Ding B, Wan Zainon WMN. Improving multimodal sentiment prediction through vision-language feature interaction. Multimed Syst. 2025;31(1):63. doi:10.1007/s00530-024-01659-4.
39. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc.; 2017. p. 6000–10.
40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2016. p. 770–8.
41. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: European Conference on Computer Vision. Cham, Switzerland: Springer; 2020. p. 213–29.
42. Niu T, Zhu S, Pang L, El Saddik A. Sentiment analysis on multi-view social data. In: International Conference on Multimedia Modeling. Cham, Switzerland: Springer; 2016. p. 15–27.
43. Xu N, Mao W. MultiSentiNet: a deep semantic network for multimodal sentiment analysis. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. New York, NY, USA: ACM; 2017. p. 2399–402.
44. Xu N, Mao W, Chen G. A co-memory network for multimodal sentiment analysis. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York, NY, USA: ACM; 2018. p. 929–32.
45. Huang F, Zhang X, Zhao Z, Xu J, Li Z. Image-text sentiment analysis via deep multimodal attentive fusion. Knowl-Based Syst. 2019;167:26–37. doi:10.1016/j.knosys.2019.01.019.
46. Yang X, Feng S, Wang D, Zhang Y. Image-text multimodal emotion classification via multi-view attentional network. IEEE Trans Multimed. 2020;23:4014–26. doi:10.1109/tmm.2020.3035277.
47. Li Z, Xu B, Zhu C, Zhao T. CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection. arXiv:2204.05515. 2022.
48. Lu X, Ni Y, Ding Z. Cross-modal sentiment analysis based on CLIP image-text attention interaction. Int J Adv Comput Sci Appl. 2024;15(2):895–903. doi:10.14569/ijacsa.2024.0150290.
49. Zhu T, Li L, Yang J, Zhao S, Liu H, Qian J. Multimodal sentiment analysis with image-text interaction network. IEEE Trans Multimed. 2022;25:3375–85. doi:10.1109/tmm.2022.3160060.
50. Lin J, Yin H, Ping W, Molchanov P, Shoeybi M, Han S. VILA: on pre-training for visual language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2024. p. 26689–99.
51. Chay-intr T, Chen Y, Viriyayudhakorn K, Theeramunkong T. LLaVAC: fine-tuning LLaVA as a multimodal sentiment classifier. arXiv:2502.02938. 2025.