*electronics*

*Article*

# Multimodal Alignment and Hierarchical Fusion Network for Multimodal Sentiment Analysis

**Jiasheng Huang †, Huan Li *,† and Xinyue Mo ***

School of Cyberspace Security (School of Cryptology), Hainan University, Haikou 570228, China; 23210839000014@hainanu.edu.cn

* Correspondence: lihuan@hainanu.edu.cn (H.L.); moxinyue@hainanu.edu.cn (X.M.)

† These authors contributed equally to this work.

**Abstract**

The widespread emergence of multimodal data on social platforms has presented new opportunities for sentiment analysis. However, previous studies have often overlooked the issue of detail loss during modal interaction fusion. They also exhibit limitations in addressing semantic alignment challenges and the sensitivity of modalities to noise. To enhance analytical accuracy, a novel model named MAHFNet is proposed. The proposed architecture is composed of three main components. Firstly, an attention-guided gated interaction alignment module is developed for modeling the semantic interaction between text and image using a gated network and a cross-modal attention mechanism. Next, a contrastive learning mechanism is introduced to encourage the aggregation of semantically aligned image-text pairs. Subsequently, an intra-modality emotion extraction module is designed to extract local emotional features within each modality. This module serves to compensate for detail loss during interaction fusion. The intra-modal local emotion features and cross-modal interaction features are then fed into a hierarchical gated fusion module, where the local features are fused through a cross-gated mechanism to dynamically adjust the contribution of each modality while suppressing modality-specific noise. Then, the fusion results and cross-modal interaction features are further fused using a multi-scale attention gating module to capture hierarchical dependencies between local and global emotional information, thereby enhancing the model's ability to perceive and integrate emotional cues across multiple semantic levels. Finally, extensive experiments have been conducted on three public multimodal sentiment datasets, with results demonstrating that the proposed model outperforms existing methods across multiple evaluation metrics. Specifically, on the TumEmo dataset, our model achieves improvements of 2.55% in ACC and 2.63% in F1 score compared to the second-best method. On the HFM dataset, these gains reach 0.56% in ACC and 0.9% in F1 score, respectively. On the MVSA-S dataset, these gains reach 0.03% in ACC and 1.26% in F1 score. These findings collectively validate the overall effectiveness of the proposed model.

**Keywords:** multimodal; multimodal sentiment analysis; multi-level fusion; gated networks; contrastive learning



Academic Editor: Ioannis Hatzilygeroudis

Received: 19 August 2025
Revised: 13 September 2025
Accepted: 25 September 2025
Published: 26 September 2025

**Citation:** Huang, J.; Li, H.; Mo, X. Multimodal Alignment and Hierarchical Fusion Network for Multimodal Sentiment Analysis. *Electronics* **2025**, *14*, 3828. https://doi.org/10.3390/electronics14193828



## 1. Introduction

In recent years, the rapid development of social platforms has established them as a primary medium for the expression of emotional opinions. At present, users express their thoughts and emotions not only through text but also through diverse modalities,

including images and videos. Initially, sentiment analysis was primarily based on text mining. However, sentiment data on contemporary social platforms frequently exist in multimodal forms, making it challenging to comprehensively capture sentiment solely through text-based mining approaches. Single-modal sentiment analysis is often constrained by the inherent information limitations and feature sparsity of individual data sources; this hinders improvements in sentiment classification accuracy. In contrast, multimodal sentiment analysis can overcome these limitations by effectively integrating complementary features from different modalities and their latent inter-modal correlations, thereby producing more reliable emotion recognition outcomes. For example, emotional tendencies in videos can be inferred by leveraging textual, visual, and auditory cues. Nonetheless, multimodal approaches introduce additional technical challenges in processing and analyzing multi-source heterogeneous information [1]. As a result, multimodal sentiment analysis has garnered significant research attention; not only does it offer more precise emotion identification, but it is also widely applicable in domains such as rumor detection [2], public opinion monitoring [3], and content recommendation [4].

A key objective of multimodal sentiment analysis is to enhance consistent sentiment cues across modalities while leveraging complementary differences among modalities to improve prediction accuracy [5]. As illustrated in Figure 1, when both image and textual data convey the same emotion, the cross-modal association should be strengthened; conversely, when discrepancies exist, such differences should be preserved to enable the complementary integration of emotional signals. Numerous studies have focused on optimizing interactive fusion techniques for heterogeneous modalities. For instance, Hu et al. [6] achieved fusion through feature concatenation; Zhu et al. [7] investigated image-text relationships using a gating mechanism; and Xiao et al. [8] employed bidirectional interaction networks to facilitate modality interactions. Shen et al. [9] addressed two key issues: clinical depression diagnosis lacks objective indicators and is prone to doctors' and patients' subjectivity, while existing EEG-based automatic depression diagnosis methods fail to resolve EEG's high individual variability and lack recognition result uncertainty estimation. They proposed the UA-DAAN, which enhances the transferability of class-related features between domains, boosts model robustness, accuracy and reliability, with experimental results confirming its effectiveness in depression recognition. However, detail loss often occurs during the interaction and fusion of multimodal data. Most existing methods overlook this issue and neglect the significance of modality-specific local sentiment cues, thereby limiting their ability to achieve effective emotional complementarity. In addition, they remain inadequate in addressing semantic misalignment and modality-specific noise. These limitations constrain the emotional recognition capabilities of sentiment analysis models.

In summary, current multimodal sentiment analysis continues to face two core challenges. On the one hand, the significance of modality-specific local sentiment information is often overlooked in achieving emotional complementarity. On the other hand, limitations remain in the design of fusion strategies, resulting in challenges for semantic alignment and reduced robustness to noise. To address these issues, a novel method, Multimodal Alignment and Hierarchical Fusion Network (MAHFNet), is proposed. Firstly, we design an attention-guided gated interaction alignment module to model the semantic interaction between text and image via a cross-modal attention mechanism. Concurrently, a contrastive learning mechanism is introduced: it encourages semantically consistent image-text pairs to cluster and inconsistent ones to separate. This design enhances the model's ability to capture cross-modal emotional consistency and improves its robustness against noisy or conflicting information. Building on this foundation, an intra-modality emotion extraction module is employed to extract local emotional features from both textual and visual modal-

ities. These features capture fine-grained emotional cues, such as sentiment-laden words in text, or expressive regions and color tones in images, which are often overlooked in cross-modal interaction modeling yet critical for emotional understanding. Subsequently, the extracted local emotional features and cross-modal interaction features from each modality are fed into a hierarchical gated fusion module. Through a cross-gating mechanism, local emotional features are fused to suppress redundant or irrelevant information while emphasizing salient emotional cues. Finally, the fusion results and cross-modal interaction features are further integrated via a multi-scale attention gating module, aiming to capture hierarchical dependencies between local and global emotional information. This integration ultimately enhances the model's capacity to perceive and integrate emotions across multiple semantic levels.

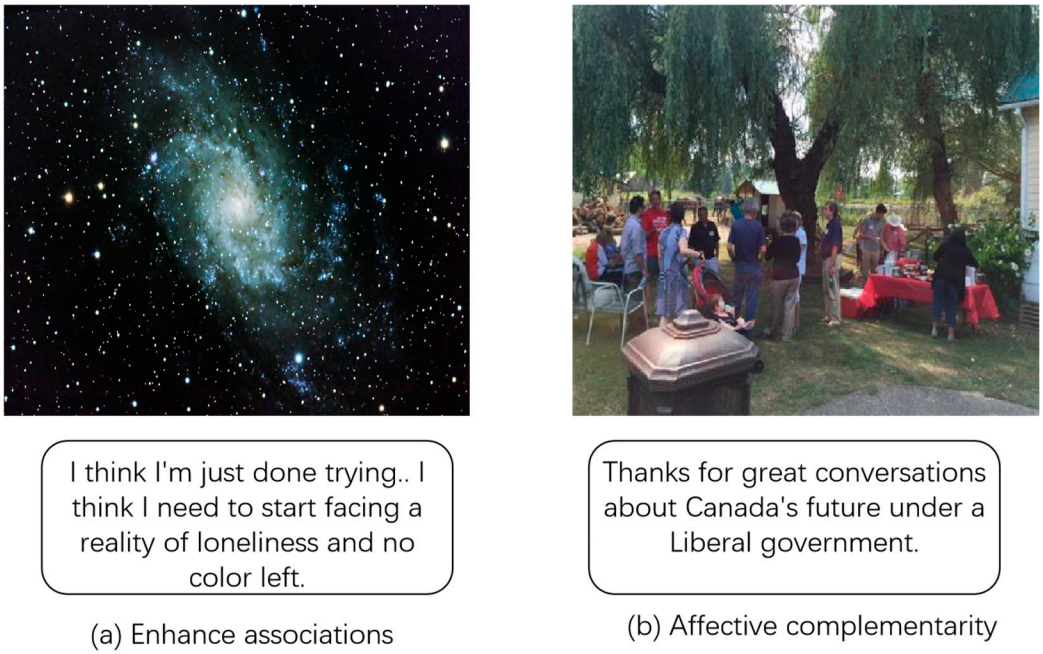


**Figure 1.** Examples of tweet data.

The main contributions of this paper are summarized as follows:

- A hierarchical fusion network with multimodal alignment is proposed, which constructs a dual-pathway structure for cross-modal alignment and modality-specific local emotion modeling. In addition, a hierarchical gated fusion mechanism is introduced to facilitate the multi-level integration of intra-modal local emotion features and shared cross-modal information, thereby enhancing emotion consistency modeling and multi-granularity emotional perception.
- Two sub-gated fusion modules are designed, namely the Cross-Gated Fusion module (CGF) and the Multi-Scale Attention Gating module (MAG). The CGF is employed to dynamically adjust emotional contributions across modalities, while the MAG integrates local and global emotion features to capture hierarchical semantic relationships in emotional expression.
- Extensive empirical studies conducted on two public multimodal sentiment analysis datasets demonstrate that the proposed model effectively captures both local and global emotional cues and facilitates comprehensive fusion, thereby improving both semantic alignment and robustness to noise. Experimental results indicate that MAHFNet achieves significant improvements over existing baseline methods.

The remainder of this paper is organized as follows: Section 2 comprehensively reviews pertinent existing literature; Section 3 details the methodology employed; Section 4 presents the experimental procedures and corresponding results; Section 5 evaluates and analyzes model performance based on the confusion matrix; Section 6 summarizes the main contributions of this paper and draws conclusions, while also pointing out the future work to be carried out.

## 2. Related Work

Effectively modeling the relationships between modalities is a central challenge in multimodal sentiment analysis. However, many existing methods rely on simplistic approaches such as direct feature concatenation or basic attention mechanisms for cross-modal interaction, which fail to adequately capture complex inter-modal dependencies, often resulting in suboptimal performance. Consequently, increasing research efforts have been devoted to developing more effective strategies for modality interaction and fusion. For example, Xu et al. [10] introduced the Multi-interactive memory network (MIMN), which captures correlations between textual and visual modalities through multiple layers of interactive attention and contextual memory.

With the emergence of Transformers [11] and numerous pre-trained models such as BERT [12] and ViT [13], multimodal sentiment analysis has experienced significant advancements. Hang et al. [14] introduced a text-centered fusion network with cross-modal attention (TeFNA). Wang et al. [15] introduced an end-to-end multimodal aspect-sentiment analysis framework involving an image-to-text conversion module that transforms visual data into BERT-compatible implicit token sequences, and incorporating an aspect-oriented filtering module for dynamic filtering of visual noise using a dual attention mechanism. Yang et al. [16] introduced a multi-grained fusion network with self-distillation model (MGFN-SD), addressing limitations such as coarse-grained semantic loss and the omission of image–aspect correlations through multi-granularity representation learning and a self-distillation mechanism. Kim et al. [17] introduced the AOBERT model for multimodal sentiment analysis. Yu et al. [18] presented a Hierarchical interactive multimodal transformer (HIMT) for modeling deep modality interactions. Chen et al. [19] introduced a Hierarchical Cross-modal Transformer (HCT) for modeling interactions between text and images using a Transformer-based architecture. Le et al. [20] introduced a Transformer-based fusion model designed to unify modality representations via joint learning.

Numerous studies have also employed Graph Convolutional Networks (GCNs) for multimodal sentiment prediction, leveraging graph structural information to model nodes and edges within multimodal data. This approach is particularly well-suited for capturing complex relationships and structural dependencies among modalities. For instance, Lu et al. [21] proposed a heterogeneous graph neural network (Hete-GNNs) model to share sentence sequence features through interactive aspect word encoding, and construct heterogeneous semantic graphs (integrating syntactic dependency trees, sentiment prior dictionaries and part-of-speech tagging) to solve the problem of semantic confounding in multi-target sentiment analysis. Lu et al. [22] proposed the Coordinated Joint Translation Fusion (CJTF) framework, which enhances textual emotion representation using an emotional interaction graph. In their approach, cross-modal masked attention is utilized to accurately identify shared semantic features, while a translation-awareness mechanism is employed to reconstruct modality-specific representations. However, the performance of GCN-based models remains highly dependent on the quality of the input graph structure. Inaccurate or incomplete connectivity among graph nodes can adversely affect overall model performance.

In addition, incorporating external knowledge has become a common strategy in multimodal sentiment analysis. Wang et al. [5] introduced the Multiple Attention-based

Multimodal Sentiment Analysis (MAMSA) framework, in which sentiment attention from a sentiment matrix is used to highlight emotion-relevant information in both text and image modalities. Zhou et al. [23] utilized SenticNet to generate sentiment scores, which were integrated into the multimodal fusion process. Dong et al. [24] proposed cross-modal feedback interactions based on a knowledge graph, which dynamically enhanced modality representations through a self-feedback feature masking mechanism and incorporated knowledge graph embeddings to supplement external semantic knowledge. Finally, multi-channel information was integrated using global feature fusion, significantly improving sentiment prediction accuracy in complex scenarios. However, approaches that rely on external knowledge typically introduce additional computational overhead and often yield limited practical gains.

Meanwhile, multi-level fusion and interaction at the segment level are becoming increasingly important in multimodal sentiment analysis. Liu et al. [25] introduced a cascaded multichannel and hierarchical fusion (CMC-HF) model, which effectively captures cross-modal interaction information through a hierarchical fusion framework. Yang et al. [26] employed stacked attention mechanisms to facilitate fusion via multiple interactions between text and image segments.

While existing works on MMSA have attempted to optimize cross-modal interaction effects through designs such as cross-attention mechanisms and multi-stage fusion, they still exhibit notable limitations in modeling complex emotional expressions. On one hand, although MAMSA achieves late-stage fusion via cross-attention and residual structures, its core focus lies on single-scale feature correlation, failing to fully account for the deep local-global interaction of emotional semantics. This results in insufficient adaptability to multi-granularity emotional expressions. On the other hand, the BIT model relies on repeated concatenation operations to complete final feature fusion. This static feature stacking approach lacks awareness of dynamic cross-modal relationships, making it unable to adaptively adjust the emotional contribution weights of different modalities according to the heterogeneity of emotional expressions, and thus prone to interference from redundant modal information.

To address these limitations, this study proposes the MAHFNet. Specifically, relying on the dual-path architecture of "cross-modal alignment—intra-modal local emotional modeling", a hierarchical gated fusion mechanism is constructed to realize the multi-level integration of intra-modal local emotional features and cross-modal shared information. To tackle the aforementioned specific issues, two modules are designed:

- Cross-Gated Fusion (CGF) Module: By dynamically adjusting the emotional contribution weights of multiple modalities, it effectively filters modal redundancy and enhances complementary information, thereby solving the problem of insufficient modeling of dynamic cross-modal relationships.
- Multi-Scale Attention Gating (MAG) Module: By integrating local and global emotional features, it establishes multi-granularity emotional semantic correlations, overcoming the limitation that single-scale attention cannot cover complex emotional expressions.
- Table 1 further compares the proposed MAHFNet with MAMSA and BIT from the perspective of core design dimensions of architectural components, clearly presenting the differentiated positioning of this study.

**Table 1.** Comparison of architectural components.

| Comparison Dimension | MAHFNet | MAMSA | BIT |
|---|---|---|---|
| Integration mechanism | Cross-Gated Fusion | Cross-attention + Residual Structure | concatenation |
| Attention mechanism | Multi-Scale Attention Gating | Cross-attention | Cross-attention |

## 3. Method

This section presents the overall architecture of the proposed MAHFNet. As shown in Figure 2, the model consists of three key components: (1) an Attention-guided Alignment and Interaction module (AGAI), (2) an Intra-modal Emotion Extraction module, and (3) a Hierarchical Gated Fusion module (HGFM). Detailed Configuration information can be found in Table 2.

The purpose of MSA tasks is to obtain sentiments by using multi-modal signals of text (T) and image (P). Generally speaking, MSA can be seen as a classification task or regression task. This task is considered as a classification task. Therefore, the model inputs text $T_i$ and image $P_i$, and then outputs a sentiment label $L_i$.

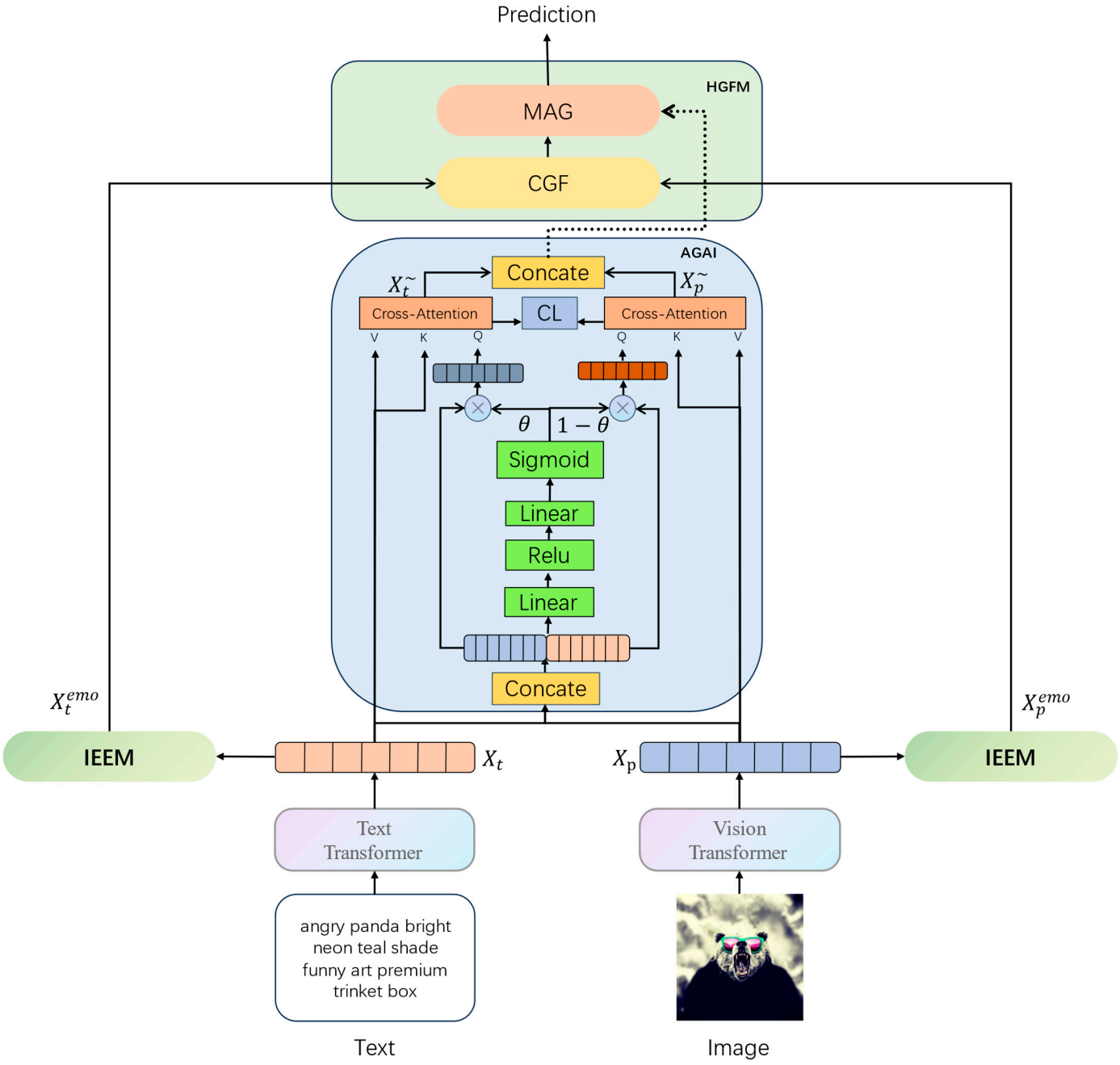


**Figure 2.** Structure diagram of MAHFNet.

**Table 2.** Configuration information.

| Configuration Information | Value |
|---|---|
| The number of attention heads | 8 |
| Layers of the IEEM | 2 |
| Model dimension d_model | 512 |
| Size of the hidden layer | 256 |
| Contrast projection layer activation function | ReLU (Contrastive Projection Layer/Dynamic Mode Weighted MLP); GELU (Cross Attention Projection Layer) |
| Normalization method | LayerNorm ($\varepsilon = 1 \times 10^{-5}$) |
| Kernel size of convolution | $3 \times 1$ |
| Padding method | Same padding |
| Dilation rate | 2 |
| order of LayerNorm | Pre-norm |
| Dropout rate | 0.1 |

### *3.1. Feature Extraction*

For the text modality, a sentence $T_i = \{W_1, \ldots W_j, \ldots W_k\}$ consists of a sequence of words. The text Transformer branch of CLIP [27] is employed; this branch has been pre-trained on 400 million image-text pairs. To acquire word-level features, the final pooled representation is discarded, while the token-level outputs are preserved. Each word is thereby encoded as a 512-dimensional feature vector:

$$X_t = TextTransformer(T_i) \tag{1}$$

where $X_t \in R^{k \times d}$, with $k \leq 77$ denoting the number of words in the sentence, and $d$ = 512 the feature dimension.

For the image modality, the Vision Transformer (ViT) architecture [13] is followed, wherein each image is divided into a sequence of 16 × 16 patches, which serve as visual tokens. To extract patch-level features, the vision branch of the CLIP model is adopted. Similarly to the text modality, the final pooled token is discarded, and the individual patch embeddings are retained. Each image patch is thus represented as a 512-dimensional feature vector:

$$X_p = VisionTransformer(P_i) \tag{2}$$

where $X_p \in R^{n \times d}$, $n$ = 196 is the number of patches, $d$ = 512 is the embedding dimension.

### *3.2. Attention-Guided Alignment and Interaction Module (AGAI)*

To capture semantic interactions between text and image while mitigating modality-specific noise and conflicts, an Attention-Guided Alignment and Interaction (AGAI) module is proposed. This module conducts fine-grained cross-modal interaction and contrastive alignment through a three-step process.

#### 3.2.1. Dynamic Modality Weighting

Given the input textual feature $X_t$ and visual feature $X_p$, they are first concatenated and processed by a multi-layer perceptron followed by a sigmoid function to produce a gating coefficient $\theta$,

$$\theta = \sigma\big(MLP\ \big([X_t; X_p]\big)\big) = \sigma\big(Linear\big(Dropout\big(ReLU\big(Linear\big[[X_t; X_p]\big]\big)\big)\big)\big) \tag{3}$$

where $\theta \in R^1$.

The gate is then used to adaptively modulate the contribution of each modality:

$$X_t' = \theta \cdot X_t \in R^{k \times d} \tag{4}$$

$$X_p' = (1 - \theta) \cdot X_p \in R^{k \times d} \tag{5}$$

This dynamic reweighting enables the network to selectively emphasize informative content from either modality based on contextual relevance. It effectively suppresses noise introduced during feature interaction and adaptively adjusts the contribution of each modality.

3.2.2. Bidirectional Cross-Modal Attention

To model the semantic alignment between modalities, two cross-attention mechanisms are applied. The textual feature attends to the visual feature and vice versa.

$$X_t^{\sim} = CrossAttn_t(Q = X_t', K = V = X_p) = softmax\left(\frac{(X_t'W_Q)(X_pW_K)^{\top}}{\sqrt{d_k}}\right)(X_pW_V) \tag{6}$$

$$X_p^{\sim} = CrossAttn_i\left(Q = X_p', K = V = X_t\right) = softmax\left(\frac{(X_t'W_Q)(X_pW_K)^{\top}}{\sqrt{d_k}}\right)(X_pW_V) \tag{7}$$

where $W_Q, W_K, W_V \in R^{d \times \mathrm{d_k}}$ is linear projection matrix. $d_k = 64$.

The attended features are concatenated and passed through a shared feed-forward projection layer to obtain the interaction feature:

$$X_c = \mathrm{Proj}([\ X_t^{\sim} : X_p^{\sim}]) = (Linear(GELU(Dropout(LayerNorm([\ X_t^{\sim} : X_p^{\sim}])))))) \tag{8}$$

where $\mathrm{X_c} \in R^d$ represents the aligned and fused representation of cross-modal information.

3.2.3. Contrastive Representation Alignment

To enhance semantic consistency and improve cross-modal discriminative capability, a contrastive learning strategy is incorporated. The attended features from each modality are pooled and projected into a shared embedding space for contrastive alignment.

$$\widetilde{x_t} = Proj_t(AvgPool\ (X_p^{\sim})) = (Linear(ReLU(Linear(LayerNorm(AvgPool\ (X_p^{\sim})))))) \tag{9}$$

$$\widetilde{x_p} = Proj_i(AvgPool\ (X_t^{\sim})) = (Linear(ReLU(Linear(LayerNorm((AvgPool\ (X_t^{\sim}))))))) \tag{10}$$

A symmetric InfoNCE-style contrastive loss is computed based on cosine similarity with temperature scaling:

$$\mathcal{L}_{con} = -\frac{1}{2}\left[log\frac{exp(sim(x_t, x_p^{+})/\tau)}{\sum_{x_i^{-}} exp\left(sim\left(x_t, x_p^{-}\right)/\tau\right)} + log\frac{exp\left(sim\left(x_p, x_t^{+}\right)/\tau\right)}{\sum_{x_t^{-}} exp\left(sim\left(x_p, x_t^{-}\right)/\tau\right)}\right] \tag{11}$$

where $sim(\cdot)$ denotes cosine similarity and $\tau$ is a temperature parameter. $x_p^{+}$ is a visual contrast feature that is derived from the same sample as $x_t$.$x_p^{-}$ is a visual contrast feature that is derived from the same sample as $x_t$.$sim(\cdot)$ represents cosine similarity.

The AGAI outputs the fused interaction feature $X_c$, which is passed to the subsequent fusion module, and the contrastive loss $\mathcal{L}_{con}$, which is used to jointly optimize modality alignment during training.

### *3.3. Intra-Modal Emotion Extraction Module (IEEM)*

While cross-modal interaction captures the high-level alignment between modalities, the fine-grained emotional cues embedded in each modality are often overlooked or lost in the process of intermodal interaction. To address this, an intra-modal emotion extraction module is applied to independently refine emotional representations within each modality.

Specifically, each modality is processed through a two-layer multi-head self-attention network to model local contextual dependencies

$$SelfAttn(X) = softmax\left(\frac{(XW_Q)(XW_K)^\top}{\sqrt{d_k}}\right)(XW_V) \tag{12}$$

$$X_t^{emo} = SelfAttn^{(2)}(X_t) \in R^{k\times d} \tag{13}$$

$$X_p^{emo} = SelfAttn^{(2)}(X_p) \in R^{n\times d} \tag{14}$$

$$SelfAttn\ (X) = softmax\left(\frac{XW_Q(XW_K)^\top}{\sqrt{d_k}}\right)(XW_V) \tag{15}$$

where $SelfAttn^{(2)}$ represents the stacking of two layers of attention.

This design enables the model to focus on intra-modality sentiment-relevant elements. For textual inputs, emotional tokens such as adjectives, negations, and intensifiers are emphasized. For visual inputs, local visual patterns such as facial expressions or warm/cold color regions are highlighted.

The outputs $X_t^{emo}$ and $X_p^{emo}$ serve as modality-enhanced emotional descriptors, which are fed into the subsequent fusion module.

### *3.4. Hierarchical Gated Fusion Module (HGFM)*

To integrate the extracted intra-modal emotion features with cross-modal interaction features, a Hierarchical Gated Fusion Module (HGFM) is proposed. This module is designed to suppress irrelevant or redundant signals and to model complementary dependencies between local and global sentiment cues. It comprises two sequential stages: Cross-Gated Fusion (CGF) and Multi-Scale Attention Gating (MAG).

#### 3.4.1. Cross-Gated Fusion (CGF)

Given the intra-modal emotion features $X_t^{emo}$ and $X_p^{emo}$, the CGF mechanism introduces mutual gating to regulate cross-modality influence:

$$X_g = \sigma(W_t X_t^{emo}) \odot X_p^{emo} + \sigma\left(W_p X_p^{emo}\right) \odot X_t^{emo} \tag{16}$$

where $W_t \in R^{d\times d}, W_p \in R^{d\times d}$ are learnable linear projections, $\sigma(\cdot)$ denotes the sigmoid function, and $\odot$ is element-wise multiplication. $X_g \in R^{n\times d}$.

This formulation enables each modality to attend to the most relevant features in the other modality, dynamically suppressing modality-specific noise. The structure of the module is shown in Figure 3.

#### 3.4.2. Multi-Scale Attention Gating (MAG)

To further integrate the cross-gated intra-modal features $X_g$ with the cross-modal interaction representation $X_C$ a Multi-Scale Attention Gating (MAG) module is designed. The MAG structure of the module is shown in Figure 4. It aims to model both temporal dependencies and channel-wise emotional salience through convolutional and attention mechanisms, followed by adaptive fusion.

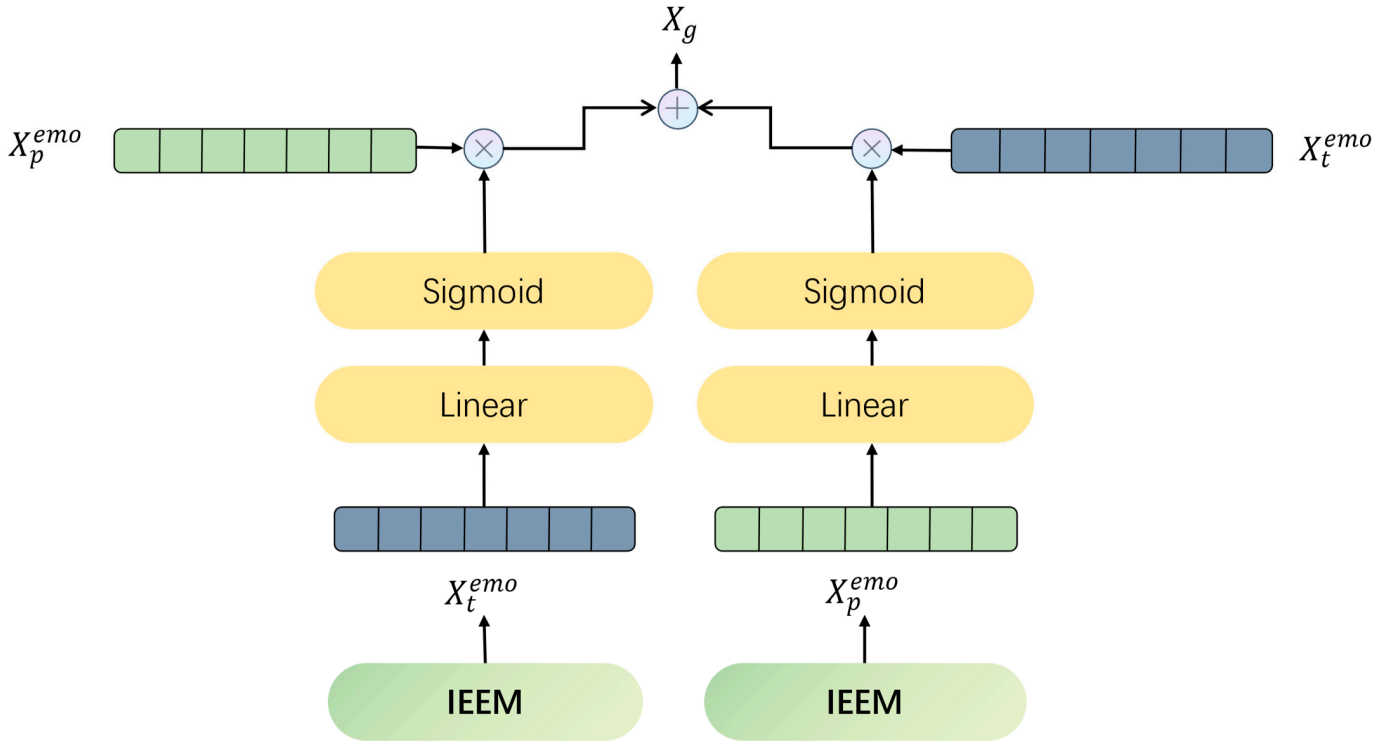


**Figure 3.** Structure diagram of CGF.

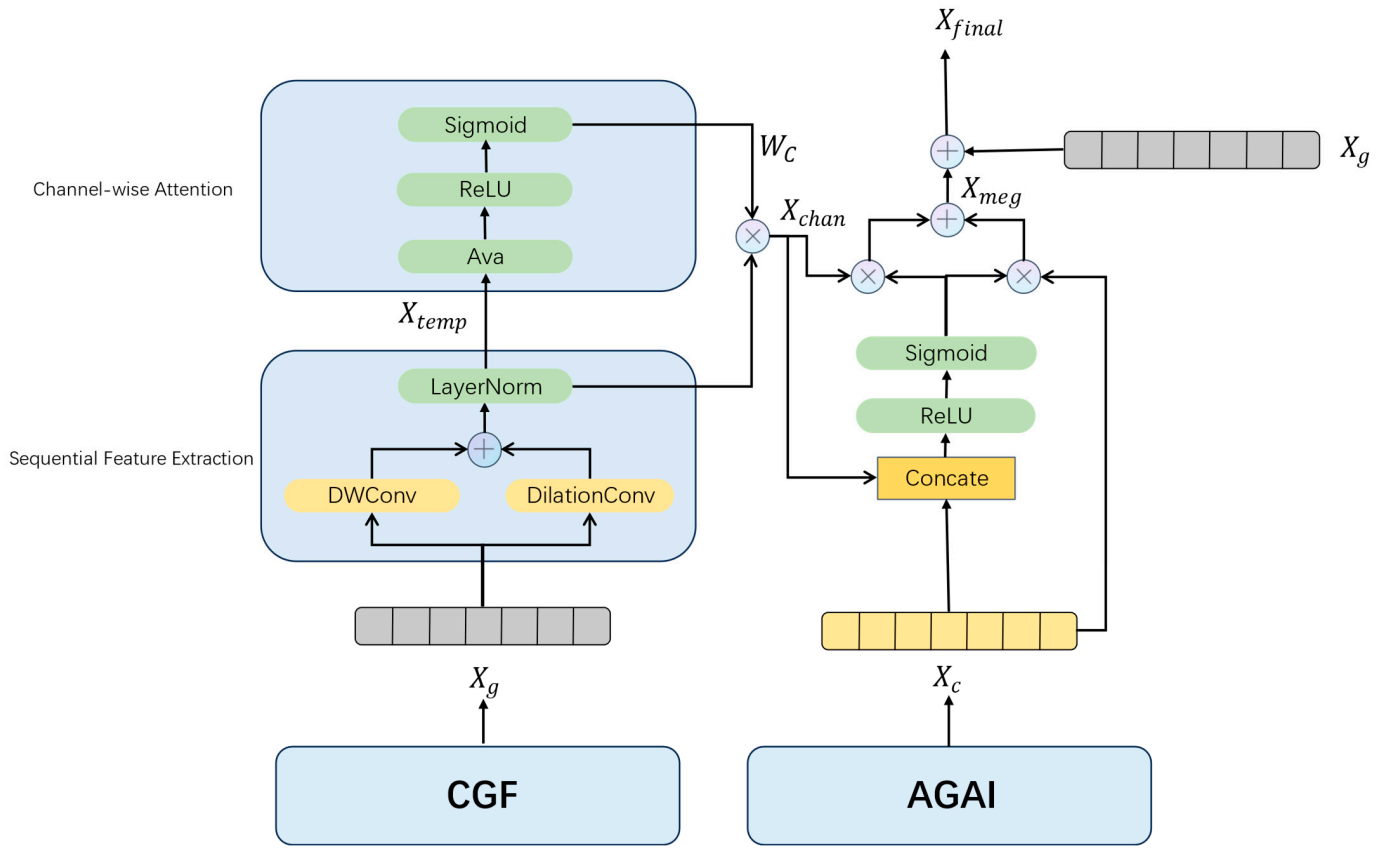


**Figure 4.** Structure diagram of MAG.

Sequential Feature Extraction: To enhance sequence-level emotion representations, a specialized feature extraction block was introduced to model ordered data sequences. This module employs multi-scale convolutional operations to capture both local associations and

broader contextual relationships within the input feature sequence. The architecture comprises two complementary components: a depthwise separable convolution, which extracts fine-grained patterns and immediate sequential dependencies, and a dilated convolution, which expands the receptive field to integrate hierarchical relationships over longer temporal spans.

The depthwise separable convolution performs channel-wise filtering using a kernel size of 3 (with no dilation), effectively capturing local sequential patterns within each feature channel independently. In parallel, the dilated convolution is applied with a dilation rate of 2, enabling the network to capture extended contextual dependencies by enlarging the receptive field without increasing the parameter count. The outputs from both convolutional paths are aggregated via element-wise addition and passed through a Layer Normalization layer to stabilize training.

This synergistic design facilitates comprehensive modeling of both short-range feature interactions and long-range dependencies, which is essential for capturing subtle emotional cues embedded in sequential data structures.

$$X_{temp} = LayerNorm\left(DWConv\left(X_g^\top\right) + DilationConv\left(X_g^\top\right)\right)^\top \tag{17}$$

where $X_g^\top$ is the transposed input. The convolution kernel size of DWConv is $3 \times 1$, without dilation. The convolution kernel size of DilationConv is $3 \times 1$, dilation rate = 2. $X_{temp} \in R^{n \times d}$.

Channel-wise Attention: To identify which semantic channels (i.e., feature dimensions) are more emotion-relevant, a squeeze-and-excitation-style channel attention is applied. The channel weights are computed as

$$w_c = \sigma\left(W_{channel} \cdot \mathrm{ReLU}\left(W_{temp} \cdot \mathrm{Avg}\left(X_{temp}\right)\right)\right) \tag{18}$$

where $Avg\left(X_{temp}\right) = \frac{1}{L}\sum_i^L X_{temp}^{(i)}$,$W_{channel} \in R^d$ represents the channel attention weight.

The refined feature after channel attention becomes

$$X_{chan} = X_{temp} \odot w_c \tag{19}$$

This step ensures that the most sentiment-relevant dimensions are amplified before fusion.

Adaptive Gated Fusion: Finally, the attended representation $X_{chan}$ and the interaction feature $X_C$ are integrated via a learnable adaptive gating mechanism. A sigmoid gate is generated from the concatenation of both inputs:

$$g = \sigma(W_k \cdot ReLU(W_{total} \cdot [X_{chan}; X_c])) \tag{20}$$

where $g \in \mathrm{R}^1$ is used to weight the relative contribution of each path:

$$X_{mag} = g \cdot X_{chan} + (1 - g) \cdot X_c \tag{21}$$

To enhance residual consistency, the final multi-modal fusion representation is computed as

$$X_{final} = X_{mag} + X_g \tag{22}$$

This fusion strategy enables the model to capture hierarchical emotional signals, spanning local temporal structure, global channel importance, and cross-modal alignment, in a flexible and learnable way.

### *3.5. Classification and Final Objective*

The final multimodal sentiment representation $X_{\text{final}}$, obtained from the hierarchical fusion module, is passed through a linear classifier to generate prediction logits:

$$\hat{y} = Softmax\left(W_{task}X_{final} + b\right) \tag{23}$$

where $W_{task} \in R^{d\times C}$, C represents the number of categories. $b \in R^{C}$ are learnable classification parameters.

We compute the standard cross-entropy loss $\mathcal{L}_{task}$ between the predicted logits and the ground truth labels. Together with the contrastive loss $\mathcal{L}_{con}$ from the AGAI, the final objective function of the model is defined as

$$\mathcal{L}_{task} = -\frac{1}{B}\sum_{i=1}^{B}\log(y_i[L_i]) \tag{24}$$

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda\mathcal{L}_{con} \tag{25}$$

where $\lambda$ is a weighting hyperparameter that balances the contribution of the contrastive alignment loss.

## 4. Results

### *4.1. Dataset*

An evaluation of our models was conducted using three publicly accessible multimodal sentiment analysis datasets sourced from social media platforms: MVSA-S [28], TumEmo [26] and HFM [29]. The HFM dataset is sourced from Twitter, while TumEmo is derived from Tumblr and contains numerous image-text pairs. TumEmo is a weakly supervised sentiment analysis dataset. Notably, both datasets capture users' sentiment expressions across a broad spectrum of topics, which closely aligns with real-world scenarios in social media sentiment analysis. The MVSA dataset is commonly used for multimodal sentiment analysis tasks. Each dataset was divided into training, validation, and testing sets using an 8:1:1 ratio, with details provided in Tables 3 and 4.

**Table 3.** Dataset statistics for the MVSA-S, HFM, and TumEmo.

| Dataset | Train | Val | Test | Total |
|---|---|---|---|---|
| HFM | 19,816 | 2410 | 2409 | 24,635 |
| TumEmo | 156,204 | 19,525 | 19,536 | 195,265 |
| MVSA-S | 3611 | 450 | 450 | 4511 |

### *4.2. Parameter Setting and Evaluation Indicators*

The Adam optimizer implemented in PyTorch (2.8.0 versions) was employed to train the models. The batch size was set to 64 for both the HFM and TumEmo datasets, with an initial learning rate of $1 \times 10^{-4}$. Learning rate scheduling configurations varied by dataset: for HFM, step size = 2 and gamma = 0.8 were used, while identical settings (step size = 2 and gamma = 0.8) were applied to TumEmo. Contrastive loss coefficients were set to 0.0001 for HFM and 0.001 for TumEmo. All experiments were conducted on NVIDIA RTX 3090 GPUs (Nvidia Corporation, Santa Clara, CA, USA). The detailed experimental parameters are shown in Table 5.

**Table 4.** Emotional Label Categories.

| | Positive | | Negative | Total |
|---|---|---|---|---|
| **HFM** | 10,560 | | 14,075 | 24,635 |
| | **Positive** | **Neutral** | **Negative** | **Total** |
| **MVSA-S** | 2683 | 470 | 1358 | 4511 |
| | **TumEmo** | | | |
| | **Angry** | | 14,554 | |
| | **Bored** | | 32,283 | |
| | **Calm** | | 18,109 | |
| | **Fear** | | 20,264 | |
| | **Happy** | | 50,267 | |
| | **Love** | | 34,511 | |
| | **Sad** | | 25,277 | |
| | **All** | | 195,265 | |

**Table 5.** Experimental parameters.

| Dataset | Epoch | Scheduler Type | Step Size | Dropout | Gamma | Batch_Size | lr | λ | τ |
|---|---|---|---|---|---|---|---|---|---|
| TumEmo | 20 | StepLR | 2 | 0.1 | 0.8 | 64 | $1-10^{-4}$ | 0.001 | 0.07 |
| MVSA-S | 40 | StepLR | 2 | 0.1 | 0.8 | 64 | $1-10^{-4}$ | 0.0001 | 0.4 |
| HFM | 40 | StepLR | 2 | 0.1 | 0.8 | 32 | $1-10^{-4}$ | 0.0005 | 0.07 |

Accuracy (ACC) and F1 score are standard evaluation metrics for sentiment analysis, quantifying model performance. ACC measures the proportion of correctly classified instances relative to the total dataset size. The F1 score provides a balanced assessment by harmonizing precision and recall, particularly valuable for imbalanced classification scenarios. For both metrics, higher values indicate superior model effectiveness.

$$\mathrm{ACC} = \frac{\mathrm{TP}}{\mathrm{N}}$$

$$\mathrm{F1-score} = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

Let P denote the number of correctly predicted samples and N the total number of samples. The accuracy is defined as the ratio of correctly predicted samples to the total number of samples, the proportion of true positives among all predictions. The recall measures the proportion of correctly predicted positive instances among all actual positive instances. The F1-score is calculated as the harmonic mean of accuracy and recall, providing a balanced measure of both metrics.

### *4.3. Baseline Model*

In this study, baseline models are categorized into three groups: text-based models, image-based models, and multimodal models.

Text-based models: CNN [30] and BiLSTM [31] are classical neural architectures widely used in text classification tasks. TGNN [32] introduces graph structures at the text level to capture non-sequential semantic relationships. BiACNN [33] integrates CNN and BiLSTM with an attention mechanism to enhance sentiment representation in textual data.

Image-based models: ResNet [34] serves as a foundational visual backbone, applied through standard pre-training and fine-tuning procedures. OSDA [26] adopts a multi-view strategy to extract affective cues from images for sentiment classification.

Multimodal models: MultiSentiNet [10] is a deep semantic network designed for joint text-image sentiment analysis. HSAN [35] employs a hierarchical attention framework that leverages image captions to align multimodal features. MGNNS [36] introduces a multi-channel graph neural network with sentiment-aware representations for multimodal sentiment detection. CLMLF [37] utilizes a multi-layer Transformer architecture for feature fusion and incorporates contrastive learning to improve text-image alignment. CIGNN [38] enhances image representation through attribute-level encoding, constructs dual graph neural networks to capture dataset-level global features, and performs sentiment analysis on the fused representations. MAMSA [1] introduces an adaptive attention mechanism to dynamically balance the contribution of text and image features, and employs hierarchical fusion guided by sentiment information to enhance multimodal representation learning. MPNAS [39] presents a unified NAS-based pruning framework with a two-stage process, first coarse-grained NAS, then fine-grained NAS guided by MSA characteristics, to tackle existing pruning limitations, showing superiority on three datasets.

### *4.4. Comparative Experiments*

To evaluate the effectiveness of the proposed model in multimodal sentiment recognition, Experiments are conducted on two publicly available datasets, and its performance is compared against both unimodal and multimodal baseline models. The detailed results are presented in Table 6. All the baseline model results are derived from MPNAS [39] and MAMSA [1]. The experimental results of this model were repeated three times under different seeds, and were presented in the form of "average value ± standard deviation". The CLIP model has not been fine-tuned. Bold values indicate the performance of the proposed MAHFNet model, while underlined values denote the best results among all baseline methods. For the TumEmo dataset, the model's runtime is 4 h, 50 min, and 49.70 s, with approximately 14.15 million trainable parameters.

**Table 6.** Experimental results of acc and f1 on three datasets.

| Modality | Model | TumEmo | | Model | HFM | | Model | MVSA-S | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc | F1 | | Acc | F1 | | Acc | F1 |
| Text | CNN | 0.6154 | 0.4774 | CNN | 0.8003 | 0.7532 | CNN | 0.6819 | 0.5590 |
| | BiLSTM | 0.6188 | 0.5126 | BiLSTM | 0.8190 | 0.7753 | BiLSTM | 0.7012 | 0.6506 |
| | BERT | - | - | BERT | 0.8389 | 0.8326 | BERT | - | - |
| | TGNN | 0.6379 | 0.6362 | - | - | - | TGNN | 0.7034 | 0.6594 |
| Image | ResNet50 | - | - | ResNet50 | 0.7277 | 0.7138 | ResNet50 | 0.6467 | 0.6155 |
| | DuIG | 0.4636 | 0.4561 | ResNet101 | 0.7228 | 0.7122 | DuIG | 0.6822 | 0.6538 |
| Image-Text | MSN | 0.6418 | 0.5692 | Concat(2) | 0.8103 | 0.7799 | MSN | 0.6984 | 0.6984 |
| | HSAN | 0.6309 | 0.5398 | Concat(3) | 0.8174 | 0.7874 | HSAN | 0.6988 | 0.6690 |
| | CoMem | 0.6426 | 0.5909 | MMSD | 0.8344 | 0.8018 | CoMem | 0.7051 | 0.7001 |
| | MGNNS | 0.6672 | 0.6669 | D&R Net | 0.8402 | 0.8060 | MGNNS | 0.7377 | 0.7270 |
| | CLMLF | - | - | CLMLF | 0.8543 | 0.8487 | CLMLF | 0.7533 | 0.7346 |
| | CIGNN | 0.6738 | 0.6706 | CIGNN | 0.8556 | 0.8492 | CIGNN | 0.7511 | 0.7333 |
| | MAMSA | 0.6745 | 0.6723 | MPMM | 0.8587 | 0.8557 | MPMM | 0.7624 | 0.7418 |
| | MAHFNet | **0.7000 ± 0.0021** | **0.6986 ± 0.0027** | MAHFNet | **0.8643 ± 0.0010** | **0.8647 ± 0.0011** | MAHFNet | **0.7627 ± 0.0022** | **0.7536 ± 0.0030** |

**Note:** The best results for each metric are highlighted in bold.

#### 4.4.1. Quantitative Analysis

As shown in Table 6, compared with TGNN, the proposed model achieves an improvement of 6.21% in both metrics on the TumEmo dataset. In comparison with DuIG, the model demonstrates a significant performance gain of 23.64% and 24.25% on the same dataset. Furthermore, relative to MAMSA, the proposed model improves by 2.55% and 2.63% on TumEmo. Additionally, compared with CIGNN, the model yields improvements of 0.87% and 1.35% on the HFM dataset. These results indicate that the proposed model

consistently achieves superior performance in both binary and fine-grained (seven-class) sentiment classification tasks. On the MVSA-S dataset. These gains reach 0.03% in ACC and 1.26% in F1 score.

#### 4.4.2. Qualitative Analysis

Compared with unimodal models, the proposed model is more effective in capturing emotional information from multiple modalities by enabling enhanced and complementary interactions across them. As shown in Table 2, relying solely on images for sentiment prediction results in low accuracy. This is primarily because many images lack explicit emotional cues, and the presence of abundant non-emotional content can interfere with the model's predictions, leading to sparse emotional signals. In contrast, our model achieves effective multimodal interaction and preserves fine-grained local features. Moreover, through the hierarchical gated fusion module, it captures the hierarchical dependencies between local and global emotional representations, thereby improving the model's ability to perceive and integrate sentiment information across different semantic levels.

### *4.5. Ablation Experiment*

To further investigate the contribution of each component to the overall model performance, Four groups of ablation experiments are conducted on the TumEmo, MVSA-S and HFM datasets. The corresponding results are presented in Table 7.

**Table 7.** Ablation experiments.

| **Model** | **TumEmo** | | **Model** | **HFM** | | **Model** | **MVSA-S** | |
|---|---|---|---|---|---|---|---|---|
| | **Acc** | **F1** | | **Acc** | **F1** | | **Acc** | **F1** |
| MAHFNet | **0.7000** | **0.6986** | MAHFNet | **0.8643** | **0.8647** | MAHFNet | **0.7621** | **0.7536** |
| -(MAG) | 0.6831 | 0.6818 | -(MAG) | 0.8597 | 0.8595 | -(MAG) | 0.7361 | 0.7325 |
| -(HGFM) | 0.6954 | 0.6942 | -(HGFM) | 0.8600 | 0.8640 | -(HGFM) | 0.7561 | 0.7440 |
| -(IEEM) | 0.6902 | 0.6886 | -(IEEM) | 0.8580 | 0.8587 | -(IEEM) | 0.7450 | 0.7325 |
| -(AGAI) | 0.6290 | 0.6278 | -(AGAI) | 0.8240 | 0.8248 | -(AGAI) | 0.7228 | 0.6864 |

**Note:** The best results for each metric are highlighted in bold.

To evaluate the contribution of each component, ablation experiments were conducted by successively removing the proposed modules: MAG, HGFM, IEEM, and AGAI. Specifically, -(MAG) refers to the removal of the MAG. The MAG integrates the outputs of the CGF and AGAI for fusion. After removing the MAG, the output of the CGF is directly used as the final output. -(HGFM) refers to the removal of the Hierarchical Gated Fusion Module, where feature representations are directly averaged for prediction. -(IEEM) denotes the additional removal of the Intra-modal Emotion Extraction Module on top of HGFM. -(AGAI) represents the further elimination of the Attention-Guided Alignment and Interaction Module, replacing it with a simple cross-attention mechanism for multimodal feature fusion.

The results clearly demonstrate that each proposed module contributes significantly to the overall performance, further validating the effectiveness and necessity of the modular design. When only one MAG is removed, only the output of the CGF is used as the final output. The CGF directly fuses the textual and graphic features through the gating mechanism without the interaction of the attention mechanism or multi-scale fusion. This results in a large amount of noise in the output features, leading to a significant decline in the effect. Upon removal of the HGFM, the model's performance decreases on both datasets, indicating that the hierarchical gated fusion mechanism effectively captures dependencies between local and global sentiment representations, thereby facilitating comprehensive

multimodal fusion. When the IEEM is further removed, the model's accuracy declines by 0.52% and 0.56% on TumEmo, and by 0.20% and 0.53% on HFM. This highlights the importance of extracting local features from individual modalities, which enables the effective complementation of emotional information and mitigates the loss of fine-grained details during modality fusion. Finally, with the removal of the AGAI, a substantial performance drop is observed, indicating that semantic interaction between text and image is effectively achieved through the cross-modal attention mechanism. Additionally, the integration of contrastive learning further enhances the model's ability to aggregate semantically consistent image-text pairs.

### *4.6. Parametric Experiments*

The correct parameter setting can effectively improve the performance of the model. In this chapter, the important parameters include the learning rate, Batch_size, and the coefficient λ of Contrastive learning.

#### 4.6.1. Learning Rate Experiments

Results of the learning rate experiments are presented in Table 8, where the optimal learning rate for three datasets is identified as 0.0001. Specifically, when the learning rate is excessively large (e.g., 0.001), the model's performance decreases sharply during training. Conversely, as the learning rate continues to decrease from this high value, the model's performance gradually improves; it reaches its peak when the learning rate is 0.0001, after which performance begins to decline if the learning rate is reduced further.

**Table 8.** Experimental results of learning rate on dataset.

| Learing Rate | TumEmo | | Learing Rate | HFM | | Learing Rate | MVSA-S | |
|---|---|---|---|---|---|---|---|---|
| | Acc | F1 | | Acc | F1 | | Acc | F1 |
| 0.0005 | 0.6855 | 0.6864 | 0.0005 | 0.8601 | 0.8608 | 0.0005 | 0.7251 | 0.7020 |
| 0.0004 | 0.6915 | 0.6925 | 0.0004 | 0.8647 | 0.8652 | 0.0004 | 0.7450 | 0.7387 |
| 0.0002 | 0.6996 | 0.6984 | 0.0002 | 0.8589 | 0.8589 | 0.0002 | 0.7384 | 0.7315 |
| 0.0001 | **0.7000** | **0.6986** | 0.0001 | **0.8643** | **0.8647** | 0.0001 | **0.7627** | **0.7544** |
| 0.00008 | 0.6947 | 0.6937 | 0.00008 | 0.8643 | 0.8641 | 0.00008 | 0.7251 | 0.7087 |
| 0.00006 | 0.6833 | 0.6884 | 0.00006 | 0.8597 | 0.8600 | 0.00006 | 0.7472 | 0.7410 |
| 0.00005 | 0.6878 | 0.6870 | 0.00005 | 0.8522 | 0.8527 | 0.00005 | 0.7583 | 0.7544 |
| 0.00001 | 0.6377 | 0.6371 | 0.00001 | 0.8277 | 0.8286 | 0.00001 | 0.6962 | 0.6540 |

**Note:** The best results for each metric are highlighted in bold.

#### 4.6.2. Batchsize

The experiments for Batchsize are shown in Table 9. When Batch_size = 64, the best results are obtained on TumEmo and HFM datasets. When Batch_size = 32,the best results are obtained on MVSA-S datasets.

**Table 9.** Experimental results of Batch_size.

| Batch_Size | TumEmo | | Batch_Size | HFM | | Batch_Size | MVSA-S | |
|---|---|---|---|---|---|---|---|---|
| | Acc | F1 | | Acc | F1 | | Acc | F1 |
| 8 | 0.6926 | 0.6926 | 8 | 0.8580 | 0.8575 | 64 | 0.7273 | 0.7265 |
| 16 | 0.6930 | 0.6930 | 16 | 0.8551 | 0.8558 | 32 | **0.7627** | **0.7544** |
| 32 | 0.6896 | 0.6898 | 32 | 0.8634 | 0.8629 | 16 | 0.7384 | 0.7218 |
| 64 | **0.7000** | **0.6986** | 64 | **0.8643** | **0.8647** | 8 | 0.7206 | 0.7250 |

**Note:** The best results for each metric are highlighted in bold.

#### 4.6.3. Contrastive Learning Loss Experiments

The effect of the contrastive learning loss coefficient is presented in Table 10. Contrastive learning plays a critical role in aligning textual and visual modality features by encouraging semantically consistent cross-modal representations to be drawn closer in the embedding space. Appropriately setting the weight coefficient of the contrastive loss is essential for optimizing the learning process, enhancing the model's discriminative capability, and fully leveraging the potential of the cross-modal alignment module. As shown in the results, a coefficient of 0.001 yields the best performance on the TumEmo dataset, while a coefficient of 0.0001 achieves the best results on the HFM dataset and a coefficient of 0.0005 achieves the best results on the MVSA-S dataset.

**Table 10.** Experimental results of $\lambda$.

| λ | TumEmo | | λ | HFM | | λ | MVSA-S | |
|---|---|---|---|---|---|---|---|---|
| | **Acc** | **F1** | | **Acc** | **F1** | | **Acc** | **F1** |
| 0.01 | 0.6947 | 0.6947 | 0.01 | 0.8605 | 0.8602 | 0.01 | 0.7339 | 0.7282 |
| 0.005 | 0.6944 | 0.6949 | 0.005 | 0.8618 | 0.8621 | 0.005 | 0.7494 | 0.7422 |
| 0.001 | **0.7000** | **0.6986** | 0.001 | 0.8634 | 0.8636 | 0.001 | 0.7361 | 0.7331 |
| 0.0005 | 0.6950 | 0.6939 | 0.0005 | 0.8501 | 0.8506 | 0.0005 | **0.7627** | **0.7544** |
| 0.0001 | 0.6959 | 0.6941 | 0.0001 | **0.8643** | **0.8647** | 0.0001 | 0.7140 | 0.6991 |

**Note:** The best results for each metric are highlighted in bold.

#### 4.6.4. Contrastive Learning of the Temperature Sweep Experiment

The effect of the temperature sweep is presented in Table 11. The effect of the temperature parameter τ in contrastive learning on model performance across the TumEmo, HFM, and MVSA-S datasets. As a critical factor in scaling similarity scores within contrastive loss, τ regulates how closely semantically consistent textual and visual features cluster in the embedding space, balancing the model's ability to distinguish between instances and tolerate semantic similarities among samples. Proper calibration of τ is essential for optimizing cross-modal alignment, too small a value may over-penalize useful semantic proximity, while too large a value can blur meaningful distinctions. From the results, a τ of 0.07 yields the best performance on TumEmo and HFM, effectively enhancing feature clustering for these datasets. In contrast, the MVSA-S dataset achieves optimal results with a higher τ of 0.4, suggesting that its more complex cross-modal semantic correlations benefit from a relaxed temperature to better capture nuanced alignments.

**Table 11.** Experimental results of $\tau$.

| τ | TumEmo | | τ | HFM | | τ | MVSA-S | |
|---|---|---|---|---|---|---|---|---|
| | **Acc** | **F1** | | **Acc** | **F1** | | **Acc** | **F1** |
| 0.07 | **0.7000** | **0.6986** | 0.07 | **0.8643** | **0.8647** | 0.07 | 0.7450 | 0.7385 |
| 0.1 | 0.6943 | 0.6935 | 0.1 | 0.8622 | 0.8621 | 0.1 | 0.7472 | 0.7437 |
| 0.2 | 0.6934 | 0.6937 | 0.2 | 0.8609 | 0.8615 | 0.2 | 0.7539 | 0.7468 |
| 0.4 | 0.6934 | 0.6934 | 0.4 | 0.8623 | 0.8639 | 0.4 | **0.7627** | **0.7544** |
| 0.8 | 0.6921 | 0.6917 | 0.8 | 0.8638 | 0.8644 | 0.8 | 0.7494 | 0.7367 |

**Note:** The best results for each metric are highlighted in bold.

## 5. Confusion Matrix

To better illustrate the alignment between predicted labels and ground truth, the confusion matrices obtained from the model's performance on both datasets are presented in Figure 5. It can be observed that the model performs robustly across most categories in the seven-class TumEmo dataset, with particularly high accuracy in identifying Anger,

Bored, Reflux, Fear, and Sad. Nevertheless, a relatively higher rate of misclassification is observed between the Happy and Love categories. This is likely attributable to their inherent semantic and emotional proximity, as expressions of love often encompass or co-occur with feelings of happiness, leading to ambiguous boundaries between the two. Such subtle emotional overlaps pose additional challenges for precise classification. In contrast, the binary classification task on the HFM dataset is less demanding, and the model achieves consistently high predictive accuracy across both classes.

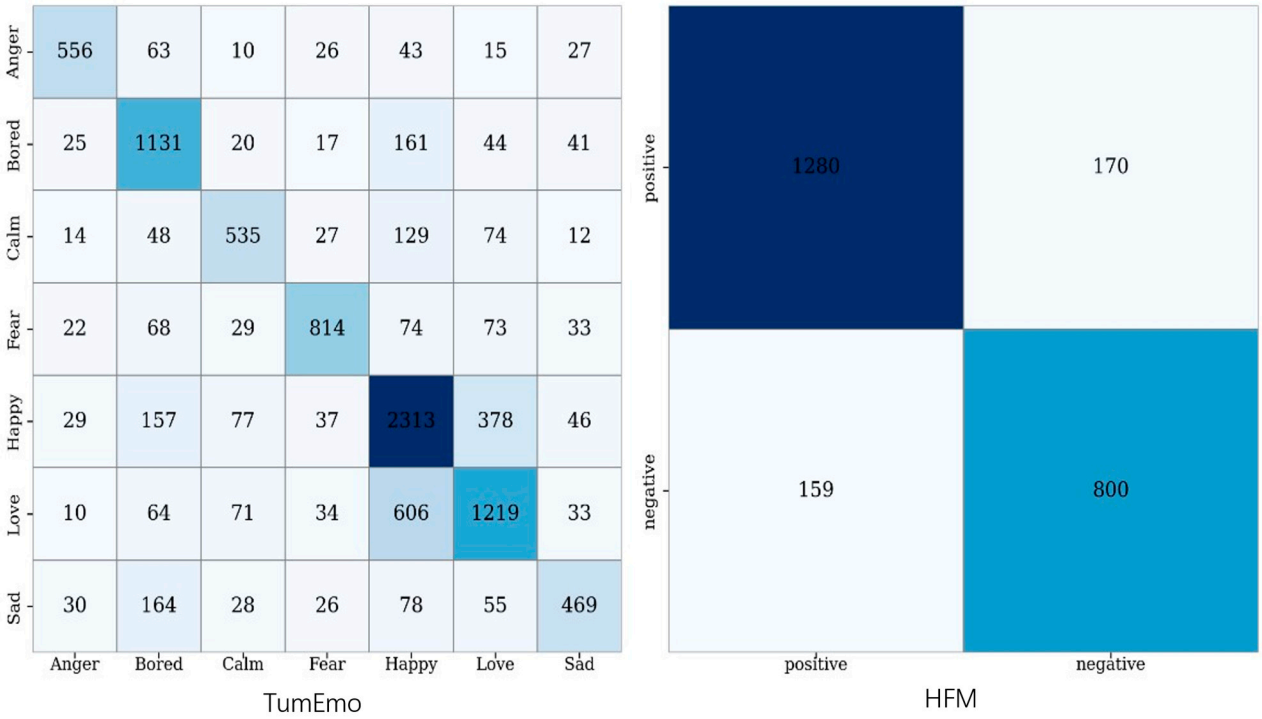


**Figure 5.** Confusion matrix results, the (**left**) figure shows the results for the Tum-Emo dataset, and the (**right**) figure shows the results for the HFM dataset.

## 6. Case Analysis

### *6.1. Case Analysis Based on Pictures and Texts*

This chapter further analyzes the effectiveness of the model in sentiment analysis by examining case studies. As shown in Figure 6 and Table 12, in Figure 6a, the image presents a town scene on a snowy day, with an overall cool tone. Visually, there are no direct "exciting and positive" cues. The text reads: "Let it snow, let it snow, let it snow?" It is snowing after 3 years? #Overjoyed" conveys a positive emotion. The prediction results of MAHFNet are consistent with the true values as "Positive". However, when the HGFM is removed, the model is unable to capture the multi-scale features in the image, and cannot dynamically achieve the association and fusion of the positive semantic of the text and the features of the image, resulting in a prediction of neutral sentiment. In Figure 6b,c, the deficiencies of the model in judging the emotions of "Happy" and "Love" are illustrated. The boundaries between the emotions of "Happy" and "Love" are blurred. For instance, in Figure 6c, although the picture mainly conveys a sense of happiness, there are words like "love" in the text, which more often express a generalized fondness. Such delicate emotions are rather difficult to capture. In Figure 6b, the overall theme revolves around "engagement anniversary", and the text also shows a clear sense of happiness. In response to these blurred boundaries, the model still has shortcomings and may make incorrect judgments.


**Figure 6.** The graphic and textual pair used for analyzing cases.

**Table 12.** Case studies in different scenarios.

| **Model** | | | |
|---|---|---|---|
| | **(a)** | **(b)** | **(c)** |
| MAHFNet | Positive | Love | Happy |
| -(HGFM) | Neutral | Love | Happy |
| Ture Label | Positive | Happy | Love |

### *6.2. Case Heat Map Analysis*

The heat map is shown in Figure 7. These heatmaps show how cross-attention works in the AGAI for multimodal sentiment analysis, helping the model focus on text-image regions with emotional connections. For the first image (inside a vehicle), the heatmap highlights crowded areas, faces, and lights; these match the text's complaint about "overcrowded transport" and negative tone (like "disgrace"). The model uses cross-attention to align this "unpleasant visual scene" with the text's negative emotion. For the second image (superhero figure), the heatmap focuses on the smile, power symbol, and thumbs-up, visual cues of positivity. The model leverages cross-attention to link these "positive visual features" with the text about "filming memories," capturing upbeat sentiment. Overall, the highlighted regions match what humans see as emotion-carrying visual parts, so cross-attention effectively guides the model to focus on areas that matter for cross-modal sentiment understanding.


**Figure 7.** Heatmap of the cross-attention mechanism.

## 7. Conclusions

The proliferation of multimodal posts across social platforms has opened new avenues for sentiment analysis, yet existing studies often suffer from detail loss during cross-modal interaction fusion, struggle with semantic alignment challenges, and remain vulnerable to modal noise. To address these limitations and boost analytical precision, this study proposes the Multimodal Alignment and Hierarchical Fusion Network (MAHFNet) for sentiment analysis tasks. While MAHFNet integrates established mechanisms, including attention, gating, and contrastive learning, its novelty lies in how these components are synergistically combined to tackle cross-modal challenges in sentiment analysis. Specifically, MAHFNet first leverages contrastive learning and attention to achieve interaction and alignment of multimodal features; then extracts local features from individual modalities to compensate for information loss during cross-modal integration; and finally employs a hierarchical gated fusion strategy to combine global and local features, enhancing emotional representations for more accurate sentiment prediction. Beyond the technical contributions, ethical considerations in multimodal sentiment analysis also merit attention. Multimodal datasets often carry inherent biases and models like MAHFNet, if deployed without safeguards, may amplify these biases in scenarios like automated content moderation or social listening. Addressing them is nonetheless essential to ensuring the model's positive and responsible real-world impact.

Experimental results on multiple public datasets demonstrate the effectiveness and robustness of MAHFNet. Despite these encouraging results, the model still faces challenges in recognizing emotions from complex images or ambiguous expressions. Future work will explore incorporating descriptive text to mitigate emotion sparsity in visually intricate scenarios, and integrate bias-mitigation strategies to enhance the model's fairness and ethical compliance.

**Author Contributions:** Conceptualization, J.H. and H.L.; methodology, J.H. and H.L.; Software: J.H.; Formal analysis and investigation: J.H. and H.L.; Writing—original draft preparation: J.H., X.M. and H.L.; Writing—review and editing: X.M. and H.L.; Funding acquisition: X.M. and H.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Hainan Provincial Natural Science Foundation of China (Grant numbers: 623RC455, 623RC457, 425QN244), the Scientific Research Fund of Hainan University (Grant numbers: KYQD(ZR)-22096, KYQD(ZR)-22097), and the Lanzhou University-Hainan University Technical Service Project (HD-KYH-2024424).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The TumEmo dataset is available at https://github.com/YangXiaocui1215/MVAN (accessed on 7 May 2025). The HFM dataset is available at https://github.com/headacheboy/data-of-multimodal-sarcasm-detection (accessed on 15 March 2025). The implementation of this work is publicly accessible at https://github.com/waibibab/Emotion(accessed on 16 September 2025).

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

## References

1. Wang, H.; Du, Q.; Xiang, Y. Image–text sentiment analysis based on hierarchical interaction fusion and contrast learning enhanced. *Eng. Appl. Artif. Intell.* **2025**, *146*, 110262. [CrossRef]
2. Wang, C.; Zhou, B.; Tu, H.; Liu, Y. Rumor detection on social media using temporal dynamic structure and emotional information. In Proceedings of the 2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC), Shenzhen, China, 9–11 October 2021; IEEE: Piscataway, NJ, USA, 2021.
3. Liu, X.; Wei, F.; Jiang, W.; Zheng, Q.; Qiao, Y.; Liu, J.; Niu, L.; Chen, Z.; Dong, H. MTR-SAM: Visual multimodal text recognition and sentiment analysis in public opinion analysis on the internet. *Appl. Sci.* **2023**, *13*, 7307. [CrossRef]
4. Malitesta, D.; Cornacchia, G.; Pomo, C.; Merra, F.A.; Di Noia, T.; Di Sciascio, E. Formalizing multimedia recommendation through multimodal deep learning. *ACM Trans. Recomm. Syst.* **2025**, *3*, 1–33. [CrossRef]
5. Wang, H.; Ren, C.; Yu, Z. Multimodal sentiment analysis based on multiple attention. *Eng. Appl. Artif. Intell.* **2025**, *140*, 109731. [CrossRef]
6. Li, C.; Hu, Z. Multimodal sentiment analysis of social media based on top-layer fusion. In Proceedings of the 2022 IEEE 8th International Conference on Computer and Communications (ICCC), Chengdu, China, 9–12 December 2022; IEEE: Piscataway, NJ, USA, 2022.
7. 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*, 3375–3385. [CrossRef]
8. Xiao, X.; Pu, Y.; Zhao, Z.; Gu, J.; Xu, D. BIT: Improving image-text sentiment analysis via learning bidirectional image-text interaction. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023; IEEE: Piscataway, NJ, USA, 2023.
9. Shen, J.; You, L.; Ma, Y.; Zhao, Z.; Liang, H.; Zhang, Y.; Hu, B. UA-DAAN: An Uncertainty-Aware Dynamic Adversarial Adaptation Network for EEG-Based Depression Recognition. *IEEE Trans. Affect. Comput.* **2025**, *16*, 2130–2141. [CrossRef]
10. Xu, N.; Mao, W.; Chen, G. Multi-interactive memory network for aspect based multimodal sentiment analysis. In Proceedings of the AAGAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019; Volume 33.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
12. Devlin, J.; Chang, M.W.; 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, Minneapolis, MN, USA, 2–7 June 2019; Volume 1.
13. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
14. Huang, C.; Zhang, J.; Wu, X.; Wang, Y.; Li, M.; Huang, X. TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis. *Knowl.-Based Syst.* **2023**, *269*, 110502. [CrossRef]
15. Wang, Q.; Xu, H.; Wen, Z.; Liang, B.; Yang, M.; Qin, B.; Xu, R. Image-to-text conversion and aspect-oriented filtration for multimodal aspect-based sentiment analysis. *IEEE Trans. Affect. Comput.* **2023**, *15*, 1264–1278. [CrossRef]
16. Yang, J.; Xiao, Y.; Du, X. Multi-grained fusion network with self-distillation for aspect-based multimodal sentiment analysis. *Knowl.-Based Syst.* **2024**, *293*, 111724. [CrossRef]
17. Kim, K.; Park, S. AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis. *Inf. Fusion* **2023**, *92*, 37–45. [CrossRef]

18. Yu, J.; Chen, K.; Xia, R. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Trans. Affect. Comput.* **2022**, *14*, 1966–1978. [CrossRef]
19. Chen, H.; Shen, F. Hierarchical cross-modal transformer for RGB-D salient object detection. *arXiv* **2023**, arXiv:2302.08052.
20. Le, H.-D.; Lee, G.S.; Kim, S.H.; Kim, S.; Yang, H.J. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access* **2023**, *11*, 14742–14751. [CrossRef]
21. Lu, G.; Li, J.; Wei, J. Aspect sentiment analysis with heterogeneous graph neural networks. *Inf. Process. Manag.* **2022**, *59*, 102953. [CrossRef]
22. Lu, Q.; Sun, X.; Gao, Z.; Long, Y.; Feng, J.; Zhang, H. Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. *Inf. Process. Manag.* **2024**, *61*, 103538. [CrossRef]
23. Zhou, R.; Guo, W.; Liu, X.; Yu, S.; Zhang, Y.; Yuan, X. Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. *arXiv* **2023**, arXiv:2306.01004.
24. Dong, S.; Fan, X.; Ma, X. Multichannel multimodal emotion analysis of cross-modal feedback interactions based on knowledge graph. *Neural Process. Lett.* **2024**, *56*, 190. [CrossRef]
25. Liu, X.; Xu, Z.; Huang, K. Multimodal emotion recognition based on cascaded multichannel and hierarchical fusion. *Comput. Intell. Neurosci.* **2023**, *2023*, 9645611. [CrossRef] [PubMed]
26. Yang, X.; Feng, S.; Wang, D.; Zhang, Y. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Trans. Multimed.* **2020**, *23*, 4014–4026. [CrossRef]
27. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PmLR, Virtual, 18–24 July 2021.
28. Niu, T.; Zhu, S.; Pang, L.; El Saddik, A. Sentiment analysis on multi-view social data. In Proceedings of the International Conference on Multimedia Modeling, Miami, FL, USA, 4–6 January 2016; Springer International Publishing: Cham, Switzerland, 2016.
29. Cai, Y.; Cai, H.; Wan, X. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2506–2515.
30. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:1408.5882. [CrossRef]
31. Wang, Z.; Yang, B. Attention-based bidirectional long short-term memory networks for relation classification using knowledge distillation from BERT. In Proceedings of the 2020 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Calgary, AB, Canada, 17–22 August 2020; IEEE: Piscataway, NJ, USA, 2020.
32. Huang, L.; Ma, D.; Li, S.; Zhang, X.; Wang, H. Text level graph neural network for text classification. *arXiv* **2019**, arXiv:1910.02356. [CrossRef]
33. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. In Proceedings of the AAGAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
34. 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, Las Vegas, NV, USA, 27–30 June 2016.
35. Xu, N. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 22–24 July 2017; IEEE: Piscataway, NJ, USA, 2017.
36. Yang, X.; Feng, S.; Zhang, Y.; Wang, D. Multimodal sentiment detection based on multi-channel graph neural networks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Volume 1.
37. Li, Z.; Xu, B.; Zhu, C.; Zhao, T. CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. *arXiv* **2022**, arXiv:2204.05515. [CrossRef]
38. Wang, H.; Ren, C.; Yu, Z. Multimodal sentiment analysis based on cross-instance graph neural networks. *Appl. Intell.* **2024**, *54*, 3403–3416. [CrossRef]
39. Zhang, B.; Ren, A.; Zhang, Z.; Duan, M.; Liu, D.; Tan, Y.; Zhong, K. MPNAS: Multimodal Sentiment Analysis Pruning via Neural Architecture Search. In Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–5.