INFORMATICA, 2026, Vol. 0, No. 0, 1–29

© 2026 *Vilnius University*
DOI: https://doi.org/10.15388/26-INFOR643

Frequency Domain Decoupling and Semantic Filtering Network for Multimodal Sentiment Analysis

Yundong LIU[1], Chengfang TAN[1], Shunxiang ZHANG[2], Yulei ZHANG[2], Kuan-Ching LI[3,*], Rubén González CRESPO[4,*]

[1] *School of Artificial Intelligence and Big Data, Suzhou University, Suzhou, China*
[2] *School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China*
[3] *School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, China*
[4] *Department of Computer Science and Technology, Universidad Internacional de La Rioja, Logroño, La Rioja, Spain*
e-mail: szxylyd@126.com, 874036730@qq.com, sxzhang@aust.edu.cn, 2230817302@qq.com, edge4xt@outlook.com, ruben.gonzalez@unir.net

Received: June 2025; accepted: August 2026

Abstract. Multimodal sentiment analysis, due to its comprehensive ability to capture user sentiment, has significant application value in areas such as public opinion analysis. Existing research, however, falls short in several aspects: (1) it inadequately models the global structural information of the image, and (2) it overlooks the potential noise impact within each modality. These limitations hinder the accurate extraction of sentiment cues from individual modalities. To address these issues, we propose a Frequency Domain Decoupling and Semantic Filtering Network for Multimodal Sentiment Analysis. This network primarily integrates frequency-domain decoupling with semantic filtering to process high- and low-frequency image information separately, thereby enhancing model performance. Specifically, we designed a Dynamic Frequency Domain Decoupling Module that applies discrete wavelet transforms for differentiated image processing. This module, combined with a Dual-Domain Loss Function, constrains consistency between the text semantic space and the frequency distribution of the optimized image features, preventing sentiment information loss from excessive filtering. The module also incorporates two key components: a Global Semantic Sentiment Component (GSSC) and a High-Frequency Filtering Component (HFFC). In the GSSC component, we designed a Hybrid Mamba to leverage text in capturing global semantic information from low-frequency image data. Furthermore, our HFFC component generates a dynamic weight matrix guided by text, enabling quantitative noise suppression. Additionally, we developed a Multi-Grained Semantic Purification Module to filter noise at the word, phrase, and sentence levels. Experimental findings from publicly accessible datasets indicate that our proposed model achieves competitive performance compared with existing methods on multimodal sentiment analysis under the adopted experimental settings and sarcasm detection tasks, validating the effectiveness of our noise-suppression method in cross-modal sentiment analysis. A key limita-

*Corresponding authors.

tion of the model is its use of DWT: downsampling-related resolution loss in low-frequency subbands and independent subband partitioning compromise the capture of large-scale global structural correlations and local-global feature modelling, which advanced transform techniques can alleviate.

Key words: sentiment analysis, multimodal, frequency domain, semantic filtering.

1. Introduction

Sentiment is a subjective response to external stimuli, and sentiment analysis primarily aims to understand human sentiment using existing knowledge (Pandey and Vishwakarma, 2024; Singh *et al.*, 2024). With the widespread use of social media, there has been an increasing number of ways to combine image and text to express sentiment. As a result, multimodal sentiment analysis has gained significant attention from researchers (Sun *et al.*, 2025; Zhu *et al.*, 2023). Compared to traditional text-based sentiment analysis, multimodal sentiment analysis requires handling more modalities, posing greater challenges to researchers (Gandhi *et al.*, 2023). Multimodal sentiment analysis has broad potential in social-media analysis, education, healthcare, and other human-centered applications (Gandhi *et al.*, 2023; Do *et al.*, 2024; Lu *et al.*, 2024; Das and Singh, 2023). By jointly modelling textual and visual evidence, these methods can capture sentiment cues that may be incomplete or ambiguous in either modality alone (Zhu *et al.*, 2022). More broadly, data-driven intelligent technologies have also been applied to shipping communication security, privacy-preserving collaborative learning, and secure maritime data management (Feng *et al.*, 2026; Li *et al.*, 2025b,a).

In the past few years, considerable advances have been made in multimodal sentiment analysis (Wang *et al.*, 2025a; Zhao *et al.*, 2025a; Wang *et al.*, 2025c). Early research primarily focused on performing sentiment analysis by examining the correlation between images and texts, which played a crucial role in improving sentiment classification accuracy (Zhao *et al.*, 2019; Truong and Lauw, 2019). Other studies have introduced cross-modal mechanisms, such as multi-level attention mechanisms and information correlation modelling, to model images globally (Xue *et al.*, 2022; Chen *et al.*, 2023; Wang *et al.*, 2025b), further enhancing the interaction between modalities, or explored the deep fusion of multimodal features through Transformer technologies (Kim and Park, 2023; Aziz *et al.*, 2025; Wu *et al.*, 2025). Some studies have also incorporated colour cues and cross-modal translation mechanisms for sentiment analysis (An and Zainon, 2023; Zhang *et al.*, 2024). Despite significant advancements in inter-modal interaction modelling, Kim *et al.* (2021) argue that networks dealing with images should be more complex than those for text. However, existing research has not deeply explored the impact of the global structure of images on sentiment understanding, nor has it adequately addressed the interference of latent noise within modalities during sentiment feature extraction. This noise may originate from irrelevant background information in the image or from redundant expressions in the text, hindering the precise extraction of sentiment features and reducing the model's accuracy (Zhou *et al.*, 2024, 2026a, 2026b).

To tackle this problem, this work proposes a new approach called **F**requency **D**omain Decoupling and **S**emantic **F**iltering **Net**work for Multimodal Sentiment Analysis (FDSF-Net). The objective is to suppress intra-modal noise interference and enhance the expressive ability of sentiment features within each modality. Specifically, we design a dynamic frequency-domain decoupling module that applies discrete wavelet transforms to the image, dividing its frequency-domain information into Low-Frequency (LF) global semantic sentiment bands and High-Frequency (HF) noise-sensitive bands, allowing us to distinguish between sentiment information and image noise. In this module, we design different components for low and high frequencies to enable differential processing. Specifically, we have developed a global semantic modelling component and an HF filtering component. The global semantic modelling component, leveraging Mamba's powerful temporal modelling capabilities, captures the image's structural information in the LF domain, enabling a thorough understanding of its potential sentiment. The HF filtering component primarily filters the HF noise-sensitive domain, thereby improving the model's ability to capture sentiment cues. To further optimize the frequency-domain decoupling effect, we introduce a dual-domain loss function. This loss function enforces consistency between the text semantic space and frequency-decoupled image features in the frequency domain during optimization, thereby preventing the loss of sentiment information that may result from excessive filtering. This design allows our model to not only accurately extract sentiment features from each modality when processing cross-modal information but also to minimize the influence of noise. Additionally, we have designed a multi-granularity semantic purification module to filter noise at multiple levels in the text. This module filters text information temporally at the word, phrase, and sentence levels, alleviating redundant or irrelevant sentiment information and improving the accuracy of text-based sentiment feature extraction. Finally, we fuse the features of both modalities using the Cross-Spectral Interaction Module. This module, through a cross-frequency-domain interaction mechanism, fully considers the interrelationship between image and text, thereby optimizing the fusion of sentiment features. Experimental validation on publicly available datasets demonstrates the effectiveness of our method.

The contributions of our work are as follows:

- To propose a multimodal sentiment analysis model based on frequency domain decoupling and semantic filtering, which improves the model's sentiment analysis capability by modelling the global structure of the image and analysing noise.
- To design a dynamic frequency domain decoupling module and a multi-granularity semantic purification module. The former processes different frequency subbands of the image differentially, while the latter performs noise suppression at multiple granularities of the text, thereby effectively enhancing the model's ability to extract sentiment cues.
- The proposed model achieves competitive performance on publicly available datasets under the reported single-run experimental setting.

2. Related Work

2.1. *Text Sentiment Analysis*

Traditional approaches primarily rely on conventional machine learning algorithms. For example, Moraes *et al.* (2013) conducted a comparison using SVM and ANN for document-level sentiment classification tasks. They used the bag-of-words model for feature selection and weighting, and discussed the classification accuracy of both methods across different contexts within a standard evaluation setting. Shunxiang *et al.* (2023) proposed a model based on sentiment intensity and PU learning. They divided reviews into different subsets based on sentiment intensity and used SCAR and Spy techniques to extract initial positive and negative samples. A semi-supervised PU learning detector was then constructed to iteratively detect fake reviews from incoming streaming data. Some studies focus on improving dictionaries and statistical methods to better understand text for sentiment analysis. For instance, Kang *et al.* (2012) addressed the issue of inadequate sentiment lexicons in restaurant reviews by proposing a new sentiment lexicon (sentilexicon) and improving the Naive Bayes algorithm. They combined unigrams and bigrams as features, narrowing the accuracy gap between positive and negative sentiment classification, and outperforming SVM and the original Naive Bayes in recall and precision. Wang *et al.* (2023b) proposed an automatic method for generating fine-grained sentiment lexicons by constructing a seed lexicon through sentiment-sentiment transfer, extending the lexicon using graph propagation, and performing multi-information fusion based on neural networks. The resulting FGSL demonstrated strong performance across a range of sentiment analysis tasks. Pashchenko *et al.* (2022), using the NRC sentiment lexicon and unsupervised learning, explored the relationship between sentiment and star ratings in hotel and tourism reviews, finding that customer feedback on different sentiment aspects varied with ratings. However, some studies suggest that relying on a single method may limit the model's expressive power; therefore, a combination of methods may be more effective (Rodríguez-Ibáñez *et al.*, 2023). For example, Bibi *et al.* (2022) presented a framework for unsupervised learning that utilizes concepts and hierarchical clustering to analyse sentiment on Twitter. The results showed that unsupervised learning was comparable to supervised learning techniques. Wang *et al.* (2019) explored Twitter sentiment analysis by combining textual information with sentiment diffusion patterns. They proposed the SentiDiff iterative algorithm, which considered the interaction between both elements to enhance sentiment analysis performance, and used sentiment diffusion patterns for the first time to improve Twitter sentiment analysis. While these methods have made significant contributions to sentiment analysis tasks, the powerful expressive capabilities of deep learning have led many studies to adopt deep learning to address issues in this field (Tai *et al.*, 2015; Tang *et al.*, 2015; Dai *et al.*, 2021). For example, Chen (2015) proposed two convolutional neural network models, Parallel CNN and Deep CNN, based on word2vec word embeddings, to identify question-answer relationship candidates in QA systems. They used convolutional layers to extract semantic features, and pooling and fully connected layers to summarize them. Usama *et al.* (2020) introduced a model based

on RNN and CNN with an attention mechanism. They first used CNN to extract sentence features, then applied an attention mechanism to compute contextual weights for these features, and finally input the features and weights into an RNN for sentiment analysis. Tian *et al.* (2020) proposed the SKEP model, which incorporates sentiment masking and three sentiment prediction targets to embed sentiment information into pre-trained representations. The model significantly outperformed baseline models, achieving new best results on most test sets. The widespread use of attention mechanisms has provided new perspectives for sentiment analysis of text. Several studies have combined attention mechanisms to understand sentiment in text. For example, Zhai *et al.* (2020) proposed the Multi-AFM model, which generates contextual representations using gating units, this model was applied to sentiment analysis of educational big data, improving classification performance. Parveen *et al.* (2023) proposed the GARN framework, which integrates RNN and attention mechanisms to extract sentiment features, perform feature selection, and conduct sentiment classification on Twitter. More recently, Zhang *et al.* (2025b) proposed a textual graph representation with syntactic weighting that combines word-position graph structure, syntactic weights, attention, and external knowledge to model implicit sentiment. Beyond sentiment analysis, Zhao *et al.* (2025b) demonstrated that hyperbolic graph attention can integrate hierarchical semantic representations with long-context information for technical keyphrase extraction, illustrating the broader value of graph-based long-context modelling for text understanding.

Despite the significant achievements of deep learning in text sentiment analysis, the widespread use of social media means that sentiment expression is often not limited to text alone. Therefore, extending the powerful capabilities of deep learning to multimodal sentiment analysis by integrating information from different modalities to more accurately identify and understand sentiment is the primary focus of this study.

2.2. *Multimodal Sentiment Analysis*

Multimodal sentiment analysis poses a greater challenge than traditional text-based sentiment analysis, as it requires integrating information from multiple modalities to uncover underlying sentiment cues. Traditional methods predominantly rely on attention mechanisms to enable models to focus on key information across different modalities, thereby enhancing sentiment analysis performance (Truong and Lauw, 2019; Xue *et al.*, 2022). For instance, Zhang *et al.* (2023) utilized the Transformer model to extract both image and text features, incorporating LSTM and attention to better emphasize important information. Li *et al.* (2023), Liang *et al.* (2025) introduced tensors to extract and fuse features from different sources, such as text, images, and audio, in order to capture the complex data patterns in multimodal sentiment analysis. Additionally, an improved bilinear fusion method was employed, in which the model dynamically adjusts the fusion weights in real time based on the features of both images and text to achieve weighted fusion. Hu *et al.* (2024) proposed a three-channel multimodal fusion framework where the second channel removes redundant information from auxiliary modalities, while the third channel enhances the significance of the primary modality, and the integration of features from the three channels

is managed by a multi-channel information fusion gate. Wang *et al.* (2025b) introduced an adaptive attention module that dynamically adjusts the contributions of text and image features by leveraging cross-modal attention to extract shared representations, thereby improving the performance of both modalities. Additionally, sentiment information is used to guide the extraction of features relevant to sentiment. Several studies have focused on effectively fusing features from different modalities to leverage their complementary information (Zhao *et al.*, 2019; Wang *et al.*, 2022). For example, Mai *et al.* (2022) employed intra-modal, cross-modal, and semi-contrastive learning concurrently, thoroughly exploring cross-modal interactions and learning the relationships between samples and categories, thereby reducing the modality gap. Moreover, they introduced refinement terms and modality intervals to learn single-modal pairs better. An and Zainon (2023) enhanced sentiment analysis accuracy by integrating semantic information and image colour cues from image-text pairs. The model included a feature extraction module to extract semantic and colour features, a feature interaction module to facilitate information exchange between features via cross-attention mechanisms, and a label prediction module to consolidate features and strengthen multimodal sentiment analysis. Cheng *et al.* (2023) proposed an Attention Temporal Convolution Network (ATCN) to enhance the representation of temporal features in a single modality, alongside a Multi-layer Feature Fusion (MFF) model to improve multimodal fusion performance, fused features of different levels based on feature correlation, and used cross-modal multi-head attention to explore relationships among low-level features.

In addition, semantic correlations and sentiment consistency between different modalities are crucial for sentiment analysis, leading some studies to incorporate contrastive learning to enhance model performance. For instance, Zhang *et al.* (2024) used a translation network encoder to capture shared concepts between vision and text, addressing the issue of missing modalities. Based on this framework, a single-modal weight adaptation strategy was introduced to leverage meta-learning from a few labelled samples to learn single-modal weights. Wang *et al.* (2025a) explores modality correlations by a multi-layer cross-modal interaction module. Subsequently, feature fusion was performed using a multimodal fusion module that incorporated contrastive learning to uncover key sentiment features across different categories and samples.

With the growing use of cutting-edge methods, including generative and graph models, more research is focusing on these areas. For example, Huan *et al.* (2023) proposed a UniMF model, which includes a translation module and a prediction module. The translation module generates missing modalities using existing modal information through Multimodal Generation Masking (MGM) and a Multimodal Generation Transformer (MGT). The prediction module integrates multimodal information via an attention mechanism to generate predictions, using a Multimodal Understanding Transformer (MUT) that incorporates Multimodal Understanding Masks (MUM) and Multimodal Sequence (MMSeq) representations for unified multimodal understanding. Zhao *et al.* (2025a) generated virtual modalities to replace missing ones and aligned the semantic space of virtual and missing modalities through contrastive loss. Wang *et al.* (2023a) addressed sentiment information in multimodal data from both global and local fine-grained perspectives. The

global perspective obtained overall sentiment representations from text-image caption pairs, while the local perspective explored fine-grained information in text and images via two graph structures. Ultimately, combining global and local sentiment information provided aspect-level sentiment polarity. Wang *et al.* (2025d) captures sentiment correlations within and across modalities through a graph-based architecture. More recently, Zhang *et al.* (2025a) proposed a Multimodal Semantic Fusion Network that uses gated attention for cross-modal alignment, graph convolutional networks to model interactions among aligned features, and the integration of explicit and implicit sentiment semantics.

Although these methods have made progress in multimodal sentiment analysis tasks, they have not fully exploited the global structural information in images and have overlooked noise interference during cross-modal interactions. Given that multimodal data often contains substantial redundancy and noise, these interferences can disrupt the model's training and degrade performance. Thus, the effective differentiation of image processing and noise mitigation between image and text data to improve the accuracy of multimodal sentiment analysis is the primary focus of this study.

3. Proposed Method

This section presents the architecture and core components of the proposed FDSF-Net framework in Fig. 1.

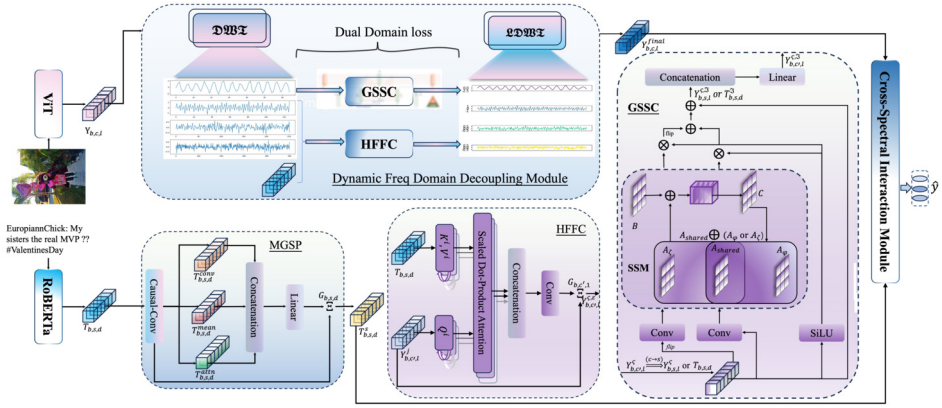


Fig. 1. Framework of FDSF-Net. The model first employs ViT and RoBERTa to extract visual features $Y_{b,c,l}$ and textual features $T_{b,s,d}$, where b denotes the batch index, c the number of image patches, l the visual embedding dimension, s the text sequence length, and d the textual embedding dimension. The Dynamic Frequency Domain Decoupling Module (dashed box 1) decomposes input images into low-frequency (LF) and high-frequency (HF) components via discrete wavelet transform (DWT), and reconstructs optimized visual features $\mathbf{Y}^{final}_{b,c,l}$ using inverse DWT (IDWT). The Multi-Grained Semantic Purification Module (dashed box 2) refines T into denoised textual features T_s by suppressing noise at word-, phrase-, and sentence-level granularities. The Cross-Spectral Interaction Module (dashed box 3) performs bidirectional cross-modal attention between $\mathbf{Y}^{final}_{b,c,l}$ and T_s, producing the fused representation F_{fused} for sentiment prediction $\hat{y}$. Solid arrows indicate the direction of information flow, while dashed boxes highlight the three core modules.

3.1. *Task Definition and Description*

Multimodal sentiment analysis primarily aims to detect the sentiment conveyed across different modalities. The multimodal data in this study consists of image and text. Given a text-image pair (X_t^j, X_i^j) along with the corresponding task label y^j, the dataset is defined as follows:

$$D = \left\{\left(X_t^j, X_i^j\right), y^j\right\}_{j=1}^{N}, \tag{1}$$

where X_t^j represents the j-th text sample in the dataset, X_i^j is the corresponding image associated with X_t^j, and y^j denotes the task-specific label. Specifically, for our core task of multimodal sentiment analysis, y^j is a binary label for the HFM dataset, with 1 indicating positive sentiment and 0 indicating negative sentiment, whereas it is a 3-dimensional one-hot vector for the MVSA datasets, indicating negative, neutral, and positive sentiment, respectively. To further improve the generalization ability of our proposed method, we also conduct experiments on multimodal sarcasm-detection datasets, where the labels are split into sarcasm and non-sarcasm. The mapping between these sarcasm detection labels and sentiment labels follows the approach of Wei *et al.* (2023): sarcasm labels are mapped to negative sentiment, and non-sarcasm labels are mapped to positive sentiment. N refers to the total number of samples in the dataset, and the text-image pair (X_t^j, X_i^j) is fed into the model F to predict the task-specific label, which realizes the corresponding multimodal classification goal:

$$F\left(X_t^j, X_i^j\right) \longmapsto y^j. \tag{2}$$

This formulation describes the prediction process of the model for each text-image pair in the dataset.

3.2. *Feature Extraction Module*

3.2.1. *Text Feature Extraction*

A pre-trained model, RoBERTa-base (Liu, 2019), trained on large-scale datasets, is used as the text encoder. First, the text is split into a sequence of tokens, and then these sequences are input into RoBERTa to obtain text features $\mathbf{T} \in \mathbb{R}^{B\times S\times d_t}$, where B is the batch size, S is the token-sequence length, and $d_t = 768$, as represented by the following formula:

$$\mathbf{T} = \text{RoBERTa}(\mathbf{X}_t) = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_S], \tag{3}$$

where $\mathbf{t}_e \in \mathbb{R}^{B\times d_t}$ represents the batch of embeddings at token position e. The equivalent element-wise notation used below is $\mathbf{T}_{b,s,d}$.

3.2.2. *Image Feature Extraction*

We employ the Vision Transformer (ViT) (Dosovitskiy *et al.*, 2020) as our image encoder. Initially, the input image X_i with a fixed size of $224 \times 224 \times 3$ (*height* $\times$ *width* $\times$ *channels*)

is segmented into a collection of m-flattened 2D patches. These patches are subsequently fed into ViT to extract image features $Y_v \in \mathbb{R}^{m \times d_v}$, as illustrated in equation (4):

$$Y_v = ViT(X_i) = [e_1, e_2, \ldots, e_m], \tag{4}$$

where $e_p \in \mathbb{R}^{d_v}$ is the embedding of the p-th image patch.

3.3. *Dynamic Frequency Domain Decoupling Module*

To accurately capture sentiment-related features within an image, our method employs the Discrete Wavelet Transform (DWT) to decompose the image data into specialized frequency domains, thereby preserving both global structure and fine-grained textures (Zhao *et al.*, 2021). Specifically, we utilize time-frequency analysis techniques—including wavelets (Mallat, 1989; Daubechies and Heil, 1992)—to minimize noise while isolating distinct visual components. Through a series of low-pass and high-pass filtering operations, our model decomposes the input into four sub-bands: LL (Low-Low), LH (Low-High), HL (High-Low), and HH (High-High). In our technical framework, the LL sub-band constitutes the Low-Frequency (LF) domain, which retains the overall contour and large-scale global information necessary for understanding the primary sentiment tendency (Huang *et al.*, 2021). Simultaneously, the combination of LH, HL, and HH sub-bands forms the High-Frequency (HF) domain, which focuses on extracting horizontal, vertical, and diagonal edge details. By decoupling the image into these domains, the model can effectively capture subtle visual cues—such as textures and corners—that are often overlooked in spatial analysis but are vital for precise sentiment judgment. The formal decomposition process, illustrated in Fig. 2, is expressed as follows:

$$\mathbf{Y}_{b,c,l} = \mathcal{DWT}(\mathbf{Y}) = \left(\mathbf{Y}^{\varsigma}_{b,c',l}, \mathbf{Y}^{h}_{b,c',l}\right), \tag{5}$$

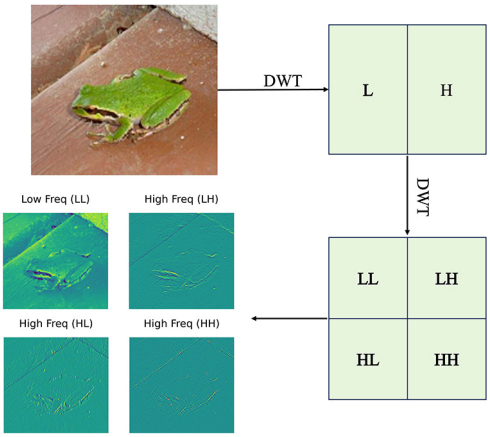


Fig. 2. Illustration of the discrete wavelet transform decomposition process used in the proposed frequency-domain decoupling module.. The input image is first subjected to low-pass and high-pass filtering in the horizontal direction, then filtered again in the vertical direction, resulting in four sub-bands: LL, LH, HL, and HH.

where $\mathbf{Y}_{b,c',l} \in \mathbb{R}^*$ is a three-dimensional tensor, where b is the batch index, c is the number of patches in the image, c' denotes the adjusted number of image patches after decomposition, and l is the embedding hidden layer dimension. The wavelet transform decomposes the input image into low-frequency feature $\mathbf{Y}^{\varsigma}_{b,c',l}$ and high-frequency feature $\mathbf{Y}^{h}_{b,c',l}$.

In this module, we implement a Global Semantic Sentiment Component (GSSC) and a High-Frequency Filtering Component (HFFC). After processing the GSSC and HFFC components separately, we use the inverse wavelet transform to synthesize the processed LF and HF parts back into the original feature space. The processed LF component, denoted as $\mathbf{Y}^{\varsigma,\Im}_{b,c',l}$ with the superscript $\Im$ indicating the output of GSSC, and the processed HF component, denoted as $\mathbf{Y}^{h,\ell}_{b,c',l}$ with the superscript ℓ indicating the output of HFFC, together reconstruct the final image. This process can be expressed by the following formula:

$$\mathbf{Y}^{final}_{b,c,l} = \mathcal{IDWT}\left(\mathbf{Y}^{\varsigma,\Im}_{b,c',l}, \mathbf{Y}^{h,\ell}_{b,c',l}\right) \tag{6}$$

which defines the corresponding representation.

3.3.1. *Global Semantic Sentiment Component*

The LF components of the visual modality often carry richer global sentiment semantics. By leveraging ViT's image sequence modelling capability, we can apply sequence enhancement methods to images. In recent years, state space models have developed rapidly in the field of sequence modelling (Gu *et al.*, 2021, 2022; Fu *et al.*, 2022), attracting the attention of many researchers due to their linear computational complexity and ability to model long-range dependencies. SSMs originated from control theory and model the dynamic evolution process of sequences through state equations, which can be expressed by the following formula:

$$h'(t) = Ah(t) + Bx(t), \qquad y(t) = Ch(t), \tag{7}$$

where $\mathbf{A} \in \mathbb{R}^{d_s \times d_s}$, $\mathbf{B}$ and $\mathbf{C}$ are projection matrices, and d_s denotes the state dimension, which is different from the dataset size N defined in Section 3.1.

However, Mamba Gu and Dao (2023) further discretizes the parameters A and B into $\bar{A}$ and $\bar{B}$ through the time scale parameter Δ. The discretization calculation can be expressed as:

$$\bar{A} = \exp(\Delta A), \bar{B} = (\Delta A)^{-1}\left(\exp(\Delta A) - I\right) \cdot \Delta B. \tag{8}$$

At this time, the continuous-time model in (7) yields the following discrete recurrence:

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t, \qquad \mathbf{y}_t = \mathbf{C}\mathbf{h}_t. \tag{9}$$

Mamba has now achieved significant success in deep learning tasks (Ge *et al.*, 2024; Pan *et al.*, 2025; Xie *et al.*, 2024). To effectively enhance the latent sentiment features

of an image, we introduce the bidirectional mamba (Zhu *et al.*, 2024). This bidirectional mamba not only processes the original features but also performs a flip operation on them, enabling a deeper comprehension of contextual information and uncovering sentiment connections between image and text.

In the meantime, inspired by Ye *et al.* (2025), we employ a collaborative Mamba architecture to handle information interaction between image and text. However, unlike the implementation in Ye *et al.* (2025), we do not use the same A matrix for both image and text. Instead, we introduce a more refined parameterization of Mamba's core state matrix A, incorporating both a shared component and a modality-specific incremental component. Specifically, we define a cross-modal shared core matrix A_{shared} for the model, which is used to learn the fundamental dynamics and long-term dependencies common across all sequential data. Building upon this, we introduce a sentiment modality-specific incremental matrix A_ζ and A_φ for each modality. This can be expressed by the following formula:

$$\begin{cases} (\bar{\mathbf{A}}'_\zeta, \bar{\mathbf{B}}'_\zeta) = \text{Discretize}(\Delta_\zeta, \mathbf{A}_{shared} + \mathbf{A}_\zeta, \mathbf{B}_\zeta), \\ (\bar{\mathbf{A}}'_\varphi, \bar{\mathbf{B}}'_\varphi) = \text{Discretize}(\Delta_\varphi, \mathbf{A}_{shared} + \mathbf{A}_\varphi, \mathbf{B}_\varphi), \end{cases} \tag{10}$$

where $\text{Discretize}(\Delta, \mathbf{A}, \mathbf{B})$ maps the continuous-time pair $(\mathbf{A}, \mathbf{B})$ to $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ using the input-dependent step size Δ. Thus, both state-space arguments are explicit. After the collaborative Mamba, the image and text representations are concatenated and linearly transformed.

3.3.2. *High-Frequency Filtering Component*

The HF information in the image often contains noise components, and direct processing may disrupt the continuous semantic information captured by ViT. While ViT captures HF information from an image, it also maintains global semantic consistency. Therefore, directly using the image's HF information for processing may lead to excessive filtering of these HF components, thereby destroying important sentiment cues. To avoid this issue, we introduce a text-based semantic filtering mechanism that leverages textual semantics to generate a weight matrix, enabling refined suppression of HF noise in images via a frequency-band-aware approach. Additionally, based on the image features filtered by text semantics, we incorporate a residual connection to preserve the image's core semantic information during filtering, thereby achieving more accurate sentiment analysis. The weighted values of the image HF $\mathbf{Y}^{h,j}_{b,c',l}$ and the text features are calculated through the Multi-Head Attention mechanism. For the Multi-Head Attention mechanism, we can expand it into a weighted sum of multiple heads:

$$\mathbf{A}^{\text{hf},j}_{b,c',l} = \sum_{q=1}^{H} \alpha_q \cdot \text{Attn}_q\left(\mathbf{Y}^{\text{hf},j}_{b,c',l}, \mathbf{T}_{b,s,d}, \mathbf{T}_{b,s,d}\right), \tag{11}$$

where H is the number of attention heads, q is the summation index, α_q is the weight of head q, and the left-hand side does not retain the dummy head index. We obtain the filter

matrix $\mathbf{g}_{b,1,c'} \in \mathbb{R}^{B\times 1\times c'}$ through a convolution operation, which can be expressed as:

$$\mathbf{g}^{j}_{b,1,c'} = \sigma\big(\mathrm{Conv}_{1d}\big(\mathbf{A}^{h,j}_{b,c',l}\big)\big), \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\mathrm{Conv}_{1d}(\cdot)$ denotes a one-dimensional convolution operation. Therefore, the denoising of the HF $\mathbf{Y}^{h,j}_{b,c',l}$ can be expressed by the following formula:

$$\mathbf{Y}^{h,j}_{b,c',l} \to \mathbf{Y}^{h,j,\ell}_{b,c',l} = \mathbf{Y}^{h}_{b,c',l} \odot \mathbf{g}^{j}_{b,c',1}. \tag{13}$$

This operation represents the denoising process of the high-frequency features. We store multiple HF features in a list, which can be represented as:

$$\mathbf{Y}^{h,j,\ell}_{b,c',l} \xrightarrow{list} \mathbf{Y}^{h,\ell}_{b,c',l}. \tag{14}$$

This formulation indicates that the denoised high-frequency features are aggregated into a list for subsequent processing.

3.4. *Multi-Grained Semantic Purification Module*

In multimodal tasks, text data often contains rich information, and appropriate processing of text is essential for these tasks (Wang *et al.*, 2024; Sun *et al.*, 2024). However, text often contains various types of noise that manifest at different granularities, negatively affecting the accuracy of sentiment analysis. To effectively address this issue, we designed a Multi-Grained Semantic Purification Module. This module aims to refine text data by filtering out noise at different granularities, thereby extracting more precise sentiment information. To further enhance temporal feature representation, we first apply a causal convolutional operation to extract local temporal dependencies.

$$\mathbf{T}^{e}_{b,s,d} = \mathcal{K}(\mathbf{T}_{b,s,d}) \odot \sigma(\mathbf{W}_g \cdot \mathbf{T}_{b,s,d} + \mathbf{b}_g), \tag{15}$$

where s is the sequence length. The causal convolution operation $\mathcal{K}$ enhances the representational capacity of temporal features. Subsequently, we generate multi-level feature representations using a multi-grained feature-extraction method that combines multi-head self-attention, convolution, and mean pooling. The formula for calculating the weight distribution for the input text features $\mathbf{T}^{e}_{b,s,d}$ is as follows:

$$\mathbf{G}_{b,s,d} = \sigma\big(LN\big[\mathcal{A}\big(\mathbf{T}^{e}_{b,s,d}\big), \mathcal{C}\big(\mathbf{T}^{e}_{b,s,d}\big), \mathcal{M}\big(\mathbf{T}^{e}_{b,s,d}\big)\big]\big), \tag{16}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, and $\mathrm{LN}(\cdot)$ denotes layer normalization (Ba *et al.*, 2016). $\mathcal{A}(\mathbf{T}_{b,s,d})$ denotes the multi-head self-attention operation, $\mathcal{C}(\mathbf{T}_{b,s,d})$ denotes the convolution operation, and $\mathcal{M}(\mathbf{T}_{b,s,d})$ denotes the mean pooling operation. Then, $\mathbf{G}_{b,s,d}$ is used to weight and adjust the features processed by causal convolution, yielding the dynamically adjusted feature representation as follows:

$$\mathbf{T}^{s}_{b,s,d} = \mathbf{T}^{e}_{b,s,d} \odot \mathbf{G}_{b,s,d}. \tag{17}$$

This operation produces the dynamically adjusted feature representation after applying the learned gating mechanism.

3.5. *Cross-Spectral Interaction Module*

After processing, crucial information in both image and text features is preserved. To capture the correlation between image and text more precisely, we employ bidirectional cross-modal attention for interaction. Bidirectional cross-modal interaction is modelled in two directions: $E_\alpha \rightarrow E_\beta$, where $(E_\alpha, E_\beta) \in \{(\mathbf{Y}^{final}_{b,c,l}, \mathbf{T}^{s}_{b,s,d}), (\mathbf{T}^{s}_{b,s,d}, \mathbf{Y}^{final}_{b,c,l})\}$. This approach improves the complementarity between image and text information, enhancing the model's understanding of deeper relationships between the two modalities.

$$\begin{aligned}\mathbf{C}_\alpha &= \mathrm{Co}_{atts}(\mathbf{E}_\alpha, \mathbf{E}_\beta) \\ &= \mathrm{softmax}_\beta\left(\frac{(\mathbf{W}_\alpha\mathbf{E}_\alpha)(\mathbf{W}_\beta\mathbf{E}_\beta)^T}{\sqrt{d_k}}\right)(\mathbf{W}_\beta\mathbf{E}_\beta),\end{aligned} \tag{18}$$

where $\mathbf{E}_\alpha$ and $\mathbf{E}_\beta$ denote the feature representations of two modalities (e.g. image and text), $\mathbf{W}_\alpha$ and $\mathbf{W}_\beta$ are learnable projection matrices, d_k denotes the scaling dimension, and $\mathrm{Co}_{atts}(\cdot)$ denotes the cross-modal attention operation. Simultaneously, to enhance sequence position awareness, we introduce rotary positional embeddings for both the input query and the key (Su *et al.*, 2024). Next, we concatenate the processed image features $\mathbf{G}_{img} \in \mathbb{R}^{b,c,l}$, text features $\mathbf{G}_{txt} \in \mathbb{R}^{b,s,d}$, and cross-attention features $\mathbf{C}_\alpha \in \mathbb{R}^{b,c,l}$ and $\mathbf{C}_\beta \in \mathbb{R}^{b,s,d}$. The concatenation operation joins these features column-wise, yielding a new combined feature:

$$\mathbf{F}_\gamma = \mathcal{C}[\mathbf{G}_{img}, \mathbf{G}_{txt}, \mathbf{C}_\alpha, \mathbf{C}_\beta], \tag{19}$$

where $\mathcal{C}$ denotes the concatenation operation, and $\mathbf{F}_\gamma \in \mathbb{R}^{b,s',d}$ represents the concatenated feature. We then compute the mean of the merged features, fusing the multimodal features into a more compact representation:

$$\mathbf{F}_{fused} = \frac{1}{S'}\sum_{j=1}^{S'}\mathbf{F}_{\gamma,:,j,:}, \tag{20}$$

where $\mathbf{F}_{fused} \in \mathbb{R}^{B\times d_f}$ is the final fused representation, S' is the total number of concatenated positions, and d_f is the common projected feature dimension. This removes the previously undefined normalizing constant $4D$.

3.6. *Optimization and Classification*

3.6.1. *Optimization*

For the loss function, we use Focal Loss (Lin *et al.*, 2017) as the primary loss, combined with the proposed dual-domain loss as an auxiliary loss. These are integrated through

learnable parameters. The Focal Loss function is defined as follows:

$$\mathcal{L}_{focal}(\mathbf{x}, y) = -\alpha_y \big(1 - \text{softmax}(\mathbf{x})_y\big)^{\gamma} \log\big(\text{softmax}(\mathbf{x})_y\big), \tag{21}$$

where $\mathbf{x}$ represents the network output (logits), y denotes the ground-truth class label, $\text{softmax}(\mathbf{x})_y$ denotes the predicted probability of class y, α_y is the class-specific balancing factor, and γ is the focusing parameter that emphasizes hard-to-classify samples. The final Focal Loss for each sample is calculated as follows:

$$\mathcal{L}_{focal} = -\alpha_y (1 - P_y)^{\gamma} \log P_y, \tag{22}$$

where P_y denotes the predicted probability of the ground-truth class. The dual-domain loss comprises a spatial gradient loss and a frequency-domain loss. The spatial term preserves structural information by comparing horizontal and vertical finite differences of the reconstructed and target representations. The frequency-domain loss measures the transformed reconstruction error. The spatial gradient loss is defined as follows:

$$\mathcal{L}_{spatial} = \sum_{b=1}^{B} \Big(\big\| \nabla_x \mathbf{X}_b^{final} - \nabla_x \mathbf{X}_b^{target} \big\|_2^2 + \big\| \nabla_y \mathbf{X}_b^{final} - \nabla_y \mathbf{X}_b^{target} \big\|_2^2 \Big), \tag{23}$$

where B denotes the batch size, ∇_x and ∇_y are first-order finite-difference operators along the two spatial axes, and $\mathbf{X}_b^{target}$ is the target representation.

The frequency domain loss is defined as follows:

$$\mathcal{L}_{freq} = \sum_{b=1}^{B} \sum_{c=1}^{C} \sum_{l=1}^{L} \big\| \mathcal{DWT}\big(\mathbf{X}_{b,c,l}^{final} - \mathbf{X}_{b,c,l}^{target}\big) \big\|_2^2, \tag{24}$$

where $\mathcal{DWT}$ represents the discrete wavelet transform, and $\| \cdot \|_2^2$ denotes the squared ℓ_2 norm. The correction here moves the DWT inside the L2 norm calculation: first compute the difference between the reconstructed and original image, then perform DWT on the difference tensor, and finally calculate the L2 norm of the transformed result to measure the reconstruction error in the frequency domain.

The total dual-domain loss is combined through adaptive weighting of the spatial and frequency domain losses:

$$\mathcal{L}_{dual} = \omega_s \mathcal{L}_{spatial} + \omega_f \mathcal{L}_{freq}, \tag{25}$$

where ω_s and ω_f are learnable weight parameters used to control the relative importance of the spatial and frequency domain losses. The final total loss is the weighted sum of the Focal Loss and the dual-domain loss:

$$\mathcal{L}_{total} = \mathcal{L}_{focal} + \lambda \mathcal{L}_{dual}, \tag{26}$$

where λ is a learnable parameter that adjusts the weight balance between the focal loss and the dual-domain loss. By learning this weighted loss, the model can simultaneously optimize both classification accuracy and image reconstruction quality.

3.6.2. *Classification*

After fusing cross-modal information through the Cross-Spectral Interaction Module, we project it into the classification space. The formulation is as follows:

$$r = \mathbf{W}_c\big(\mathrm{LN}(\mathbf{F}_{fused})\big) + \mathbf{b}_c, \tag{27}$$

$$\hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{r}), \tag{28}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization, $\mathbf{W}_c \in \mathbb{R}^{K\times d_f}$ and $\mathbf{b}_c \in \mathbb{R}^K$ are classifier parameters distinct from the purification-gate parameters $\mathbf{W}_g$ and $\mathbf{b}_g$, d_f is the fused feature dimension, and K is the number of classes ($K = 3$ for MVSA and $K = 2$ for HFM). The softmax output is consistent with the multiclass Focal Loss.

4. Experimentation

4.1. *Datasets*

All our experiments were conducted on the publicly available datasets MVSA-Single, MVSA-Multiple (Niu *et al.*, 2016), and HFM (Cai *et al.*, 2019). MVSA-Multiple is an enhanced version of MVSA-Single, featuring a larger number of image-text pairs suitable for sentiment analysis. We adopted the same preprocessing strategy as Xu and Mao (2017). The HFM dataset is primarily used for multimodal sarcasm detection, a task with two sentiment polarities: positive and negative. For a fair comparison, we used the consistent data preprocessing method described in Cai *et al.* (2019). A statistical analysis of the datasets is presented in Table 1.

MVSA-Single and MVSA-Multiple are available at.[1] The HFM dataset can be downloaded from.[2]

Table 1
Overview of data statistics.

Dataset	Negative	Neutral	Positive	Total
MVSA-S	1358	470	2683	4511
MVSA-M	1298	4408	11318	17024
HFM	14075	–	10560	24635

[1] https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/
[2] https://github.com/headacheboy/data-of-multimodal-sarcasm-detection

4.2. *Hyperparameter Settings and Evaluation Metrics*

4.2.1. *Hyperparameters*

The number of output categories is set according to the target dataset, i.e. 3 for MVSA-Single and MVSA-Multiple, and 2 for HFM, and the maximum text length is 100. The base learning rate is set to 1e-6 following validation-based tuning, while the GSSC module uses a larger learning rate of 3e-5. In the Focal Loss, the balancing factor α is set to a vector of all ones with a length of 3, and the focusing parameter γ is set to 2. The batch size is configured to 12, balancing training efficiency with memory usage. Furthermore, to enhance computational efficiency, avoid local optima, and achieve faster convergence, this paper utilizes the Adam optimizer. We train the model for 50 epochs in total. For the pre-trained model, we adopt RoBERTa-base (12 layers, hidden size 768) and ViT-B/16 (12 layers, patch size 16×16) as the text and image encoders, respectively.

4.2.2. *Hardware Configuration*

All experiments are conducted on a server equipped with an Intel(R) Xeon(R) Gold 6230 CPU @ 2.10 GHz, 256 GB RAM, and 1 NVIDIA RTX 4090 GPU with 24 GB memory. The runtime environment is Python 3.12, PyTorch 2.3.0, and CUDA 11.6. The average training time per epoch is approximately 12 minutes for the MVSA-Single dataset, 28 minutes for the MVSA-Multiple dataset, and 45 minutes for the HFM dataset.

4.2.3. *Evaluation Metrics*

We adopt four widely used evaluation metrics from existing works: *Accuracy*, *Precision*, *Recall*, and *F1-score*.

4.3. *Comparison of Experimental Results*

4.3.1. *Baselines*

To validate the feasibility of our approach, we compare it with various existing methods, detailed as follows:

Text-modality mode.

- CNN (Kim, 2014): A Convolutional Neural Network that achieves its results by repeatedly convolving over local features, making it suitable for scenarios involving local pattern modelling.
- BiLSTM (Zhou *et al.*, 2016): A variant of RNN that has achieved excellent performance in text processing through its sophisticated gating mechanisms.
- BERT (Devlin *et al.*, 2019): A pre-trained model learned from a large corpus of knowledge, offering stronger feature representation capabilities compared to traditional neural networks.

Image-modality mode.

- ResNet-50 (He *et al.*, 2016): A pre-trained model based on a Convolutional Neural Network architecture.

Table 2
Experimental results of multiple comparative models on MVSA and HFM datasets.

Modality	Method	MVSA-single		MVSA-multiple		HFM	
		Acc.	F1.	Acc.	F1.	Acc.	F1.
Text	CNN	0.6819	0.559	0.6564	0.5766	0.8003	0.7532
	BiLSTM	0.7012	0.6506	0.679	0.679	0.819	0.7753
	BERT	0.7111	0.697	0.6759	0.6624	0.8389	0.8326
Image	ResNet-50	0.6467	0.6155	0.6188	0.6098	0.7277	0.7138
	ViT	0.6378	0.6226	0.6194	0.6119	0.7309	0.7152
Multimodal	MultiSentiNet	0.6984	0.6984	0.6886	0.6811	0.8103	0.7799
	HSAN	0.6988	0.669	0.6796	0.6776	0.8174	0.7874
	Co-MN-Hop6	0.7051	0.7001	0.6892	0.6883	0.8344	0.8018
	MGNNS	0.7377	0.727	0.7249	0.6934	0.8402	0.8060
	CLMLF	0.7533	0.7346	0.7200	0.6983	0.8543	0.8487
	MVCN	0.7606	0.7455	0.7207	0.7001	0.8568	0.8523
	CTMWA	0.7591	0.7574	0.7402	0.7384	–	–
	Ours	**0.7711**	**0.7680**	**0.7406**	**0.7400**	**0.8589**	**0.8405**

- ViT (Dosovitskiy *et al.*, 2020): A Transformer-based model that processes images by treating them as sequences.

Multi-Modal mode.

- MultiSentiNet (Xu and Mao, 2017): Identify sentiment by extracting vocabulary that significantly influences the sentiment of the entire tweet.
- HSAN (Xu, 2017): The model leverages image captions to derive visual features, providing extra context for the text.
- Co-MN-Hop6 (Xu *et al.*, 2018): Modelling the interaction between image and text and using shared memory to assist in the analysis.
- MGNNS (Yang *et al.*, 2021): A multi-channel graph neural network has been designed specifically for sentiment detection.
- CLMLF (Li *et al.*, 2022): It aligns and fuses multimodal features based on a Transformer encoder, and employs two contrastive learning tasks to assist in modelling.
- ITIN (Zhu *et al.*, 2022): By utilizing cross-modal alignment and gating mechanisms, the performance of the model has been improved.
- MVCN (Wei *et al.*, 2023): A new Multi-View Calibration Network designed to tackle the issues of modality heterogeneity.
- CTMWA (Zhang *et al.*, 2024): The reliability of the model has been enhanced in cases of modality absence and information imbalance through a cross-modal translation network and a single-modal weight adaptation strategy.

4.3.2. *Comparative Experimental Results and Analysis*

To effectively assess our model's performance, we conducted extensive comparative experiments, as shown in Table 2. Additionally, we performed the following experimental analyses:

1) Experimental data consistently show that the text modality generally outperforms the image modality in all tests. Specifically, the BERT model outperformed the

image-based ViT and ResNet-50 models. For instance, in the MVSA-Single task, BERT achieved an accuracy of 0.7111 and an F1-score of 0.6970, whereas ViT's accuracy was only 0.6378 and its F1-score was 0.6226, significantly lower than BERT's. The superior performance of the text modality is primarily attributed to the rich contextual information inherent in text data, which is crucial for sentiment analysis. BERT can capture deeper semantic associations and sentiment features, thus exhibiting stronger expressive capabilities in sentiment analysis tasks. In contrast, sentiment features in the image modality are often more implicit, and extracting image features is influenced by factors such as content and resolution, which can lead to inferior performance in sentiment analysis. Therefore, the powerful semantic modelling capability of the text modality provides a more distinct advantage in sentiment analysis tasks, while the image modality performs relatively weaker due to its limitations.

2) Compared to single-modality approaches, multimodal methods generally exhibit superior performance. For example, in the MVSA-Single task, the multimodal method CTMWA achieved an accuracy of 0.7591 and an F1-score of 0.7574, while the single-text modality BERT had an accuracy of 0.7111 and an F1-score of 0.6970, and the single-image modality ResNet-50 performed even worse. Multimodal methods leverage the strengths of both images and text by integrating them. Text provides rich semantic information, while an image can offer complementary visual cues, especially for sentiment with visual manifestations. For instance, in tasks with clear sentiment polarity, an image might help the system understand the sentiment context, while text provides precise sentiment words and contextual information. The complementarity of the two modalities enables multimodal methods to provide more comprehensive information in sentiment analysis tasks, thereby improving the model's accuracy and robustness. Consequently, multimodal methods effectively enhance performance in sentiment analysis tasks through cross-modal feature fusion, capturing sentiment information more comprehensively than single modalities and thus improving overall performance.

3) FDSF-Net achieves competitive, but not uniformly superior, performance relative to the compared multimodal methods. In Table 2, it obtains the highest reported accuracy and F1-score on MVSA-Single and the highest reported values on MVSA-Multiple by small margins under the adopted single-run setting. On HFM, it achieves the highest accuracy (0.8589), whereas its F1-score (0.8405) is lower than those of CLMLF (0.8487) and MVCN (0.8523). Accordingly, these results support the competitiveness of the proposed design, but they do not establish uniform superiority. Because repeated-run variance and significance tests are unavailable, the small numerical differences, particularly on MVSA-Multiple, should be interpreted cautiously and not as statistically significant improvements.

The HFM results further suggest that FDSF-Net can be applied to multimodal sarcasm detection. Its accuracy is competitive with the compared methods, while its F1-score does not exceed the strongest baselines. We therefore regard this experiment as preliminary evidence of cross-task applicability rather than proof of superior generalization.

Table 3
Ablation study results, MFS denotes equally fusing the image and text features as Li *et al.* (2022).

Model	MVSA-single		MVSA-multiple		HFM	
	Acc.	F1.	Acc.	F1.	Acc.	F1.
BERT	0.7111	0.697	0.6759	0.6624	0.8389	0.8326
ViT	0.6378	0.6226	0.6194	0.6119	0.7309	0.7152
MFS	0.7217	0.7205	0.7063	0.6851	0.8434	0.8375
Only I	0.6988	0.669	0.6796	0.6776	0.7309	0.7152
Only T	0.7051	0.7001	0.6892	0.6883	–	–
w/o DS	0.7511	0.7458	0.7294	0.7284	0.8468	0.8286
w/o DS-H	0.7644	0.7590	0.7329	0.7319	0.8522	0.8339
w/o DS-L	0.7578	0.7524	0.7318	0.7308	0.8506	0.8323
w/o DS-D	0.7556	0.7502	0.7312	0.7302	0.8489	0.8307
w/o C	0.7622	0.7568	0.7371	0.7361	0.8535	0.8352
w/o M	0.7667	0.7613	0.7388	0.7378	0.8551	0.8367
w/o S	0.7689	0.7634	0.7400	0.7390	0.8572	0.8388
All	**0.7711**	**0.7680**	**0.7406**	**0.7400**	**0.8589**	**0.8405**

4.3.3. *Ablation Study Results and Analysis*

To thoroughly validate the effectiveness of our proposed method, we conducted ablation experiments on the two MVSA datasets and HFM, as shown in Table 3. The experimental scenarios are primarily as follows:

- Only I: Indicates using only image representation.
- Only T: Indicates using only text representation.
- w/o DFDD: Represents removing the Dynamic Frequency Domain Decoupling module to observe its impact on model performance.
- w/o DFDD-H: Removes the image HF-domain operation.
- w/o DFDD-L: Removes the image LF-domain processing operation.
- w/o DFDD-D: Removes both the image LF- and HF-domain operations, retaining only the frequency transformation.
- w/o C: Removes the Cross-Spectral Interaction Module.
- w/o M: Removes the Multi-Granularity Semantic Purification module.
- w/o S: Removes the Spectral Domain Dual Domain Loss, retaining only the classification loss.
- All: Our proposed complete model.

The ablation results show the largest observed decrease when the Dynamic Frequency Domain Decoupling (DFDD) module is removed, followed by the DFDD-D, DFDD-L, and DFDD-H variants. Removing the complete DFDD module eliminates differentiated LF/HF processing; removing DFDD-D retains the transform but removes the decoupling operations; and the DFDD-L and DFDD-H variants isolate the contributions of low- and high-frequency processing. These are descriptive single-run comparisons and are not presented as statistically significant differences.

Furthermore, we analysed the contributions of the Multi-Granularity Semantic Purification Module (M), Cross-Spectral Interaction Module (C), and Spectral Domain Dual

Domain Loss (S) to model performance. The experimental results show that removing the M module had the most significant impact on model performance, followed by the C and S modules. Specifically, after removing "w/o M", the model lost its ability to perform multi-granularity purification of text information. This module optimizes text features via multi-level semantic extraction, enabling the model to better understand the complex relationship between images and text. Without the M module, the model's accuracy in multimodal understanding significantly decreased, particularly in deep fusion of image and text. After removing "w/o C", the cross-spectral interaction module no longer established an effective connection between text and image features, leading to a noticeable weakening of their fusion capability and affecting the model's overall performance in cross-modal tasks. Finally, after removing "w/o S", although this loss function helps enhance the consistency between image and text in the frequency domain, allowing them to align better, its impact on overall performance improvement was relatively limited compared to other modules. Overall, despite the unique role of the S module, its contribution to performance improvement was less significant than that of the M and C modules, indicating that the precise extraction and efficient fusion of text and image features are crucial for the success of multimodal tasks.

In summary, the ablation results indicate that each module contributes to the reported single-run performance. In particular, removing the Dynamic Frequency Domain Decoupling module produces the largest observed decrease. These descriptive differences support the usefulness, rather than establish the statistical superiority, of the complete model.

4.4. *Hyperparameter Analysis*

In deep learning, the choice of hyperparameters significantly impacts model performance. For the Multimodal Sentiment Analysis Model with Frequency-Domain Decoupling and Semantic Filtering proposed in this paper, the order of the Discrete Wavelet Transform (DWT) is a crucial hyperparameter. To thoroughly understand its influence on model performance, a series of experimental analyses was conducted, as shown in Fig. 3. The results indicate that the model performs best when using the db2 wavelet basis, while db1 performs significantly worse, and the performance of db3, db4, and db5 decreases sequentially.

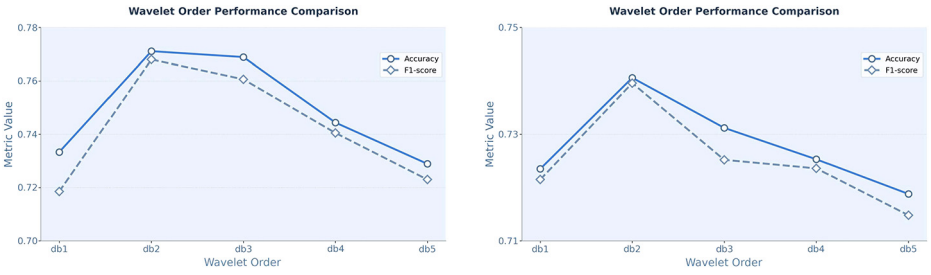


Fig. 3. Wavelet order analysis.

Specifically, the advantage of the db2 wavelet basis lies primarily in its balanced frequency-domain decoupling, which enables it to effectively process both LF global semantic information and HF noise in an image. db2 can not only precisely extract important semantic features from an image but also avoid excessive filtering that could lead to the loss of sentiment information, thereby enhancing overall model performance. In contrast, while the db1 wavelet basis is relatively simpler, its frequency-domain decomposition ability is weaker, failing to effectively handle complex image details. Our method provided insufficient noise suppression, leading to poorer model performance. Furthermore, although db3, db4, and db5 may offer advantages in processing certain frequency-domain details, their overly detailed decomposition methods can lead to excessive filtering and loss of image information, especially in finding a suitable balance between noise reduction and feature retention, ultimately causing a gradual decline in model performance. Therefore, db2 performs optimally in this model because it can suppress unnecessary noise while preserving effective information, thus maximizing the model's effectiveness.

4.5. *Visual Analysis*

To gain a deeper, more intuitive understanding of how our model processes multimodal data, we employed various visualization methods to observe it from different perspectives.

4.5.1. *Heatmap Analysis*

To further explore the sentiment information our proposed model learns from multimodal data, we conducted a visual analysis of four sample sets from the MVSA dataset using Grad-CAM. We computed each token's contribution to the classification result using the model's last attention layer's output and gradients. These contributions were then mapped back to the original image's spatial locations, generating heatmaps as shown in Fig. 4.

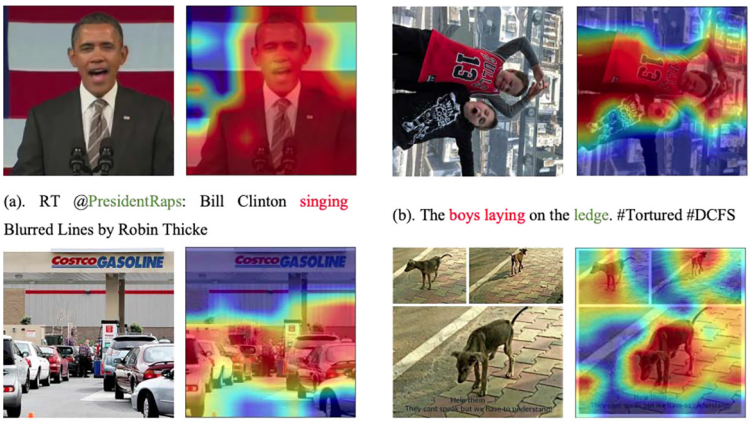


Fig. 4. Sample heatmap of MVSA dataset.

The heatmaps reveal that when key semantic features in the text are taken into account, critical regions of the image receive more attention. Based on the model's capture of the image's global structure, irrelevant information did not hinder attention to these crucial regions. This is primarily thanks to the model's understanding of the image's overall structure, and also due to our method's ability to, to some extent, mitigate the interference of potential noise in the image on the model's comprehension. For example, as shown in the first example of Fig. 4(a), when keywords like "PresidentRaps" and "singing" appear in the text, the model focuses more on the person's area in the image, making it easier to extract sentiment cues and enhance the understanding of user sentiment. When considering the potential sentiment information contained in "treat them" in Fig. 4(d), the "them" in the image was effectively attended to by the model, which is attributed to our model's comprehensive understanding of the image's global structure.

4.5.2. *Feature Distribution Analysis*

To more intuitively demonstrate the spatial distribution of features before and after processing by the Dynamic Frequency Domain Decoupling Module, we used t-SNE to reduce the dimensionality of features from the MVSA dataset and visualized the results as 2D scatter plots, as shown in Fig. 5.

By comparing the visualization effects of both scenarios, we can clearly observe that before processing by the Dynamic Frequency Domain Decoupling Module, features of the same class were relatively dispersed in space, with significant overlap between different classes. This sparsity in feature distribution and inter-class overlap made it difficult for the model to distinguish among categories, resulting in lower performance. However, after processing by the Dynamic Frequency Domain Decoupling Module, we observed that features of the same class tended to cluster spatially, and the overlap between classes decreased significantly. This optimized feature distribution provided the model with clearer class boundaries, enabling it to more easily identify and distinguish different sentiment categories, thereby significantly improving model performance in sentiment analysis tasks.

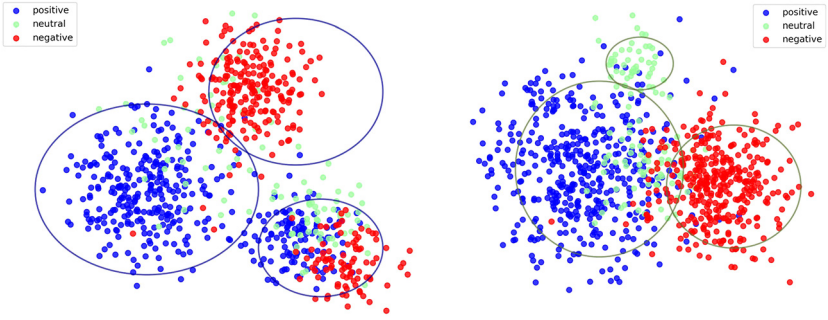


Fig. 5. The feature space distribution of the MVSA dataset before and after processing by the DSDD module.

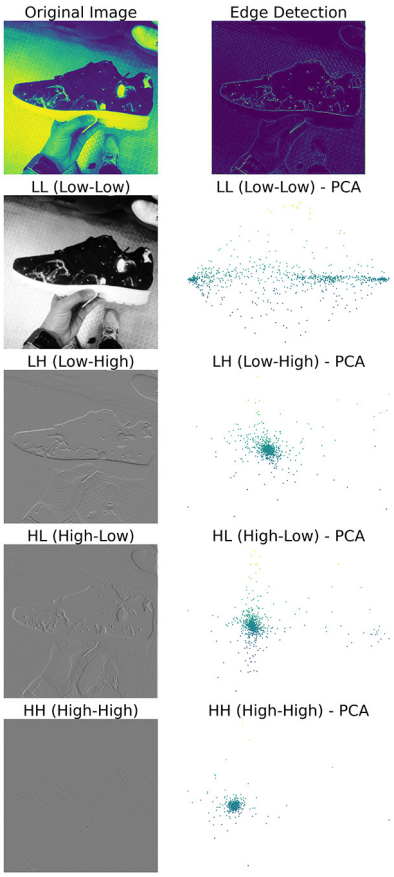


Fig. 6. PCA dimensionality reduction visualization of image frequency domain transformation process.

4.5.3. *Visual Analysis of Different Sub Bands*

In the frequency-domain analysis of an image, LF components carry the main content and global structural information, typically reflecting its overall outline, basic shapes, and general features. These LF components are crucial to an image's overall perception, as they convey its primary structure and visual characteristics. In the frequency domain, the LF region is often the key area that best represents the image's global structure. Therefore, by extracting LF components, we can intuitively obtain a global view of the image.

As shown in Fig. 6, as the frequency gradually increases, the HF components of the image begin to emerge. This HF information contains the detailed parts of the image, such as textures, edges, and minute structural features. The HF provides finer image details, making the contours and layers of these details clearer. However, an increase in frequency not only provides more detail but also introduces additional noise components, which typically manifest as chaotic HF variations in the image. As the frequency increases, the global structural information in the image gradually blurs, while details become more complex and contain more irrelevant noise. Therefore, in image processing, balancing the

Image	Text	Probability	Label
	View from #theshard only floor 32 and am feeling #dizzy lol	0.5249	negative
	Where theres a whisk, theres a way. #powerless #bakersgonnabake	0.8713	neutral
	@MrsRachelHudson (-) ella- Te quiero....- susurra mordindose el labio inferior mirndola a los ojos-	0.9958	positive

Fig. 7. Case study.

extraction of LF and HF information, especially effective noise suppression, is a critical issue in image analysis. Through frequency domain analysis, we can clearly distinguish the main structural components from noise in an image, thereby enhancing the model's sentiment analysis capabilities.

4.6. *Case Study*

To further explore the performance of our proposed model in sentiment analysis, we selected several representative cases for in-depth analysis, as shown in Fig. 7. This aims to comprehensively evaluate the model's robustness and accuracy. The analysis results indicate that when image information is unclear or somewhat ambiguous, the model struggles to determine sentiment polarity, primarily because it lacks the ability to extract and understand the global sentiment tone. For instance, in the first case, despite explicit sentiment cues such as "feeling dizzy lol" in the text, the model's confidence in its judgment was low due to the ambiguity of the image's sentiment.

In contrast, in the other two cases in the table, the model demonstrated higher confidence in its judgment. For example, in the third case, the prominent flowers in the image mostly convey positive sentiment. Concurrently, the model effectively suppresses noise information outside the target object in the image, preventing it from interfering with sentiment analysis. The results above indicate that our proposed model has strong capabilities for extracting sentiment information from both images and text, thereby enhancing its overall performance in sentiment analysis.

5. Concluding Remarks and Future Work

This work introduces the Frequency Domain Decoupling and Semantic Filtering Network (FDSF-Net) for multimodal sentiment analysis. The model addresses multimodal noise

through differentiated processing of image-frequency components and semantic filtering of text. Under the reported single-run setting, FDSF-Net achieves competitive results on the evaluated sentiment-analysis datasets and shows potential applicability to multimodal sarcasm detection. These results should not be interpreted as statistically significant differences because repeated-run variance estimates are not available.

However, our method has limitations. When the sentiment information in an image is unclear or ambiguous, the model struggles to determine sentiment polarity. Its discriminative performance is particularly limited when the global sentiment tone is indistinct or when there is significant interfering information in the image. This might be due to our model's insufficient handling of global sentiment tone information. Future work will focus on more refined image processing methods. To address the model's limitations when image sentiment is unclear, we plan to incorporate more advanced visual understanding techniques to improve its sensitivity to subtle sentiment expressions in images. Additionally, future efforts could involve introducing additional sentiment-related multimodal data and optimizing data preprocessing methods to further improve the model's robustness and accuracy across complex, varied scenarios.

References

An, J., Zainon, W.M.N.W. (2023). Integrating color cues to improve multimodal sentiment analysis in social media. *Engineering Applications of Artificial Intelligence*, 126, 106874.

Aziz, A., Chowdhury, N.K., Kabir, M.A., Chy, A.N., Siddique, Md.J. (2025). MMTF-DES: a fusion of multimodal transformer models for desire, emotion, and sentiment analysis of social media data. *Neurocomputing*, 623, 129376.

Ba, J.L., Kiros, J.R., Hinton, G.E. (2016). Layer Normalization. arXiv preprint arXiv:1607.06450.

Bibi, M., Abbasi, W.A., Aziz, W., Khalil, S., Uddin, M., Iwendi, C., Gadekallu, T.R. (2022). A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognition Letters*, 158, 80–86.

Cai, Y., Cai, H., Wan, X. (2019). Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2506–2515.

Chen, D., Su, W., Wu, P., Hua, B. (2023). Joint multimodal sentiment analysis based on information relevance. *Information Processing & Management*, 60(2), 103193.

Chen, Y. (2015). *Convolutional Neural Network for Sentence Classification*. Master's thesis, University of Waterloo.

Cheng, H., Yang, Z., Zhang, X., Yang, Y. (2023). Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion. *IEEE Transactions on Affective Computing*, 14(4), 3149–3163.

Dai, J., H., Yan, T., Sun, P., Liu, X., Qiu (2021). Does syntax matter? A strong baseline for aspect-based sentiment analysis with RoBERTa. arXiv preprint arXiv:2104.04986.

Das, R., Singh, T.D. (2023). Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s), 1–38.

Daubechies, I., Heil, C. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*, pp. 4171–4186.

Do, H.N., Phan, H.T., Nguyen, N.T. (2024). Multimodal sentiment analysis using deep learning and fuzzy logic: a comprehensive survey. *Applied Soft Computing*, 167, 112279.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. (2020). An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Feng, Z., Han, D., Li, J., Shi, S., Li, K.-C. (2026). Intrusion detection system for shipping communication networks based on federated distillation learning. *Expert Systems with Applications*, 299, 129966.

Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A., Ré, C. (2022). Hungry Hungry Hippos: Towards language modeling with state space models. arXiv preprint arXiv:2212.14052.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A. (2023). Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91, 424–444.

Ge, Y., Chen, Z., Yu, M., Yue, Q., You, R., Zhu, L. (2024). MambaTSR: you only need 90k parameters for traffic sign recognition. *Neurocomputing*, 599, 128104.

Gu, A., Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Gu, A., Gupta. A., Goel. K., Ré. C. (2022). On the parameterization and initialization of diagonal state space models. Advances in Neural Information Processing Systems 35.

Gu, A., Goel, K., Ré, C. (2021). Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Hu, R., Yi, J., Chen, A., Chen, L. (2024). Multichannel cross-modal fusion network for multimodal sentiment analysis considering language information enhancement. *IEEE Transactions on Industrial Informatics*, 20(7), 9814–9824.

Huan, R., Zhong, G., Chen, P., Liang, R. (2023). UniMF: a unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE Transactions on Multimedia*, 26, 5753–5768.

Huang, S.-C., Shen, L., Lungren, M.P., Yeung, S. (2021). GLoRIA: a multimodal global-local representation learning framework for label-efficient medical image recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3922–3931.

Kang, H., Yoo, S.J., Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5), 6000–6010.

Kim, K., Park, S. (2023). AOBERT: all-modalities-in-one BERT for multimodal sentiment analysis. *Information Fusion*, 92, 37–45.

Kim, W., Son, B., Kim, I. (2021). ViLT: vision-and-language transformer without convolution or region supervision. In: *International Conference on Machine Learning*. PMLR.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1746–1751.

Li, J., Han, D., Shi, S., Xin, X., Li, K.-C., Chang, C.-C. (2025a). An active client selection scheme based on blockchain for federated learning in shipping. *IEEE Transactions on Intelligent Transportation Systems*, 26(11), 20669–20684.

Li, J., Han, D., Weng, T.-H., Wu, H., Li, K.-C., Castiglione, A. (2025b). A secure data storage and sharing scheme for port supply chain based on blockchain and dynamic searchable encryption. *Computer Standards & Interfaces*, 91, 103887.

Li, Y., Liang, W., Xie, K., Zhang, D., Xie, S., Li, K. (2023). LightNestle: quick and accurate neural sequential tensor completion via meta learning. In: *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, pp. 1–10.

Li, Z., Xu, B., Zhu, C., Zhao, T. (2022). CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection. arXiv preprint arXiv:2204.05515.

Liang, C., Liang, W., Chen, Y., Li, K., Zomaya, A.Y. (2025). FedTCTF: tensor completion-based federated learning for device heterogeneity. *IEEE Transactions on Sustainable Computing*, 10(6), 1227–1239.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal loss for dense object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, Y. (2019). RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXivurl1907.11692.

Lu, Q., Sun, X., Long, Y., Gao, Z., Feng, J., Sun, T. (2024). Sentiment analysis: comprehensive reviews, recent advances, and open challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11), 15092–15112.

Mai, S., Zeng Y., Zheng S., Hu H. (2022). Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3), 2276–2289.

Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 674–693.

Moraes, R., Valiati, J.F., Neto, W.P.G. (2013). Document-level sentiment classification: an empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2), 621–633.

Niu, T., Zhu, S., Pang, L., El Saddik, A. (2016). Sentiment analysis on multi-view social data. In: *Proceedings of the 22nd International Conference on Multimedia Modeling*. Springer, Cham, pp. 15–27.

Pan, Z., Li, C., Plaza, A., Chanussot, J., Hong, D. (2025). Hyperspectral image classification with Mamba. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 5602814.

Pandey, A., Vishwakarma, D.K. (2024). Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: a survey. *Applied Soft Computing*, 152, 111206.

Parveen, N., Chakrabarti, P., Hung, B.T., Shaik, A. (2023). Twitter sentiment analysis using hybrid gated attention recurrent network. *Journal of Big Data*, 10(1), 50.

Pashchenko, Y., Rahman, M.F., Hossain, S., Uddin, K., Islam, T. (2022). Emotional and the normative aspects of customers' reviews. *Journal of Retailing and Consumer Services*, 68, 103011.

Rodríguez-Ibáñez, M., Casánez-Ventura, A., Castejón-Mateos, F., Cuenca-Jiménez, P.-M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223, 119862.

Shunxiang, Z., Aoqiang Z., Guangli Z., Zhongliang W., KuanChing L. (2023). Building fake review detection model based on sentiment intensity and PU learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10), 6926–6939.

Singh, U., Abhishek, K., Azad, H.K. (2024). A survey of cutting-edge multimodal sentiment analysis. *ACM Computing Surveys*, 56(9), 1–38.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y. (2024). RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing*, 568, 127063.

Sun, H., Li, W., Liu, J., Chen, H., Pei, R., Zou, X., Yan, Y., Yang, Y. (2024). CoSeR: bridging image and language for cognitive super-resolution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Sun, H., Niu, Z., Wang, H., Yu, X., Liu, J., Chen, Y.-W. (2025). Multimodal sentiment analysis with mutual Information-based disentangled representation learning. *IEEE Transactions on Affective Computing*, 16(3), 1606–1617.

Tai, K.S., Socher, R., Manning, C.D. (2015). Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.

Tang, D., Qin, B., Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., Wang, H., Wu, F. (2020). SKEP: sentiment knowledge enhanced pre-training for sentiment analysis. arXiv preprint arXiv:2005.05635.

Truong, Q.-T., Lauw, H.W. (2019). VistaNet: visual aspect attention network for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 305–312.

Usama, M., Ahmad, B., Song, E., Hossain, M.S., Alrashoud, M., Muhammad, G. (2020). Attention-based sentiment analysis using convolutional and recurrent neural network. *Future Generation Computer Systems*, 113, 571–578.

Wang, D., Liu, S., Wang, Q., Tian, Y., He, L., Gao, X. (2022). Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25, 4909–4921.

Wang, D., Tian, C., Liang, X., Zhao, L., He, L., Wang, Q. (2023a). Dual-perspective fusion network for aspect-based multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26, 4028–4038.

Wang, H., Du, Q., Xiang, Y. (2025a). Image–text sentiment analysis based on hierarchical interaction fusion and contrast learning enhanced. *Engineering Applications of Artificial Intelligence*, 146, 110262.

Wang, H., Ren, C., Yu, Z. (2025b). Multimodal sentiment analysis based on multiple attention. *Engineering Applications of Artificial Intelligence*, 140, 109731.

Wang, J., Yang, Y., Jiang, Y., Ma, M., Xie, Z., Li, T. (2024). Cross-modal incongruity aligning and collaborating for multi-modal sarcasm detection. *Information Fusion*, 103, 102132.

Wang, L., Niu, J., Yu, S. (2019). SentiDiff: combining textual information and sentiment diffusion patterns for Twitter sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 32(10), 2026–2039.

Wang, X., Wang, M., Cui, H., Zhang, Y. (2025c). Manifold knowledge-guided feature fusion network for multimodal sentiment analysis. *Expert Systems with Applications*, 280, 127537.

Wang, Y., Huang, G., Li, M., Li, Y., Zhang, X., Li, H. (2023b). Automatically constructing a fine-grained sentiment lexicon for sentiment analysis. *Cognitive Computation*, 15(1), 254–271.

Wang, Y., Jian, H., Zhuang, J., Guo, H., Leng, Y. (2025d). SSLMM: semi-supervised learning with missing modalities for multimodal sentiment analysis. *Information Fusion*, 120, 103058.

Wei, Y., Yuan, S., Yang, R., Shen, L., Li, Z., Wang, L., Chen, M. (2023). Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 5240–5252.

Wu, H., Kong, D., Wang, L., Li, D., Zhang, J., Han, Y. (2025). Multimodal sentiment analysis method based on image-text quantum transformer. *Neurocomputing*, 637, 130107.

Xie, X., Cui, Y., Tan, T., Zheng, X., Yu, Z. (2024). FusionMamba: dynamic feature enhancement for multimodal image fusion with Mamba. *Visual Intelligence*, 2(1), 37.

Xu, N. (2017). Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In: *Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI'17)*, pp. 152–154.

Xu, N., Mao, W. (2017). MultiSentiNet: a deep semantic network for multimodal sentiment analysis. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2399–2402.

Xu, N., Mao, W., Chen, G. (2018). A co-memory network for multimodal sentiment analysis. In: *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 929–932.

Xue, X., Zhang, C., Niu, Z., Wu, X. (2022). Multi-level attention map network for multimodal sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(5), 5105–5118.

Yang, X., Feng, S., Zhang, Y., Wang, D. (2021). 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*, pp. 328–339.

Ye, J., Zhang, J., Shan, H. (2025). DepMamba: progressive fusion mamba for multimodal depression detection. In: *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 1–5.

Zhai, G., Yang, Y., Wang, H., Du, S. (2020). Multi-attention fusion for sentiment analysis of educational big data. *Big Data Mining and Analytics*, 3(4), 311–319.

Zhang, B., Yuan, Z., Xu, H., Gao, K. (2024). Crossmodal translation based meta weight adaption for robust image-text sentiment analysis. *IEEE Transactions on Multimedia*, 26, 9949–9961.

Zhang, S., Liu, J., Jiao, Y., Zhang, Y., Chen, L., Li, K. (2025a). A multimodal semantic fusion network with cross-modal alignment for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(10), 1–22. https://doi.org/10.1145/3744648.

Zhang, S., Li, J., Li, S., Duan, W., Wei, Z., Li, K.-C. (2025b). Textual graph representation with syntactic weighting for implicit sentiment analysis. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(6), 4288–4299. https://doi.org/10.1109/TETCI.2025.3550515.

Zhang, Y., Wang, L., Wang, H. (2023). Cross-modal sentiment analysis based on transformer and image-text collaborative interaction. In: *2023 International Conference on Computer Engineering and Distance Learning (CEDL)*. IEEE.

Zhao, S., Yao X., Yang J., Jia G., Ding G., Chua T.-S., Schuller B.W., Keutzer K. (2021). Affective image content analysis: two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 6729–6757.

Zhao, X., Poria, S., Li, X., Chen, Y., Tang, B. (2025a). Toward robust multimodal sentiment analysis using multimodal foundational models. *Expert Systems with Applications*, 276, 126974.

Zhao, Y., Li, K.-C., Zhang, S., Ye, T. (2025b). Hyperbolic graph attention network fusing long-context for technical keyphrase extraction. *Information Fusion*, 120, 103061. https://doi.org/10.1016/j.inffus.2025.103061.

Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Heng Chua, M.C., Liu, M. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management*, 56(6), 102097.

Zhou, P., Shi, W., Tian, J., Qi Z., Li B., Hao H., Xu B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 207–212.

Zhou, S., Chen, Y., Liang, W., Li, K., Zomaya, A.Y. (2026a). MultiSecDFL: multi-metric filtering-based secure aggregation method for decentralized federated learning. *IEEE Transactions on Consumer Electronics*. https://doi.org/10.1109/TCE.2026.3688316.

Zhou, S., Chen, Y., Liang, W., Li, K., Meng, W. (2026b). TrustHFL: an efficient aggregation method for trustworthy hierarchical federated learning. *IEEE Internet of Things Journal*, 13(11), 24618–24631.

Zhou, S., Li, K., Chen, Y., Yang, C., Liang, W., Zomaya, A.Y. (2024). TrustBCFL: mitigating data bias in IoT through blockchain-enabled federated learning. *IEEE Internet of Things Journal*, 11(15), 25648–25662.

Zhu, L., Zhu, Z., Zhang, C., Xu, Y., Kong, X. (2023). Multimodal sentiment analysis based on fusion methods: a survey. *Information Fusion*, 95, 306–325.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X. (2024). Vision Mamba: efficient visual representation learning with bidirectional state space model. In: *International Conference on Machine Learning*.

Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., Qian, J. (2022). Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia*, 25, 3375–3385.

Y. Liu received the master of science degree in educational technology from East China Normal University, China. He is currently an associate professor in the Computer Science and Technology program at Suzhou University, China. His research interests include image recognition and artificial intelligence.

C. Tan received the master of science degree in educational technology from Nanjing Normal University, China. She is currently an associate professor in the Computer Science and Technology program at Suzhou University, China. Her research interests include image recognition and artificial intelligence.

S. Zhang received the PhD degree from the School of Computer Engineering and Science, Shanghai University, Shanghai, China, in 2012. He is currently a professor with the School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China. His research interests include intelligent information processing, data mining, big data analytics, and sentiment analysis.

Y. Zhang received the master of science degree from Anhui University of Science and Technology, Huainan, China. He is currently with the School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China. His research interests include multimodal sentiment analysis, multimodal sarcasm detection, deep learning, and multimodal information fusion.

K.-C. Li received the PhD degree in electrical engineering from the University of São Paulo, Brazil. He is currently a distinguished professor with the Department of Computer Science and Information Engineering, Providence University, Taiwan. He is also affiliated with the School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, China. His research interests include cloud computing, GPU computing, big data, and parallel programming.

R. Crespo received the PhD degree in computer science engineering from Universidad Pontificia de Salamanca, Spain. He is currently a vice-rector and full professor of Computer Science and Artificial Intelligence at Universidad Internacional de La Rioja (UNIR), Spain. His research interests include artificial intelligence, Industry 4.0, project management, and accessibility.