Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.072286

ARTICLE

A Multimodal Sentiment Analysis Method Based on Multi-Granularity Guided Fusion

Zilin Zhang[1], Yan Liu[1,*], Jia Liu[2], Senbao Hou[3], Yuping Zhang[1] and Chenyuan Wang[1]

[1]Henan Key Laboratory of Cyberspace Situation Awareness, Key Laboratory of Cyberspace Security, Ministry of Education, Information Engineering University, Zhengzhou, 450001, China
[2]State Key Laboratory of Mathematical Engineering and Advanced Computing, Information Engineering University, Zhengzhou, 450001, China
[3]Henan Key Laboratory of Imaging and Intelligent Processing, Information Engineering University, Zhengzhou, 450001, China

*Corresponding Author: Yan Liu. Email: ms.liuyan@foxmail.com

Received: 23 August 2025; Accepted: 29 September 2025; Published: 09 December 2025

ABSTRACT: With the growing demand for more comprehensive and nuanced sentiment understanding, Multimodal Sentiment Analysis (MSA) has gained significant traction in recent years and continues to attract widespread attention in the academic community. Despite notable advances, existing approaches still face critical challenges in both information modeling and modality fusion. On one hand, many current methods rely heavily on encoders to extract global features from each modality, which limits their ability to capture latent fine-grained emotional cues within modalities. On the other hand, prevailing fusion strategies often lack mechanisms to model semantic discrepancies across modalities and to adaptively regulate modality interactions. To address these limitations, we propose a novel framework for MSA, termed Multi-Granularity Guided Fusion (MGGF). The proposed framework consists of three core components: (i) Multi-Granularity Feature Extraction Module, which simultaneously captures both global and local emotional features within each modality, and integrates them to construct richer intra-modal representations; (ii) Cross-Modal Guidance Learning Module (CMGL), which introduces a cross-modal scoring mechanism to quantify the divergence and complementarity between modalities. These scores are then used as guiding signals to enable the fusion strategy to adaptively respond to scenarios of modality agreement or conflict; (iii) Cross-Modal Fusion Module (CMF), which learns the semantic dependencies among modalities and facilitates deep-level emotional feature interaction, thereby enhancing sentiment prediction with complementary information. We evaluate MGGF on two benchmark datasets: MVSA-Single and MVSA-Multiple. Experimental results demonstrate that MGGF outperforms the current state-of-the-art model CLMLF on MVSA-Single by achieving a 2.32% improvement in F1 score. On MVSA-Multiple, it surpasses MGNNS with a 0.26% increase in accuracy. These results substantiate the effectiveness of MGGF in addressing two major limitations of existing methods—insufficient intra-modal fine-grained sentiment modeling and inadequate cross-modal semantic fusion.

KEYWORDS: Multimodal sentiment analysis; cross-modal fusion; cross-modal guided learning

1 Introduction

With the rapid advancement of social media, smart devices, and multimodal sensing technologies, users are increasingly generating vast amounts of multimodal data—such as text, images, and audio—during daily communication. These data not only carry rich semantic content but also deeply reflect individual emotional expressions, thereby offering critical support for building more natural and emotionally intelligent

human-computer interaction systems. Against this backdrop, Multimodal Sentiment Analysis (MSA) [1] has emerged as a pivotal task in the field of affective computing and has attracted growing interest across interdisciplinary domains, including Natural Language Processing (NLP), Computer Vision (CV), and Artificial Intelligence (AI). MSA aims to effectively integrate information from heterogeneous modalities to accurately recognize users' emotional states, enhancing the system's capability to perceive, understand, and respond to human affect.

Despite the emergence of numerous MSA methods and their impressive results on various benchmark tasks, there remain two critical limitations in existing research. First, most approaches heavily rely on modeling global features of each modality while overlooking the rich local structural information embedded within modalities. For example, the method proposed by Tsai et al. [2] focuses primarily on global image features and neglects the extraction of fine-grained emotional cues from local image regions. Similarly, although Zong et al. [3] achieve improvements in cross-modal collaborative modeling, their method is still based on holistic modal embeddings, lacking effective representation of spatial and temporal local structures. Second, existing fusion strategies are often restricted to token-level static alignment and weighted concatenation, lacking the capability to perceive semantic discrepancies across modalities or to adaptively regulate their integration. For instance, the word-level fusion method by Chen et al. [4] performs reasonably well in modality alignment but remains constrained by token-based concatenation strategies and fails to model local temporal dynamics and fine-grained cross-modal dependencies. Overall, most mainstream MSA methods tend to ignore intra-modal local emotional cues, which often arise from localized token combinations—such as facial expressions or hand gestures in images—and carry crucial emotional information. Therefore, how to effectively extract local features and achieve cooperative modeling between global and local representations remains a major challenge in multimodal sentiment analysis.

To address the above challenges, we propose a novel framework for multimodal sentiment analysis, termed Multi-Granularity Guided Fusion (MGGF). The framework is composed of three key modules: (i) Multi-Granularity Feature Extraction Module: This module jointly extracts both global and local emotional features from each modality and performs intra-modal fusion of multi-granularity information to construct richer unimodal representations. (ii) Cross-Modal Guidance Learning Module (CMGL): A novel mechanism based on Softmax operations and Kullback-Leibler (KL) divergence is introduced to measure the distributional divergence and complementarity between image and text modalities. This information serves as a guidance signal to inform subsequent fusion strategies, especially in contexts of modality agreement or conflict. (iii) Cross-Modal Fusion Module (CMF): We design a bidirectional interactive attention mechanism to capture fine-grained semantic dependencies between modalities. This allows for deeper emotional information exchange and contributes to the construction of more discriminative fused representations. The main contributions of this paper are summarized as follows:

- We propose a multi-granularity guided fusion framework for multimodal sentiment analysis, which jointly models global and local features from both image and text modalities and achieves coordinated intra- and inter-modal information fusion.
- We design a CMGL module that introduces a feature distribution divergence modeling method based on Softmax and KL divergence. This module effectively quantifies the distributional bias between modalities and guides the fusion process in both modality-consistent and modality-conflicting scenarios.
- We propose a CMF with a novel bidirectional semantic interaction attention mechanism, capable of capturing fine-grained semantic dependencies across modalities and improving the sentiment discriminative power of the fused representation.
- We conduct comprehensive experiments on two widely-used MSA datasets—MVSA-Single and MVSA-Multiple. Results show that MGGF surpasses the current state-of-the-art model CLMLF on

MVSA-Single by 2.32% in F1 score, and outperforms MGNNS on MVSA-Multiple with a 0.26% improvement in accuracy. These findings validate the effectiveness and potential of MGGF in multimodal sentiment analysis tasks.

2 Related Work on Multimodal Sentiment Analysis

Early research on multimodal sentiment analysis predominantly focused on two primary fusion strategies: early fusion and late fusion. Early fusion approaches typically involve the direct concatenation of feature representations from different modalities (e.g., visual and textual) at the feature level. While such methods are relatively simple to implement and can retain a substantial amount of raw information, they often fail to model the intricate and subtle semantic correlations between modalities adequately. This limitation can lead to the inclusion of redundant or noisy features, thereby compromising the overall performance of sentiment recognition. In contrast, late fusion methods perform sentiment classification independently within each modality and subsequently integrate the outputs at the decision level.

Although this strategy can enhance the accuracy of intra-modal modelling to a certain extent, the lack of a collaborative cross-modal learning mechanism makes it difficult to capture deep inter-modal interactions, ultimately constraining further performance gains. To address these limitations, recent research has increasingly shifted toward intermediate fusion strategies. These approaches aim to introduce inter-modal interaction mechanisms during the process of deep feature representation learning, enabling a more fine-grained and dynamic modelling of cross-modal sentiment representations. Recent work by Wang et al. [5] demonstrates that cross-modal hierarchical fusion with multi-task learning achieves superior performance compared with existing models on CH-SIMS, CMU-MOSI, and CMU-MOSEI datasets. Li et al. [6] proposed a Fine-grained Multimodal Fusion Network (FMFN), which integrates learnable denoising tokens, token-level cross-modal alignment, and correlation-aware fusion to improve the performance of multimodal sentiment analysis. Yadav and Vishwakarma [7] proposed DMLANet, which integrates dual attention across image channels and spatial dimensions, alongside semantic and self-attention networks for fine-grained fusion of image-text emotional features. Zhu et al. [8] introduced ITIN, aligning region-word pairs and using gating mechanisms to achieve fine-grained cross-modal sentiment modelling by integrating visual and textual contexts. Huang et al. [9] developed TeFNA, a text-centric fusion framework incorporating cross-modal attention to address alignment and fusion challenges in MSA. Wang et al. [10] proposed TETFN, which employs a text-guided multi-head attention and cross-modal mapping structure, emphasising the dominant role of text to improve both modality consistency and semantic diversity. By enabling deeper interaction between modalities, intermediate fusion achieves a balance between representation richness and alignment precision, making it a powerful alternative to early and late fusion approaches in complex sentiment analysis tasks.

3 Method

3.1 Problem Definition

MSA primarily aims to comprehensively utilize the information embedded in different modalities to predict the sentiment polarity or intensity of the input samples. Formally, let us define a multimodal dataset $D = \{X, Y\}$, where each sample $(x, y) \in D$ consists of multimodal input data and the corresponding sentiment label. This can be denoted as $(x, y) = \{U_m, y\}$, where U_m represents the sequential information of the m-th modality, and y denotes the sentiment label. In general scenarios, the modality set can include text, images, videos, or audio.

The core research problem addressed in this paper is how to effectively leverage the distributional differences between any two modalities to guide the fusion strategy. More specifically, for a given pair of modality sequences U_1 and U_2, we first extract both global and local features, denoted as (x_1^g, x_1^l) and (x_2^g, x_2^l). These features are then fused to obtain unimodal representations X_1 and X_2. Subsequently, by quantifying the distributional discrepancy between these two modality-specific representations, we derive a difference factor α, which serves as a dynamic weighting signal to modulate the cross-modal fusion process. This mechanism allows us to preserve more unimodal-specific features when the modalities are consistent, while enhancing cross-modal interactive features in the case of conflicting modalities. Finally, the fused representation and the unimodal representations are jointly input into a classifier through a hierarchical fusion operation to predict sentiment polarity or intensity.

In this study, we specifically focus on the task of sentiment analysis involving text and image modalities. For ease of notation, we denote the textual sequence as U_t and the image sequence as U_v.

3.2 Overview

This paper proposes a MGGF, designed to enhance the modeling capability of complex cross-modal interactions in sentiment analysis tasks. MGGF strengthens adaptive perception of modality discrepancies and enhances expressive representation by guiding the fusion of global and local features from textual and visual modalities. As illustrated in Fig. 1, the overall framework is composed of four key modules: (i) Multi-granularity Feature Extraction Module: This module employs pretrained modality encoders to obtain both global and local representations for each modality, thereby capturing comprehensive semantic information as well as fine-grained contextual dependencies. (ii) CMGL: By quantifying distributional discrepancies between unimodal representations, this module measures the heterogeneity between textual and visual modalities. The resulting cross-modal discrepancy score is then used to dynamically regulate subsequent fusion strategies. (iii) CMF: Based on the extracted multi-granularity features within each modality, this module introduces a cross-modal semantic interaction mechanism. It guides the fusion of unimodal vectors with cross-modal complementary features, thereby adapting to varying degrees of modality discrepancies and constructing a unified multimodal representation. (iv) Classifier: Finally, the fused multimodal representations and unimodal features are jointly fed into a classifier. Under the guidance of the cross-modal discrepancy score, the classifier performs the final prediction of sentiment polarity or sentiment intensity.

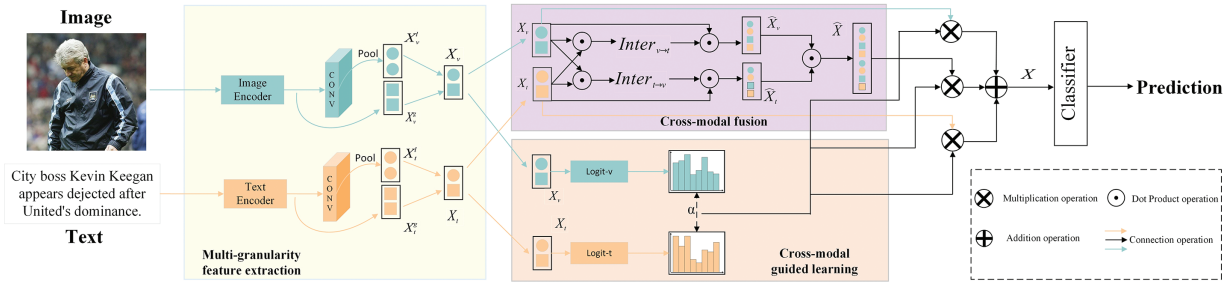

Figure 1: The overview of our proposed aspect-oriented model MGGM

3.3 Multi-Granularity Feature Extraction

The module extracts global and local semantics from text and images using pretrained Transformer encoders (BERT, ViT). As extraction is not the focus, standard models are used. Global and local features of modality $m \in \{t, v\}$ are denoted m as X_m^g and X_m^l.

3.3.1 Global Feature Extraction

To capture overall semantics, we model global features by taking the [*CLS*] token (denoted as CLS^{-1}) from the final encoder layer of modality m, denoted X_m^g, and projecting it via a feedforward layer into a unified feature space.

$$X_m^g = FF(CLS^{-1}(U_m)), \quad m \in \{t, v\} \tag{1}$$

where U_m denotes the original input sequence of modality m.

3.3.2 Local Feature Extraction

To preserve fine-grained semantics, local features are extracted from all BERT tokens and ViT patches ($BERT^{-1}$, ViT^{-1}). These embeddings are refined via Conv1D and aggregated with AdaMaxPool1d, yielding the local representations:

$$X_t^l = AdaMaxPool1d(Conv1D(BERT^{-1})), \quad X_v^l = AdaMaxPool1d(Conv1D(ViT^{-1})) \tag{2}$$

To obtain rich modality-specific features, global and local representations of modality m are fused via nonlinear projection to form the unimodal representation:

$$X_m = MLP(CONCAT(X_m^g, X_m^l)), \quad m \in \{t, v\} \tag{3}$$

where X_m^g and X_m^l denote the global and local feature representations of modality m, respectively.

3.3.3 Intra-Modal Features Fusion

To enrich semantics, global X_m^g and local feature X_m^l features of modality m are concatenated and fused via an MLP, producing the unified unimodal representation X_m^l:

$$X_m = MLP(CONCAT(X_m^g, X_m^l)), \quad m \in \{t, v\} \tag{4}$$

where X_m^g and X_m^l denote the global and local representations of modality m, respectively, and $m \in \{t, v\}$ corresponds to the textual and visual modalities.

3.4 Cross-Modal Guidance Learning

To measure representational differences between text and image, features X_t and X_v are projected into probability distributions P_t and P_v via a Softmax layer:

$$P_t = Softmax(X_t), \quad P_v = Softmax(X_v) \tag{5}$$

The average KL divergence between P_t and P_v is used to quantify modality discrepancy, defined as:

$$a_1 = KL(p \parallel q) = \frac{1}{n}\sum_{i=1}^{n} p_v \log \frac{p_v}{p_t}, \quad a_2 = KL(q \parallel p) = \frac{1}{n}\sum_{i=1}^{n} p_t \log \frac{p_t}{p_v} \tag{6}$$

where n denotes the sample size, and p_v and p_t represent the probability distributions of visual and textual modalities, respectively.

To bound the discrepancy score in $[0,1]$ and improve learnability, the averaged KL divergence is passed through a Sigmoid function, producing the final cross-modal score α:

$$a = sigmoid\left(\frac{1}{2}(a_1 + a_2)\right) \tag{7}$$

The score α indicates modality divergence: lower values imply aligned representations, higher values reflect greater discrepancies. During inference, α serves as a dynamic weight—emphasizing cross-modal interaction when divergence is high, and preserving unimodal features when low—enabling adaptive semantic representation.

3.5 Cross-Modal Fusion

CMF captures semantic interactions between modalities, enhancing multimodal sentiment analysis, especially under modality inconsistencies. It takes X_t and X_v as input and computes a cross-modal attention matrix $InterW$ to model semantic correlations:

$$InterW_{v\to t} = softmax\left(\frac{[X_t][X_v]^T}{\sqrt{dim}}\right), \quad InterW_{t\to v} = softmax\left(\frac{[X_v][X_t]^T}{\sqrt{dim}}\right) \tag{8}$$

where dim denotes the feature dimension, used to scale the dot product for training stability. The **softmax** function ensures that the attention weights are normalized.

Using the weight matrices, cross-modal features are constructed to enable dynamic semantic transfer and fusion across modalities, yielding the interaction representations:

$$\hat{X}_t = InterW_{v\to t} \times X_t, \quad \hat{X}_v = InterW_{t\to v} \times X_v \tag{9}$$

To further capture more complex interactions between the two modalities, we apply an outer product operation between $\hat{X}_t$ and $\hat{X}_v$, forming the final fusion representation:

$$\hat{X} = \hat{X}_t \otimes \hat{X}_v \tag{10}$$

here, $\otimes$ denotes the outer product, which explicitly encodes bilinear relationships between modalities.

3.6 Classifier

In the classification stage, the model constructs its input via adaptive hierarchical fusion of unimodal features—derived from intra-modal fusion of global and local representations—and cross-modal features generated by the CMF, which capture semantic interactions between modalities:

$$X = (a \times \hat{X}) \oplus ((1-a) \times X_t) \oplus ((1-a) \times X_v) \tag{11}$$

where X_m represents features from the visual modality, X_t represents features from the textual modality, and a denotes the cross-modal discrepancy score computed by the cross-modal guidance module, which quantifies the degree of semantic divergence between modalities. The symbol $\oplus$ refers to hierarchical fusion operations.

This design adaptively balances feature contributions: high cross-modal discrepancy boosts reliance on cross-modal features to resolve conflicts, while low discrepancy favors unimodal features to reduce fusion noise. The final fused representation X is then passed to an MLP for sentiment prediction:

$$\hat{y} = MLP(X) \tag{12}$$

3.7 Optimization Object

We adopt a joint optimization strategy combining regression and classification losses to improve both fine-grained sensitivity and coarse-grained accuracy. The total loss is:

$$L = L_{reg} + L_{cls} \tag{13}$$

here, L_{reg} uses L_1 loss (MAE) to capture subtle sentiment variations and resist outliers:

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{14}$$

L_{cls} applies cross-entropy to enhance class separability and discriminative power:

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N} -y_i \log(\hat{y}_i) \tag{15}$$

This multi-task approach improves robustness and generalization by jointly learning fine- and coarse-grained sentiment signals.

4 Experiment

4.1 Set up

4.1.1 Datasets

In our study, to ensure the validity and reliability of experimental results, we adopt two widely used benchmark datasets for MSA: MVSA-Single [11] and MVSA-Multiple. Both datasets consist of image-text pairs, with each pair annotated with three sentiment labels: Positive, Neutral, and Negative. The detailed statistics are provided in Table 1. The two datasets differ primarily in their annotation mechanisms: MVSA-Single is annotated by a single annotator for each image-text pair, whereas MVSA-Multiple is annotated independently by three annotators, with no mutual influence among their judgments. To guarantee the consistency of sentiment annotations across modalities and ensure the effectiveness of fused representations, we performed cleaning and preprocessing on both datasets.

Table 1: Sentiment label distribution in MVSA-Single and MVSA-Multiple

Datasets	Positive	Neutral	Negative	Total
MVSA-Single	2680	466	1365	4511
MVSA-Multiple	11,445	4185	1394	17,024

Specifically, for the MVSA-Single dataset, we first removed samples with sentiment conflicts between the image and text modalities (e.g., where the image expresses positive sentiment while the text conveys negative sentiment). Next, for samples labeled as "Neutral," we retained only those where the overall tendency was non-neutral, assigning the dominant sentiment as the final label. After filtering, a total of 4511 sentiment-consistent image-text pairs were preserved. For the MVSA-Multiple dataset, we followed the same procedure to remove sentiment-conflicted samples. Then, a majority voting mechanism was applied: if at least two out of three annotators agreed on the sentiment label of a sample, that label was adopted as the final annotation; otherwise, the sample was discarded. Ultimately, we obtained 17,024 high-consistency labeled image-text pairs.

To mitigate potential training instability caused by distributional imbalance or domain bias across datasets, we partitioned each dataset into training, validation, and testing subsets with a ratio of 8:1:1. This ensures the stability and generalization capability of the model during training, hyperparameter tuning, and evaluation. The sample distributions across subsets are reported in Table 2.

Table 2: Dataset split statistics for MVSA-Single and MVSA-Multiple

Datasets	Training	Testing	Validation	Total
MVSA-Single	3609	451	451	4511
MVSA-Multiple	13,620	1702	1702	17,024

4.1.2 Evaluation Guidelines

In our experiments, to comprehensively evaluate the model's performance and practical applicability in multimodal sentiment analysis (MSA) tasks, we adopt two commonly used but essential evaluation metrics: (1) Accuracy (Acc): representing the overall correctness of sentiment analysis, used to measure the model's ability to make accurate predictions across all samples. (2) F1-score (F1): which considers both Precision and Recall in classification tasks, reflecting the balance of performance across different sentiment categories. Together, these metrics provide a well-rounded evaluation of the model's multimodal sentiment analysis capability. By leveraging both benchmark datasets, we aim to achieve a more precise and comprehensive understanding of the model's effectiveness in MSA tasks.

4.1.3 Implementation Details

In this study, we employ BERT and ViT as the feature extractors for textual and visual modalities, respectively, to capture semantic information from multimodal inputs. The image input size is standardized to 224 × 224, and the hidden layer dimension is set to 768. Model training is performed using the Adam optimizer with an initial learning rate of 2e–5. All experiments are conducted under the PyTorch 2.1.0 framework, with the runtime environment configured on Ubuntu 22.04, utilizing CUDA 12.1 and an NVIDIA GeForce RTX 4090 GPU to accelerate training. The batch size is set to 128, and the training process runs for 20 epochs. Under this configuration, the final model achieves its optimal performance.

4.1.4 Baseline Comparisons

To evaluate the effectiveness of the proposed MGGF model, we compare it with several representative and widely used baselines in multimodal sentiment analysis:

- Co-MemNet [12]: A co-memory attention model using iterative memory hops to establish semantic associations—text guides image region localisation, while images help extract sentiment keywords from text.
- MVAN [13]: Proposes a multi-view attention network trained on the TumEmo dataset, extracting global, object, and scene-level image features, and jointly modelling them with text via multimodal attention.
- MGNNS [14]: A graph neural network model that fuses scene and object information from images with latent textual sentiment, achieving deep-level cross-modal feature interaction and aggregation.
- CLMLF [15]: Employs Transformer-based cross-layer fusion to align textual and visual features, and introduces a contrastive learning task guided by labels and data to capture sentiment-related commonalities in multimodal inputs.

- MVCN [16]: proposed a Multi-View Calibration Network, which addresses the challenges of modality fusion, feature misalignment, and label inconsistency. To this end, the framework sequentially introduces a text-guided fusion module, a feature constraint task based on sentiment consistency, and an adaptive loss calibration strategy.
- MFGFN [17]: proposed a Multi-Granularity Feature Gated Fusion Network that integrates fine-grained unimodal features extracted by BERT and ResNet with coarse-grained multimodal features obtained via CLIP. By introducing a co-attention encoder and a gated fusion mechanism, the model effectively enables cross-modal and multi-granularity information interaction, while adaptively assigning feature weights to enhance fusion performance.

4.2 Experimental Results

To assess the effectiveness of the proposed MGGF model, we compare it against several state-of-the-art multimodal sentiment analysis methods on two widely used benchmarks: MVSA-Single and MVSA-Multiple. As summarised in Table 3, MGGF consistently outperforms baseline models, demonstrating superior multimodal sentiment discrimination. On MVSA-Single, MGGF exceeds the strongest baseline, CLMLF, by 1.25% in accuracy and 2.32% in F1-score. On MVSA-Multiple, it surpasses MGNNS and MVAN, achieving improvements of 0.63% in accuracy and 0.26% in F1-score.

Table 3: Performance comparison of different methods on MVSA-Single and MVSA-Multiple

Methods	MVSA-Single		MVSA-Multiple	
	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)
Co-MemNet	70.51	70.01	69.92	69.83
MVAN	72.98	72.98	72.36	72.30
MGNNS	73.77	72.70	72.49	69.34
CLMLF	75.33	73.46	72.00	69.83
MFGFN	76.22	75.38	70.82	69.94
MGGF (ours)	**76.58**	**75.78**	**73.12**	**72.56**

CLMLF integrates multi-layer features and adopts contrastive learning to enhance cross-modal alignment; however, it fails to handle modality conflicts and neglects local emotional cues, resulting in decreased accuracy. MGNNS, while based on graph neural networks, suffers from a static graph structure that lacks the capacity for dynamic adjustment and does not capture fine-grained information, leading to unstable performance. ITIN employs a region-word alignment strategy for cross-modal interaction; however, its rigid alignment mechanism introduces noise under modality inconsistency, thereby undermining recognition accuracy. DMLANet prioritizes image-side modeling while overlooking textual representations, which limits its effectiveness on text-dominant samples. TeFNA, by adopting a fixed text-centered fusion paradigm, lacks adaptability, especially in image-dominant or modality-conflicted scenarios. In contrast, the proposed MGGF combines global and local emotional features, enhancing intra-modal representational capacity. It further introduces a KL divergence-based discrepancy scoring mechanism to regulate the cross-modal fusion strategy dynamically. This design enables MGGF to effectively adapt to both modality-consistent and modality-conflicting scenarios, thereby significantly improving both accuracy and robustness, and addressing the structural limitations observed in existing approaches. These gains confirm MGGF's ability to effectively capture high-level semantic interactions between text and image modalities, enhancing both accuracy and generalisation.

4.3 Ablation Study

To further investigate the effectiveness of each component in MGGF, we conduct three sets of experiments.

4.3.1 Effectiveness of Each Component

We conducted a comprehensive ablation study on the MVSA-Single and MVSA-Multiple datasets to evaluate the individual contributions of each component within the MGGF model. The detailed experimental settings and results are summarised in Table 4, showing performance changes as key modules are progressively removed.

Table 4: Ablation study on the architecture design of MGGF on two datasets

ID	Module Settings				MVSA-Single		MVSA-Multiple	
	Global	Local	CMGL	CMF	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)
1	✓		✓	✓	76.28	74.83	72.77	72.06
2		✓	✓	✓	74.67	74.22	71.94	71.48
3	✓	✓		✓	75.13	74.79	71.96	71.73
4	✓	✓	✓		74.75	73.96	72.27	71.82
5	✓	✓	✓	✓	**76.58**	**75.78**	**73.12**	**72.56**

In the first set of experiments, we only utilized the global features of each modality, excluding the local features. The results show that removing local features led to performance degradation: specifically, on the MVSA-Single dataset, the metrics decreased by 0.3% and 0.95%, while on the MVSA-Multiple dataset, the drops were 0.35% and 0.5%, respectively. These findings indicate that although global features serve as the primary carriers of information, local features still provide valuable complementary cues that contribute positively to fine-grained sentiment representation.

In the second set of experiments, we only retained the local features of each modality. The results revealed a more severe performance drop compared to the first set of experiments. Specifically, on the MVSA-Single dataset, the performance decreased by 1.91% and 0.95%, while on the MVSA-Multiple dataset, the declines were 1.18% and 1.08%, respectively. These findings suggest that global features play a dominant role in capturing overall semantic information, whereas local features primarily serve as complementary cues to enrich fine-grained sentiment expressions.

In the third set of experiments, we removed the CMGL, treating unimodal and cross-modal features as equally important. The results show that, on the MVSA-Single dataset, performance dropped by 1.45% and 0.99%, while on the MVSA-Multiple dataset, it decreased by 1.16% and 0.83%. This indicates that unimodal and cross-modal features contribute differently in terms of expressive content, and the KL divergence-based mechanism for quantifying distributional discrepancies is indispensable in guiding the fusion process.

In the fourth set of experiments, we removed the original CMF and instead adopted a simple concatenation strategy to combine features from both modalities. The results show that, on the MVSA-Single dataset, the performance dropped by 1.8% and 1.82%, while on the MVSA-Multiple dataset, the decreases were 0.85% and 0.74%. These outcomes demonstrate that the fusion features constructed through the cross-modal interaction mechanism capture semantic dependencies and complementary relationships more effectively than simple concatenation operations.

4.3.2 Cross-Modal Guided Learning Analysis

To evaluate the impact of different distance measurement methods on model performance, we designed two variants of the MGGF framework, each employing a different similarity calculation strategy to model the relationship between textual and visual features. Specifically, MGGF-COS utilizes cosine similarity as the distance function, while MGGF-DIS is based on Euclidean distance. Both approaches estimate distances directly in the feature space without involving probability modeling. As shown in Table 5, all three variants achieve competitive performance on multimodal sentiment analysis tasks, further validating the effectiveness of the cross-modal guidance mechanism in sentiment representation learning. Notably, MGGF-DL outperforms both MGGF-COS and MGGF-DIS. The superiority of MGGF-DL can be attributed to the use of the KL divergence mechanism, which models the discrepancies between modalities in the probability distribution space. This enables the framework to effectively capture the uncertainty and fine-grained distributional differences of cross-modal features. In contrast, although MGGF-COS and MGGF-DIS demonstrate efficiency and directness in feature space computation, their reliance solely on fixed vector similarity metrics limits their ability to model distributional information. As a result, they struggle to fully reflect the nuanced modality discrepancies underlying sentiment expressions.

Table 5: Performance comparison of different distance measurement methods in guided learning methods

Methods	MVSA-Single		MVSA-Multiple	
	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)
MGGF-COS	75.37	75.08	72.77	72.16
MGGF-DIS	74.66	73.94	72.45	71.92
MGGF-KL	**76.58**	**75.78**	**73.12**	**72.56**

4.3.3 Cross-Modal Fusion Methods

To assess the impact of different cross-modal fusion strategies on model performance, we designed two variants of the MGGF framework, each replacing the original cross-modal fusion module with an alternative mechanism: 1) MGGF-CAT: directly concatenates unimodal features after fusing global and local representations. 2) MGGF-CNN: introduces convolutional neural networks (CNNs) to fuse multimodal features, capturing local dependencies between modalities through receptive fields. The experimental results, presented in Table 6, demonstrate that both alternative approaches consistently underperform compared to the original MGGF model, further confirming the effectiveness and necessity of the proposed CMF. Specifically, MGGF-CAT shows significant performance degradation, suggesting that simple concatenation of multimodal features fails to establish deep semantic interactions, thereby limiting the discriminative power of fused representations. Although MGGF-CNN leverages convolutional kernels to model local interactions between modalities, its limited receptive field restricts the capture of broader cross-domain dependencies, resulting in suboptimal fusion outcomes.

Table 6: Performance comparison between different crossmodal fusion methods

Methods	MVSA-Single		MVSA-Multiple	
	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)
MGGF-CAT	74.75	73.96	72.27	71.82
MGGF-CNN	75.48	74.69	72.74	72.31
MGGF	**76.58**	**75.78**	**73.12**	**72.56**

4.3.4 Complexity Analysis

We evaluate the computational overhead and inference efficiency across its key components. The majority of the computational cost arises from the modality-specific encoders and the cross-modal interaction modules. Specifically, BERT and ViT are employed as the text and image encoders, respectively, both exhibiting a complexity of $\mathcal{O}(L^2 d)$, where L denotes the input sequence length and d represents the hidden dimension. Local feature enhancement is implemented via 1D convolution followed by adaptive pooling, incurring relatively low computational cost. The cross-modal guided learning module maps unimodal features into probability distributions and estimates their semantic discrepancy using KL divergence, with a total complexity of $\mathcal{O}(d^2)$. The cross-modal fusion module leverages an attention mechanism with complexity $\mathcal{O}(L_t L_v d)$ and introduces an outer product operation to capture high-order interactions across modalities, contributing an additional $\mathcal{O}(d^2)$ complexity, making it the most computationally intensive part of the model. Under the experimental configuration, the average inference time per sample is approximately 100–160 ms, which is acceptable for most real-time or near-real-time applications.

5 Conclusion

This paper proposes a Multi-Granularity Guided Fusion (MGGF) framework for multimodal sentiment analysis, aiming to improve cross-modal feature interaction and sentiment classification accuracy. By jointly constructing global and local sentiment representations within each modality and integrating them through CMGL and CMF, MGGF adaptively fuses unimodal and cross-modal features, effectively addressing misclassification caused by semantic inconsistencies across modalities. Overall, MGGF offers a robust solution to the core challenge of insufficient feature representation and ineffective fusion strategies in current multimodal sentiment analysis research.

6 Discussion

Although the proposed MGGF framework achieves competitive accuracy and robustness in multimodal sentiment analysis tasks, its overall computational overhead remains relatively high. In particular, the high-order interaction operations within the cross-modal fusion module significantly impact inference efficiency. To enhance the model's applicability in real-world deployment scenarios, future work will focus on model lightweighting. Specifically, we plan to explore low-rank interaction modeling, knowledge distillation, and parameter-sharing strategies, aiming to substantially reduce inference latency and resource consumption while maintaining strong performance.

Acknowledgement: Not applicable.

Funding Statement: This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3102904, in part by the National Natural Science Foundation of China under Grant No. U23A20305 and No. 62472440.

Author Contributions: Zilin Zhang: Conceptualization, Methodology, Formal Analysis, Writing—Original Draft. Yan Liu: Conceptualization, Supervision, Writing—Review & Editing. Jia Liu: Software, Implementation, Data Curation, Experiments. Senbao Hou: Validation, Visualization, Writing—Review & Editing. Yuping Zhang: Resources, Investigation, Dataset Processing. Chenyuan Wang: Supervision, Project Administration, Funding Acquisition. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The dataset used in this study is publicly available at https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/ (accessed on 28 September 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Singh U, Abhishek K, Azad HK. A survey of cutting-edge multimodal sentiment analysis. ACM Comput Surv. 2024;56(9):1–38. doi:10.1145/3652149.
2. Tsai YHH, Li T, Liu W, Liao PY, Salakhutdinov R, Morency LP. Integrating auxiliary information in self-supervised learning. arXiv:2106.02869. 2021.
3. Zong D, Ding C, Li B, Xu R, Huang X. Acformer: an aligned and compact transformer for multimodal sentiment analysis. In: Snoek C, Ngo CW, Mei T, Sebe N, editors. Proceedings of the 31st ACM International Conference on Multimedia (MM'23); 2023 Oct 29–Nov 3; Ottawa, ON, Canada. New York, NY, USA: Association for Computing Machinery; 2023. p. 833–42. doi:10.1145/3581783.3611974.
4. Chen M, Wang S, Liang PP, Baltrušaitis T, Zadeh A, Morency LP. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Schuller B, Zhang Y, Weninger F, editors. Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI'17); 2017 Nov 13–17; Glasgow, UK. New York, NY, USA: Association for Computing Machinery; 2017. p. 163–71. doi:10.1145/3136755.3136801.
5. Wang L, Peng J, Zheng C, Zhao T, Zhu LA. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. Inf Process Manag. 2024;61(3):103675. doi:10.1016/j.ipm.2024.103675.
6. Li X, Zhang H, Dong Z, Cheng XF, Liu Y, Zhang XM. Learning fine-grained representation with token-level alignment for multimodal sentiment analysis. Expert Syst Appl. 2025;269(2):126274. doi:10.1016/j.eswa.2024.126274.
7. Yadav A, Vishwakarma DK. A deep multi-level attentive network for multimodal sentiment analysis. ACM Trans Multimedia Comput Commun Appl. 2023;19(1):1–19. doi:10.1145/3517139.
8. Zhu T, Li L, Yang J, Zhao S, Liu H, Qian JS. Multimodal sentiment analysis with image-text interaction network. IEEE Trans Multimedia. 2022;25:3375–85. doi:10.1109/tmm.2022.3160060.
9. 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(4):110502. doi:10.1016/j.knosys.2023.110502.
10. Wang D, Guo X, Tian Y, Liu J, He LH, Luo X. TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognit. 2023;136(2):109259. doi:10.1016/j.patcog.2022.109259.
11. Niu T, Zhu S, Pang L, Li J, Cao Y. Sentiment analysis on multi-view social data. In: Schoeffmann K, Chalupsky V, Hung H, Ngo CW, O'Connor NE, editors. Multimedia Modeling. Proceedings of the 22nd International Conference on Multimedia Modeling (MMM 2016); 2016 Jan 4–6; Miami, FL, USA. Cham, Switzerland: Springer International Publishing; 2016. p. 15–27. doi:10.1007/978-3-319-27674-8_2.
12. Xu N, Mao W, Chen G. A co-memory network for multimodal sentiment analysis. In: Nie J-Y, Baeza-Yates R, Croft WB, editors. In: Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval; 2018 Jul 8–12; Ann Arbor, MI, USA. New York, NY, USA: ACM; 2018. p. 929–32.
13. Yang X, Feng S, Wang D, Zhang Y. Image-text multimodal emotion classification via multi-view attentional network. IEEE Trans Multimedia. 2020;23:4014–26. doi:10.1109/tmm.2020.3035277.
14. Yang X, Feng S, Zhang Y, Wang S, Zhang D. Multimodal sentiment detection based on multi-channel graph neural networks. In: Moens M-F, Huang X, Specia L, Yih W-T, editors. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2021 Aug 1–6; Online. Stroudsburg, PA, USA: Association for Computational Linguistics; 2021. p. 328–39. doi:10.1162/coli_r_00312.
15. Li Z, Xu B, Zhu C, Zhao T. CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection. arXiv:2204.05515. 2022.

16. Wei Y, Yuan S, Yang R, Shen L, Li Z, Wang L, et al. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In: Bouma G, Merlo P, Nivre J, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2023 Jul 9–14; Toronto, ON, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics; 2023. p. 5240–52.
17. Yu B, Li C, Shi Z. Multi-grained feature gating fusion network for multimodal sentiment analysis. Knowl Inf Syst. 2025;67(8):6879–905. doi:10.1007/s10115-025-02446-x.