RESEARCH-ARTICLE

Multimodal Sentiment Analysis with Multi-Perspective Thinking via Large Multimodal Models

JUHAO MA, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

SHUAI XU, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

YICONG LI, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China

XIAOMING FU, The University of Göttingen, Gottingen, Niedersachsen, Germany

Open Access Support provided by:

The University of Göttingen

Nanjing University of Aeronautics and Astronautics

Published: 10 November 2025

Citation in BibTeX format

CIKM '25: The 34th ACM International Conference on Information and Knowledge Management
November 10 - 14, 2025
Seoul, Republic of Korea

Conference Sponsors:
SIGWEB
SIGIR

Multimodal Sentiment Analysis with Multi-Perspective Thinking via Large Multimodal Models

Juhao Ma
Nanjing University of Aeronautics and Astronautics
Nanjing, China
majuhao2002@nuaa.edu.cn

Shuai Xu*
Nanjing University of Aeronautics and Astronautics
Nanjing, China
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
xushuai7@nuaa.edu.cn

Yicong Li
Nanjing University of Aeronautics and Astronautics
Nanjing, China
liyicong123@outlook.com

Xiaoming Fu
University of Göttingen
Göttingen, Germany
fu@cs.uni-goettingen.de

Abstract

Multimodal sentiment analysis (MSA) is attracting increasing attention from researchers. Existing studies on MSA typically rely on surface-level feature extraction and fusion that can be directly obtained from multimodal data, which may often ignore the underlying semantic connection between images and texts. Recent progress in large multimodal models (LMMs) has demonstrated their impressive reasoning abilities, which can be leveraged to improve traditional MSA approaches by providing a deeper understanding of the sematic connection of the modalities. Toward this issue, in this paper, we propose a novel framework called **MPT** that combines traditional MSA approaches with **M**ulti-**P**erspective **T**hinking from LMMs to improve prediction outcomes. Specifically, MPT instructs the traditional multimodal deep learning models to understand multiple-perspective rationales for different sentiment polarities, augmenting its knowledge base and enhancing its ability to make more accurate predictions. Extensive experiments on four refined datasets show that MPT can not only deliver better performance compared with existing methods, but also demonstrate good cross-modal understanding ability for recognizing user sentiment. The codes and datasets can be accessed here: https://github.com/RMJHQwQ/MPT.

CCS Concepts

• **Information systems** → **Multimedia information systems**; **Social networks**.

*Corresponding author

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761328

Keywords

Multimodal Sentiment Analysis, Large Multimodal Model, Multi-Perspective Thinking, Contrastive Learning

ACM Reference Format:
Juhao Ma, Shuai Xu, Yicong Li, and Xiaoming Fu. 2025. Multimodal Sentiment Analysis with Multi-Perspective Thinking via Large Multimodal Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746252.3761328

1 Introduction

Social networks like Twitter[1] and Xiaohongshu[2] have become an indispensable part of our daily lives, and we have become accustomed to expressing our views or emotions toward specific things on social networks. Currently, most of the generated content on social networks consists of multimodal data involving images and texts [26]. Given a post that contains an image and the corresponding text, multimodal sentiment analysis (MSA) aims to infer the user's sentiment polarity [5]. In Fig.1, we display representative tweet examples that convey users' positive, neutral or negative sentiment.

Existing studies basically rely on designing complicated deep neural network architectures to extract features from different data modalities, and further attempt to predict the user sentiment hidden in the multimodal data through cross-modality feature interaction [8, 10, 21]. However, the operating mechanism of these traditional approaches is basically like a black box which is unable to provide reasons or reasoning processes for making predictions and therefore lacks interpretability to a large extent. Moreover, traditional approaches are generally limited to extracting surface-level features from different modalities and trying to align them, with little consideration given to the deep semantic connection between the image and the text. We take the post shown in Fig. 1 (c) as an example, where the user complains in the text that her Christmas gift is only a red hat, even though she looks smiling in the image.

[1]www.twitter.com
[2]www.xiaohongshu.com

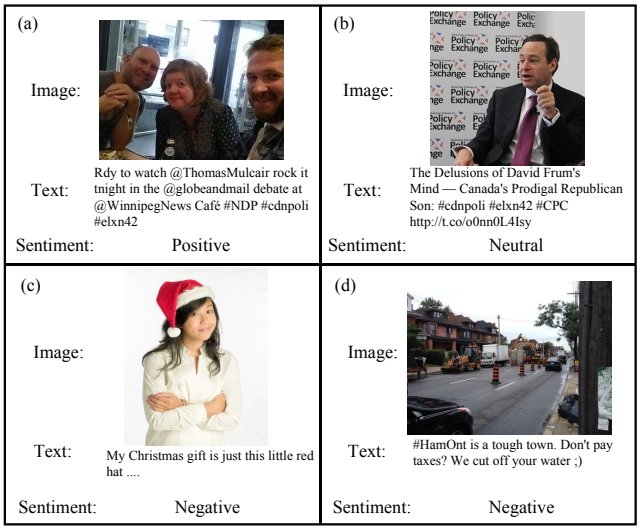


Figure 1: Examples of multimodal tweets conveying positive, neutral or negative sentiment polarities.

Obviously, the sentiment label of this post is negative considering both the text and the image. However, if judged solely from the image, the post should convey positive emotion because the user looks smiling. This means that if only surface-level features of different modalities are considered, traditional approaches will face some contradictions (at least in this example and similar cases), making it difficult to accurately identify user sentiment. In such a circumstance, it is essential to incorporate more deeper semantic connection between the image and the text for MSA tasks.

Recently, large multimodal models (LMMs) like GPT-4o and Qwen-2.5-VL have demonstrated impressive reasoning (or thinking) capabilities in cross-modality information retrieval tasks, particularly adept at capturing deep semantic connections between different data modalities [20]. Therefore, if traditional approaches can be combined with LMMs, the performance of MSA is expected to be improved. To achieve this goal and address the above issue, in this paper, we conceive a novel framework called MPT, which integrates the feature extraction and fusion capabilities of traditional deep learning models with the multi-perspective thinking capability of LMMs. Fig. 2 illustrates an overview of MPT. The characteristics of this framework are two-fold. First, we make modifications to existing multimodal datasets such as MVSA-Single / MVSA-Multiple [22], Memotion [23] and CH-Mits [21]. Specifically, we introduce indicative prompts corresponding to different sentiment polarities and collect responses from LMMs, the primary objective of which is to allow LMMs to analyze textual and visual information from multiple sentiment perspectives. By prompting LMMs to interpret the same text-image pair through different sentimental lenses, our approach is able to understand multiple rationales for different sentiment polarities. Second, to learn the fused feature representation of a specific sentiment label and differentiate the feature representations of different sentiment labels, we deploy label-based contrastive learning after a lightweight attention-based feature fusion.

In brief, contributions of this paper are summarized as follows.

- LMMs are leveraged to strengthen existing deep learning models for multimodal sentiment analysis, where multi-perspective thinking plays the role of instructing traditional approaches to understand multiple rationales for different sentiment polarities, augmenting their knowledge base and enhancing their ability to make more accurate predictions.
- A meticulously designed network architecture is used to gradually fuse text and image input as well as multiple-perspective rationales for different sentiment polarities from LMMs. The fused feature is not only used for classification, but also taken as input by a multi-label contrastive learning mechanism, which is employed to better learn representations of different sentiment labels.
- Various multimodal datasets are refined, where each image-text pair is enriched by multi-perspective thinking results from LMMs with regard to different sentiment labels. Based on the curated datasets, extensive experiments are carried out, and the results verify that the MPT framework can not only perform better than state-of-the-art approaches with regard to quantitative metrics, but also demonstrate good cross-modal understanding ability for recognizing user sentiment. The refined datasets and codes are all released.

2 Related Works

2.1 Multimodal Sentiment Analysis

The earliest visual-textual sentiment analysis models are feature-based [34]. In SentiBank [2], 1200 adjective-noun pairs were extracted as visual features and SentiStrength was used to calculate text features in the multimodal tweet sentiment analysis task. Convolution Neural Networks (CNNs) were also employed to extract features from texts and images in [39]. In recent years, MSA has also attracted much attention. In [36], the authors proposed the TumEmo dataset and the MVAN model for MSA. In [16], contrastive learning and data augmentation were utilized to align and fuse token-level text and image. Yang et al. [33] disentangled representation learning to reduce the distribution gap and the redundancy of information that exists between heterogeneous modalities. PEMNet [21] applies a parallel feature extraction method to obtain a richer semantic representation. The attention mechanism also achieves great performance on MSA tasks. CMMT [35] addressed the limitations of existing methods by incorporating aspect / sentiment-aware intramodal representations and a Text-Guided Cross-Modal Interaction Module to dynamically control the contributions of visual information. AoM [43] introduced an aspect-aware attention module to selectively align relevant textual tokens and image regions. In terms of feature fusion, [9, 25] tried effective ways to take into account interactions between different modalities.

Although existing approaches relying on complicated deep neural networks have shown effectiveness in MSA tasks, they still focus on surface-level feature extraction from different data modalities, with less attention paid to the information interaction hidden between different modalities.

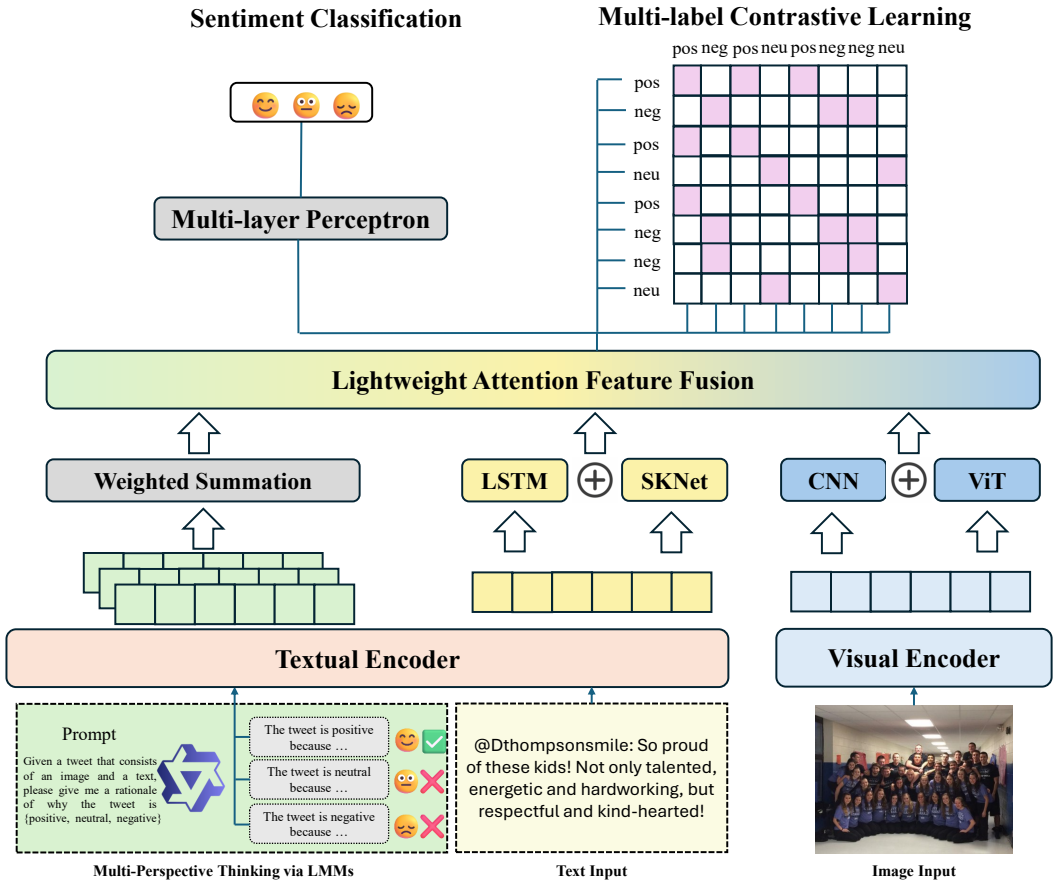


Figure 2: An overview of the proposed MPT framework.

2.2 LMMs for Multimodal Data Processing

Li et al. [13] proposed BLIP-2 which bridges the modality gap using a lightweight Querying Transformer, which is pre-trained in two stages. Liu et al. [19] introduced LLaVa which combined the CLIP [7] visual encoder and the Vicuna language model through a two-stage instruction tuning. Bai et al. [1] introduced Qwen-VL which excels in understanding both texts and images. Chen et al. [3] proposed Janus-Pro that achieves significant advances in both multimodal understanding and text-to-image instruction-following capabilities.

LMMs have been applied in various tasks [17, 18, 24]. Their effectiveness and power are also widely recognized in MSA tasks [4]. Traditional MSA models only analyzed the superficial information of features, which limits their ability to exploit deeper semantic-level information. WisdoM [28] utilized LMMs to generate contextual world knowledge to aid predict sentiments. Inspired by the success of prompt-based fine-tuning approaches in a few-shot scenario, several multimodal prompt-based fine-tuning methods such as UP-MPF [41] and PVLM [40] were proposed. Wu et al. [29] introduces a novel fine-tuning framework for large language models in few-shot multimodal aspect-based sentiment classification. [15] firstly applies Chain-of-Thought reasoning in multimodal sentiment analysis using the novel MM-PEAR-CoT framework, enhancing text representation with high-level reasoning and cross-modal filtering.

However, despite their powerful reasoning capabilities, LMMs still produce incorrect judgments or entirely unsatisfactory outputs due to limitations inherent in their training paradigm and decoder-only architecture. Consequently, relying solely on LLMs for sentiment analysis is insufficient.

In this work, we employ a **multi-perspective thinking** mechanism via LMMs to enhance traditional multimodal sentiment analysis. Specifically, we prompt the LMM to analyze textual-visual inputs from different sentiment perspectives (positive, neutral, and negative) to infer the user's emotional stance. The insights generated from LMM reasoning process are then incorporated into a meticulously designed network to enrich the feature space of traditional deep learning models. As far as we know, we are the first to adopt such multi-perspective thinking strategy via LMMs for the MSA task.

3 Problem Formulation

As is shown in Fig. 1, we denote a multimodal image-text pair as $m_i = \{v_i, t_i\}$, where v_i is the visual content and t_i is the textual content posted by a user. Each image-text pair m_i has a label p_i that belongs to an element of the sentiment polarity set $\{positive, neutral, negative\}$.

Given a set of multimodal image-text pairs $\mathcal{M} = \{m_1, m_2, ..., m_n\}$ and the corresponding labels $\mathcal{P} = \{p_1, p_2, ..., p_n\}$ where n indicates the number of samples, our objective is to learn a ternary classifier $f(\cdot)$, so that for any given image-text pair $m_j = \{v_j, t_j\}$, we can recognize its sentiment polarity by predicting its label, i.e., $f(m_j) \rightarrow \{positive, neutral, negative\}$.

4 Our Approach

4.1 Multi-perspective Thinking via LMMs

Large multimodal models (LMMs) exhibit a certain degree of reasoning ability, enabling them to analyze various inputs logically and generate reasonable interpretations. To achieve more effective sentiment classification, it is necessary to integrate LMMs with traditional multimodal deep learning sentiment analysis networks. The detailed methodology is as follows:

Given an input of an image and the corresponding text $m_i = \{v_i, t_i\} \in \mathcal{M}$, we prompt the LMMs to generate three competing rationales from three polarities $p_i \in \{positive, neutral, negative\}$, the flow can be depicted as Fig. 3.

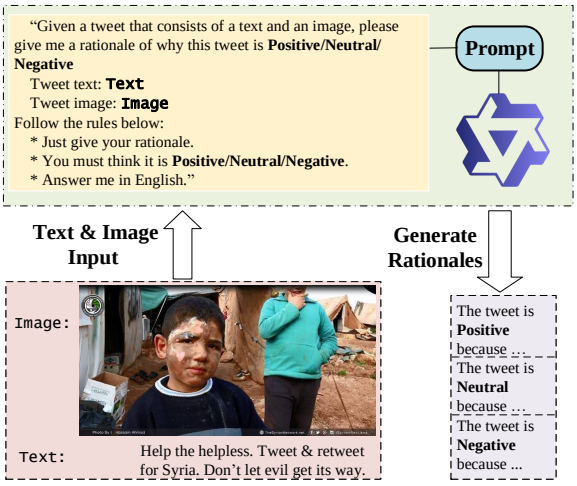


Figure 3: The utilization of the LMMs.

Using this strategy, we can obtain rationales from three different perspectives, namely r_{pos}, r_{neu} and r_{neg}. Equipped with these rationales, we leverage LMMs to analyze the input from diverse perspectives, generating a broader range of background knowledge. Since the generated knowledge includes both useful information for sentiment analysis and irrelevant or unsupported content, it allows other components of the model to access richer contextual information. As a result, this facilitates a more comprehensive assessment of the sentiment expressed in the input from multiple perspectives.

4.2 Sentiment Classification

Feature extraction plays a crucial role in multimodal sentiment analysis, as it directly impacts the model's ability to learn from input features. For text input, we refer to PEMNet [21], which uses BERT encoding for text and employs a parallel feature extraction method combining self-attention, LSTM, and SKNet [14] to process the text, thus efficiently obtaining the feature representation spatially and temporally. In this work, since BERT has already performed extensive calculations of token-level interactions through self-attention, we consider removing the self-attention mechanism from the text feature extraction in PEMNet to accelerate model training and reduce computational overhead. Subsequent comparison and ablation experiments will also demonstrate the effectiveness of this step.

Specifically, given a text input $t_i \in m_i$, we extract the feature of it in parallel. Let $SKNet(*)$ be the spatial feature extractor while $LSTM(*)$ be the temporal feature extractor. The output of these two models can be depicted as h_1 and h_2:

$$h1 = SKNet(t_i), h_2 = LSTM(t_i), t_i \in m_i \tag{1}$$

Then, we take the concatenation of h_1 and h_2 as the output of the text feature representation:

$$t^{'} = h_1 \oplus h_2 \tag{2}$$

In terms of image features, we consider integrating both semantic and spatial features. Therefore, we use the Vision Transformer (ViT) and CNN for feature extraction and concatenate the output results to form the feature representation of the image input. Given an image input $v_i \in m_i$, applying the same parallel extraction method as the text feature extractor, we can derive the spatial and temporal features h_3 and h_4 of the image at the same time utilizing $CNN(*)$ and $ViT(*)$:

$$h_3 = CNN(v_i), h_4 = ViT(v_i), v_i \in m_i \tag{3}$$

Then, to align the visual and textual space, we employ a projection layer to match the dimensionality of the image features with that of the text features, where $v^{'}$ is the final representation of the visual feature, W_p and b_p are the parameters of the projection layer:

$$v^{'} = W_p(h_3 \oplus h_4) + b_p \tag{4}$$

To integrate the rationales generated by large models, we first embed the three rationales r_{pos}, r_{neu} and r_{neg} produced by the LMMs, treating them as the "third modality" in the multimodal fusion process.

$$r^{'} = TextEncoder(R), R = \{r_{pos}, r_{neu}, r_{neg}\} \tag{5}$$

Then, a lightweight attention mechanism [9] is applied to compute the importance weights for the three modalities, followed by a weighted summation to obtain the final fused representation of the input image-text pair and the LMMs' rationale-based predictions.

$$\bar{f} = \sum_{i=1}^{3} \alpha_i f_i \tag{6}$$

where $f_i \in \{v^{'}, t^{'}, r^{'}\}$ and $\alpha_i = Softmax(LinearProjection(f_i))$

Finally, we employ a multi-layer perceptron to make predictions, where the cross-entropy loss is applied as the loss function of the classification task.

4.3 Contrastive Learning

To better learn the fused feature representations of multimodal inputs from a specific class of labels and differentiate the features from different labels, we employ a contrastive learning approach.

Based on the labels of the multimodal data, we categorize the samples into specific classes. For positive samples (the purple squares in Fig. 2), all other samples except the other positive ones are treated as negative samples (the white squares in Fig. 2). The same principle applies to neutral and negative samples. We then compute the similarity of the feature and obtain the contrastive learning loss. This loss is subsequently combined with the loss from the classification task to form the overall loss of the model, which is used for backpropagation. The detailed algorithm for getting the mask of MLCL is presented in Algorithm 1.

Algorithm 1 Multi-label Contrastive Learning Mask Generation

Input: Labels of image-text pair **L**
Output: Multi-label Contrastive Learning Mask
Stage 1: Compute label mask:
Create an empty matrix mask of size $N \times N$
for $i = 1$ **to** N **do**
  for $j = 1$ **to** N **do**
    if $\mathrm{L}[i] = \mathrm{L}[j]$ **then**
      $\mathrm{mask}[i, j] = 1$
    else
      $\mathrm{mask}[i, j] = 0$
    end if
  end for
end for
return *mask*

Inspired by supervised contrastive learning [11], after obtaining the mask, we are able to compute the multi-label contrastive learning loss, the process of which is given in Algorithm 2.

Algorithm 2 Computation of Multi-label Contrastive Learning Loss

Input: Text encoder feature vector $\mathbf{f}_t$, Image encoder feature vector $\mathbf{f}_i$
Output: Contrastive loss $\mathcal{L}_{\mathrm{MLCL}}$
$\mathbf{z}_t \leftarrow$ Text Encoder($\mathbf{f}_t$)
$\mathbf{z}_i \leftarrow$ Image Encoder($\mathbf{f}_i$)
for each pair of features $(\mathbf{z}_t, \mathbf{z}_i)$ **do**
  Compute cosine similarity:

$$s[t, i] = \frac{\mathbf{z}_t \cdot \mathbf{z}_i}{\|\mathbf{z}_t\| \|\mathbf{z}_i\|}$$

  If pair is positive **then**
    Compute positive cosine similarity:
    $\mathbf{s}_+[\mathbf{t}, \mathbf{i}] \leftarrow \mathrm{s}[t, i] * mask[t, i]$
  Else
    Compute negative cosine similarity:
    $\mathbf{s}_+[\mathbf{t}, \mathbf{i}] \leftarrow \mathrm{s}[t, i] * (1 - mask[t, i])$
end for
Compute the Multi-label Contrastive Learning:

$$\mathcal{L}_{\mathrm{MLCL}} = -\sum_{i}^{N} \frac{1}{|C(i)|} \sum_{t}^{N} \log \frac{\exp(s_+[t, i])}{\sum_j^N \mathbf{1}_{j \neq i} \exp(s[t, j]) + \epsilon}$$

return $\mathcal{L}_{\mathrm{MLCL}}$

4.4 Model Training

We adopt the cross-entropy loss for sentiment classification and adopt the multi-label contrastive learning loss for contrastive learning. To enable effective multi-task learning, we assign a specific weight to each loss term, ensuring that the model updates its parameters appropriately through backpropagation. By balancing these loss functions, we aim to optimize the joint learning process and enhance the overall performance. The total loss function for our multi-task framework can be formulated as follows:

$$\mathcal{L}(\Theta) = \lambda_1 \cdot \mathcal{L}(\Theta)_{CE} + \lambda_2 \cdot \mathcal{L}(\Theta)_{MLCL} \tag{7}$$

We denote Θ as the parameters in our model to be learned, $\mathcal{L}_{CE}$ as the cross-entropy loss and the $\mathcal{L}_{MLCL}$ as the multi-label contrastive loss. The weights λ_1 and λ_2 determine the degree of emphasis placed on each loss term during the training process, where $0 < \lambda_2 < \lambda_1 \leq 1$ due to the fact that the weight of contrastive learning should not exceed the weight of the main task, i.e., the classification task.

5 Experiments

To validate the effectiveness of our model in the multimodal sentiment analysis task, we perform comparative experiments on four different datasets. The results demonstrate that our model achieves outstanding performance across all datasets. To further investigate the structural effectiveness of our model, we perform an ablation study, systematically removing components of the model to understand their individual contributions. In addition, we conduct a case study to provide more intuitive insights into how our model captures multimodal sentiment factors. This case study allows us to analyze how well the model integrates and interprets the information from various modalities, such as text, images, and other auxiliary data, to better predict sentiment in real-world scenarios.

5.1 Datasets

Experiments are conducted using four datasets listed below:

- **MVSA-Single** and **MVSA-Multiple** [22] are image-text pairs collected from Twitter. Each pair contains an iamge and a corresponding text with a unified label. To ensure fair comparisons, we preprocess the dataset the same way in MultiSentiNet [31].
- **Memotion** [23] has 10k memes which are obtained from social media. Each meme image is accompanied by a corresponding text. By combining the text within the meme and the meme itself, a unified label is assigned to represent the sentiment stance of the person who posted the meme.
- **CH-Mits** [21] is a Chinese multimodal dataset derived from one of the most famous Chinese social media the Xiaohongshu, which contains a large amount of multimodal content with the blogger's emotional stances.

To reduce computational costs during both training and inference, we freeze the LMM parameters and construct a dataset that contains readily available outputs based on each of the aforementioned dataset. We split all the dataset into three parts: training set, validation set, and test set with a ratio of 8:1:1. Statistics of the datasets are shown in Table 1.

Table 1: Statistics of the selected datasets.

Dataset	# Train	# Val	# Test	# Total
MVSA-Single	3,611	450	450	4,511
MVSA-Multiple	13,624	1,700	1,700	17,024
Memotion	5,594	699	699	6,992
CH-Mits	1,620	202	202	2,024

5.2 Experimental Setup

All of our experiments are conducted on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU (24GB memory).

In order to verify the effect of LMMs on MSA task, we employ the Qwen2.5-VL-7B-Instruct[3] as the instructor of the traditional models. For each dataset, we set the optimal parameter combinations to ensure the best performance of our model. The details are shown in Table 2. We evaluate the performance of the model using two metrics: Accuracy (Acc) and Weighted F1 Score. Accuracy measures the overall proportion of correctly classified samples, providing a general evaluation of the model's ability for sentiment classification. Weighted F1 Score, on the other hand, accounts for class imbalances by computing the F1-score for each class and weighting them by the class distribution, which ensures that the evaluation reflects the model's effectiveness across all categories, even in imbalanced datasets.

Table 2: Parameter settings for different datasets.

Parameters	MVSA-S	MVSA-M	MEMOTION	CH-Mits
Learning Rate	1×10^{-5}	5×10^{-6}	1×10^{-5}	5×10^{-5}
Optimizer	AdamW			
Batch Size	16	8	16	16
Scheduler	Cosine Annealing			
Epoch	20			
Embedding Size	768			

5.3 Baselines

We compare our framework MPT with several unimodal and multimodal baselines.

5.3.1 Unimodal Baselines. CNN [12] and LSTM [42] are employed in text classification tasks. BERT [6] captures the bidirectional context information, improving the model's understanding of input texts. For images, we utilize CNN and Visual Transformer [7] to evaluate the model's capability in capturing spatial and semantic features from the visual modality.

5.3.2 Multimodal Baselines. We also compare MPT with following MSA models: CoMN [32], MVAN [37], MGNNS [38], CLMLF [16], ITIN [44], CGAFT [30] and CiteNet [27].

[3]https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

- **CoMN** [32] incorporates context-aware mechanisms to enhance cross-modal understanding for sentiment classification, enabling better alignment between modalities and emotional context.
- **MVAN** [37] employs cross-modal attention to effectively capture complementary information across visual, audio, and textual modalities, improving robustness in diverse emotional scenarios.
- **PEMNet** [21] utilizes the parallel feature extraction method to obtain the spatial and textual features of the input.
- **MGNNS** [38] uses graph neural networks to structure multimodal data and performs semantic-level fusion for sentiment reasoning, facilitating better relational understanding among multimodal features.
- **CLMLF** [16] introduces contrastive learning to guide multilevel fusion of multimodal features, enhancing sentiment discriminability through more informative feature representations.
- **ITIN** [44] employs an image-text interaction network to explore the intricate associations between affective image regions and textual words for emotion detection.
- **CGAFT** [30] designs an adaptive fine-tuning method that integrates fine-grained interactions between image patches and text words for sentiment prediction.
- **CiteNet** [27] introduces a novel multimodal sentiment prediction approach that uses an extraction-estimation-fusion paradigm to improve accuracy by addressing sentiment incongruities and employing cross-modal fusion techniques.

5.4 Experimental Analysis

5.4.1 Comparisons with baselines. We conduct experiments with other multimodal baselines on four datasets. Moreover, we explore the performances between uni-modality and multi-modality sentiment analysis. The overall results are summarized in table 3.

The experimental results demonstrate that our proposed framework, MPT, consistently outperforms all baseline methods across all four benchmark datasets. In particular, MPT achieves the highest performance on both the MVSA-Single and CH-Mits datasets, which are English and Chinese datasets respectively. This not only highlights MPT's superior capability in handling diverse data distributions and multimodal inputs but also demonstrates its potential for effective cross-lingual multimodal sentiment analysis. These results strongly validate the robustness and analytic ability of MPT across different languages and modalities.

Specifically, on the MVSA dataset, MPT achieves the best accuracy, significantly outperforming the strongest baseline CiteNet. The F1 score also improved, reflecting better feature alignment and sentiment discrimination for subtle visual-textual cues.

For MEMOTION, despite the relatively lower overall Accuracy (58.37%) and F1 (48.16%), MPT still surpasses all baselines, indicating its effectiveness in handling complex emotion expressions and multimodal inputs with implicit or weak emotional semantics.

In particular, on the CH-Mits Chinese-language dataset, our model achieves 98. 02% accuracy, outperforming existing approaches and highlighting the multilingual adaptability of MPT.

Table 3: Experimental results on four datasets, where bolded fonts indicate the best performance.

Modality	Model	MVSA-Single		MVSA-Multiple		MEMOTION		CH-Mits	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Text	CNN [12]	0.6819	0.5590	0.6564	0.5766	0.4657	0.4508	0.7185	0.7183
	LSTM [42]	0.7012	0.6506	0.6790	0.6790	0.5036	0.4536	0.8735	0.8733
	BERT [6]	0.7111	0.6970	0.6759	0.6624	0.5572	0.4608	0.9136	0.9138
Image	CNN [12]	0.6526	0.6364	0.6662	0.6623	0.4943	0.4494	0.8865	0.8863
	ViT [7]	0.6715	0.6226	0.6765	0.5864	0.5401	0.4528	0.9260	0.9270
Image+Text	Co-Mem [32]	0.7051	0.7001	0.6992	0.6883	0.5354	0.4608	0.9580	0.9581
	MVAN [37]	0.7298	0.7298	0.7183	0.7038	0.5583	0.4683	0.9679	0.9679
	PEMNet [21]	0.7317	0.7177	0.7109	0.6921	0.5618	0.4727	0.9778	0.9780
	MGNNS [38]	0.7377	0.7270	0.7249	0.6934	0.5518	0.4665	0.9629	0.9629
	CGAFT [30]	0.7416	0.7305	0.7233	0.6992	0.5715	0.4697	0.9703	0.9703
	ITIN [44]	0.7456	0.7437	0.7215	0.6975	0.5768	0.4754	0.9653	0.9653
	CLMLF [16]	0.7533	0.7346	0.7200	0.6983	0.5761	0.4731	0.9703	0.9702
	CiteNet [27]	0.7609	0.7467	0.7289	0.7035	0.5794	0.4755	0.9752	0.9752
	MPT(ours)	**0.7795**	**0.7562**	**0.7294**	**0.7107**	**0.5837**	**0.4816**	**0.9802**	**0.9802**

The overall comparative experimental results validate the effectiveness of our proposed model, MPT. Across four different datasets of two languages, the LMM-Instructed multimodal sentiment analysis framework outperforms all other traditional deep learning approaches. These results highlight the advantages of leveraging large multimodal models for sentiment analysis tasks, demonstrating their potential to significantly enhance performance in complex multimodal scenarios.

5.4.2 Comparisons with the LMM itself. We also compare MPT with the discrimination capabilities of the LMM itself (Qwen2.5-VL). Fig. 4 presents the performance comparison between MPT and the LMM Qwen2.5-VL across four datasets. The experimental results demonstrate that our model significantly outperforms Qwen2.5-VL in sentiment classification, achieving superior performance on all four datasets. Notably, our model shows substantial improvements on datasets like MEMOTION, which require understanding both the surface and deeper semantics of multimodal inputs. These results highlight that the reasoning ability of large multimodal models can benefit traditional deep learning multimodal sentiment analysis frameworks. Moreover, they suggest that traditional sentiment analysis networks can produce more accurate predictions under multi-perspective instructions from LMMs.

As shown in Fig. 5, the confusion matrix of MPT exhibits more true positives along the diagonal compared to Qwen2.5-VL, and the predicted values are overall closer to the ground truth. For example, both false negatives (FN) and false positives (FP) are significantly reduced. These findings further confirm that our approach achieves a notable improvement in sentiment analysis performance over the LMM itself, indicating that MPT can more accurately identify user sentiment from multimodal content compared to using Qwen2.5-VL alone.

5.4.3 When traditional deep learning models meet LMMs. To verify the effectiveness of the utilization of LMMs, we also conduct experiments with or without the multi-perspective thinking on traditional deep learning models. The results are shown in Table 4. We can observe that all models' performance get improved when thinking

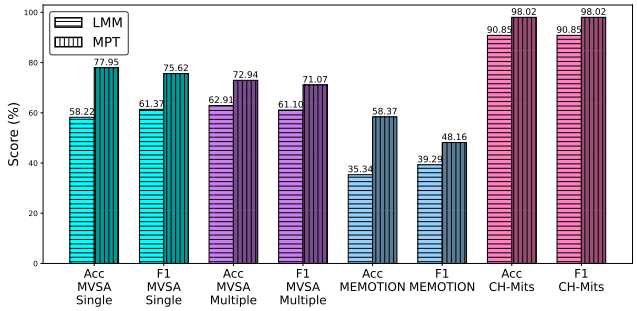
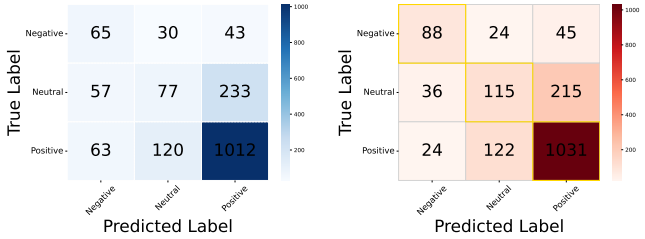


Figure 4: Comparisons between MPT and the LMM Qwen2.5-VL.



(a) LMM **(b) MPT**

Figure 5: Confusion matrices comparisons between MPT and the LMM Qwen2.5-VL.

from multiple perspectives via the LMM. This demonstrates that multiple-perspective thinking is able to strengthen traditional deep learning models in MSA tasks.

5.4.4 When sentimental polarities are opposite. In the dataset MEMOTION and CH-Mits, there exist 1,941 and 105 image-text pairs that deliver opposite sentiments. We also conduct the experiments specifically on those cases. The results are summarized in Table 5.

Table 4: Performance of deep learning models w/ or w/o multiple-perspective thinking for MSA task.

Model	LMM	Acc	F1	Acc	F1
		MVSA-Single		MVSA-Multiple	
CoMN [32]	w/o MPT	0.7051	0.7001	0.6992	0.6883
	w/ MPT	0.7183	0.7106	0.7040	0.6992
CLMLF [16]	w/o MPT	0.7533	0.7346	0.7200	0.6983
	w/ MPT	0.7595	0.7386	0.7228	0.7019
ITIN [44]	w/o MPT	0.7456	0.7437	0.7215	0.6975
	w/ MPT	0.7584	0.7482	0.7239	0.7066

We see that MPT maintains a strong predictive ability even in instances of sentiment inconsistency between text and image, indicating that our model can identify sarcastic scenarios to some extent. Furthermore, achieving such a strong capability on both Chinese and English datasets demonstrates that our model possesses robust prediction abilities across datasets in different languages.

Table 5: Performance on text-image pairs with opposite sentimental polarities.

Model	Acc	F1	Acc	F1
	MEMOTION		CH-Mits	
Co-Mem	0.5309	0.4419	0.9238	0.9238
MVAN	0.5361	0.4446	0.9333	0.9333
PEMNet	0.5490	0.4467	0.9524	0.9524
MPT	**0.5644**	**0.4514**	**0.9619**	**0.9619**

5.5 Ablation Study

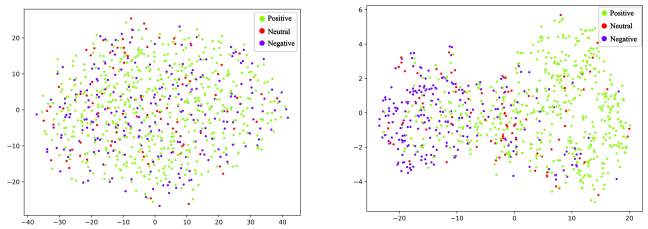
We also conduct the ablation study on the dataset MVSA-Single and MVSA-Multiple. The result of the experiment is depicted in the Table 6.

Table 6: Ablation study results on the MVSA-Single and MVSA-Multiple datasets. "✓" indicates the module is enabled, and "✗" indicates it is disabled.

R_{LMM}	MLCL	Model	Acc	F1	Acc	F1
			MVSA-Single		MVSA-Multiple	
✗	✗	Baseline	0.7206	0.7117	0.7098	0.6886
✓	✗	+R_{LMM} Only	0.7528	0.7347	0.7224	0.7012
✗	✓	+MLCL Only	0.7317	0.7177	0.7192	0.6935
✓	✓	Full Model	**0.7795**	**0.7562**	**0.7294**	**0.7107**

As shown in the table, both the R_{LMM} (Rationales from the LMMs) and MLCL (Multi-label Contrastive Learning) modules contribute positively to the overall model performance. When both modules are enabled in the Full Model, the highest accuracy and F1 scores are achieved across both datasets, confirming the synergy and complementary benefits of the two modules. In summary, R_{LMM} focuses on semantic enhancement, while MLCL strengthens similar representations of vectors from the same label, and their combination effectively improves model's performance, validating the effectiveness of the guidance of LMMs on the traditional MSA task.

To demonstrate the effectiveness of MLCL and LMM in our framework, we perform t-SNE dimensionality reduction visualizations on the results using contrastive learning alone and a combination of contrastive learning with LMM in Fig. 6. Figures (a) and (b) show the visualizations obtained using contrastive learning alone. After 10 iterations, the features of data with different labels become more similar after vectorization and dimensionality reduction, indicating that the contrastive learning module effectively brings data of the same label closer together. Figures (c) and (d) illustrate the results when LMM instructions are incorporated into the contrastive learning framework. After just one iteration, the model exhibits an initial ability to accurately discern user sentiment. After 10 iterations, it is evident that the addition of LMM further increases the separation between data of different labels in the 2-D plane. This highlights the complementary and mutually reinforcing nature of the contrastive learning and LMM modules. These results further validate the effectiveness of the proposed model.



(a) +MLCL, Epoch=1 **(b) +MLCL, Epoch=10**

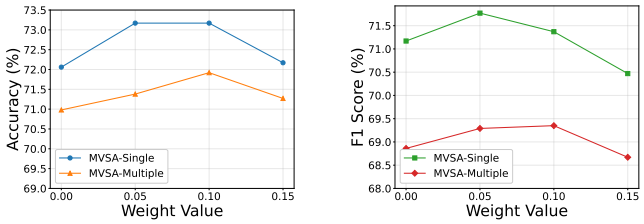
(c) +MLCL, +LMM, Epoch=1 **(d) +MLCL, +LMM, Epoch=10**

Figure 6: t-SNE Visualization on MVSA-Single dataset.

To evaluate the effectiveness of the contrastive learning module in multimodal sentiment analysis, we also conduct experiments on MVSA-Single and MVSA-Multipleb by varying the contrastive learning loss weight λ_2, using accuracy and F1 score as metrics as well. The results are displayed in Fig. 7, from which we see that on both MVSA-Single and MVSA-Multiple datasets, introducing a small contrastive loss weight (0.05 to 0.10) can improve model performance significantly: both accuracy and F1 score can be boosted. This observation indicates that contrastive learning can help extract more robust features in smaller datasets. However, further increasing the weight to 0.15 leads to performance degradation, which is likely due to contrastive loss that overwhelms the main objective of the task.

5.6 Case Studies

To demonstrate the reasoning ability of the MPT framework for cross-modal semantic connection, we collect some typical cases



(a) Accuracy w.r.t. λ_2 **(b) F1 score w.r.t. λ_2**

Figure 7: Impact of contrastive learning weight λ_2.

where MPT can make correct predictions while other models like traditional neural network models (NN) and single large multimodal models (LMM) cannot provide correct predictions (illustrated in Table 7). Many multimodal input pairs contain complex semantics that cannot be accurately interpreted from surface-level information alone. Due to the limitations of decoder-only models, the LMM may struggle to capture the full contextual information necessary for accurate sentiment prediction. In such cases, it becomes essential to combine both approaches to leverage their complementary strengths. The example in the table demonstrates a scenario where both the traditional deep learning model and the LMM individually fail to make the correct prediction, while their combination – our proposed MPT framework – successfully predicts the correct sentiment. This highlights the advantage of integrating two models for more robust and accurate understanding.

In addition, we visualize the attention weights from the final layer of the Transformer structure to further interpret the model's decision-making process. As illustrated in Fig. 8, the highlighted areas represent the regions or tokens that contribute most significantly to the model's prediction. The left side of Fig. 8 presents a positive sentiment example. It can be observed that the model successfully identifies key visual cues such as the subject's facial expression and the trophy held in hand, which indicate a positive emotional state. In the textual modality, words like "celebrate" and "success" are assigned higher attention weights, further supporting a correct sentiment classification. The right side of Fig. 8 showcases a negative sentiment case. The model effectively captures the subject's negative emotion from the image and accurately attends to indicative words in the text, such as "gloomy", which conveys a strong sense of negativity.

This visualization demonstrates that our model is capable of correctly identifying and leveraging critical features from both visual and textual modalities, thereby validating the effectiveness of our proposed approach in multimodal sentiment understanding and analysis.

6 Conclusion

In this study, we introduce multi-perspective thinking (MPT) via large multimodal models as a novel approach to strengthen conventional models for MSA tasks. The proposed framework effectively integrates traditional MSA models with advanced LMM inference capabilities, allowing for a more nuanced understanding of sentiment across various modalities, leading to improved performance in the sentiment classification tasks.

Table 7: Examples where MPT can make correct predictions while other models make incorrect predictions.

Image	Text	NN	LMM	MPT
	8am and she's already beaten @JohnCena. When does school start?	Positive	Positive	**Negative**
	Bitter reality of #VyapamScam so @ChouhanShivraj must go @geetv79	Positive	Neutral	**Negative**
	My cat is sad because he arrived in the room and found everyone talking about how his bleak outlook brings them down.	Negative	Neutral	**Positive**

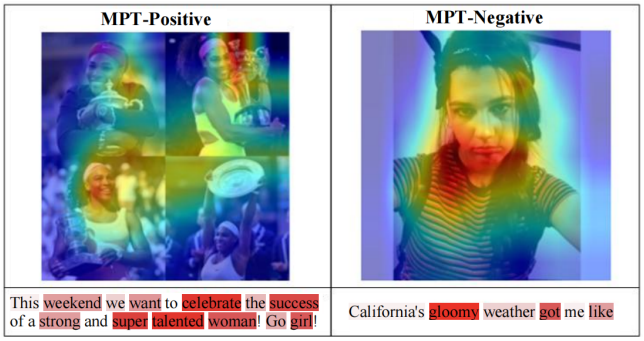


Figure 8: Attention weights visualization of selected samples.

For future work, we could utilize the cross-attention mechanism or contrastive learning to align the visual and textual modality to let them "see" each other, which would enable the model to more accurately capture the intricate relationships between modalities. In addition, we consider conducting user sentiment analysis on short video platforms (e.g., TikTok), where visual, textual, and audio signals can be comprehensively explored.

Acknowledgments

This work is supported by the Natural Science Foundation of China under Grant No. 62302213.

GenAI Disclosure

There was no use of GenAI toos whatsoever in any stage of the research.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966 [cs.CV] https://arxiv.org/abs/2308.12966

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*. 223–232.

[3] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811* (2025).

[4] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems* 37 (2024), 110805–110853.

[5] Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *Comput. Surveys* 55, 13s (2023), 1–38.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[8] Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3394–3402.

[9] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. 2022. Lightweight attentional feature fusion: A new baseline for text-to-video retrieval. In *European Conference on Computer Vision*. Springer, 444–461.

[10] Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 4395–4405.

[11] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.

[12] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*. PMLR, 19730–19742.

[14] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. 2019. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 510–519.

[15] Yan Li, Xiangyuan Lan, Haifeng Chen, Ke Lu, and Dongmei Jiang. 2025. Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 9 (2025), 1–23.

[16] Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF:A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection. arXiv:2204.05515 [cs.CL] https://arxiv.org/abs/2204.05515

[17] Hongzhan Lin, Zixin Chen, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. 2024. CofiPara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. *arXiv preprint arXiv:2405.00390* (2024).

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems* 36 (2023), 34892–34916.

[20] Pengfei Luo, Jingbo Zhou, Tong Xu, Yuan Xia, Linli Xu, and Enhong Chen. 2025. ImageScope: Unifying Language-Guided Image Retrieval via Large Multimodal Model Collective Reasoning. In *Proceedings of the ACM on Web Conference 2025*. 1666–1682.

[21] Juhao Ma, Shuai Xu, Yilin Liu, and Xiaoming Fu. 2024. CH-Mits: A Cross-Modal Dataset for User Sentiment Analysis on Chinese Social Media. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 5390–5394.

[22] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*. Springer, 15–27.

[23] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. SemEval-2020 Task 8: Memotion Analysis–The Visuo-Lingual Metaphor! *arXiv preprint arXiv:2008.03781* (2020).

[24] Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. LawLLM: Law large language model for the US legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4882–4889.

[25] Yongxin Tong, Xuchen Pan, Yuxiang Zeng, Yexuan Shi, Chunbo Xue, Zimu Zhou, Xiaofei Zhang, Lei Chen, Yi Xu, Ke Xu, et al. 2022. Hu-fu: Efficient and secure spatial queries over data federation. *Proceedings of the VLDB Endowment* 15, 6 (2022), 1159.

[26] Yongxin Tong, Jieying She, Bolin Ding, Libin Wang, and Lei Chen. 2016. Online mobile micro-task allocation in spatial crowdsourcing. In *2016 IEEE 32nd International Conference on Data Engineering*. IEEE, 49–60.

[27] Jie Wang, Yan Yang, Keyu Liu, Zhuyang Xie, Fan Zhang, and Tianrui Li. 2024. CiteNet: Cross-modal incongruity perception network for multimodal sentiment prediction. *Knowledge-Based Systems* 295 (2024), 111848.

[28] Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2282–2291.

[29] Hao Wu, Danping Yang, Peng Liu, and Xianxian Li. 2025. Chain of Thought Guided Few-Shot Fine-Tuning of LLMs for Multimodal Aspect-Based Sentiment Classification. In *International Conference on Multimedia Modeling*. Springer, 182–194.

[30] Xingwang Xiao, Yuanyuan Pu, Zhengpeng Zhao, Rencan Nie, Dan Xu, Wenhua Qian, and Hao Wu. 2023. Image–text sentiment analysis via context guided adaptive fine-tuning transformer. *Neural Processing Letters* 55, 3 (2023), 2103–2125.

[31] Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2399–2402.

[32] Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 929–932.

[33] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1642–1651.

[34] Hao Yang, Yanyan Zhao, Yang Wu, Shilong Wang, Tian Zheng, Hongbo Zhang, Zongyang Ma, Wanxiang Che, and Bing Qin. 2024. Large language models meet text-centric multimodal sentiment analysis: A survey. *arXiv preprint arXiv:2406.08068* (2024).

[35] Li Yang, Jin-Cheon Na, and Jianfei Yu. 2022. Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing & Management* 59, 5 (2022), 103038.

[36] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia* 23 (2020), 4014–4026.

[37] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2021. Image-Text Multimodal Emotion Classification via Multi-View Attentional Network. *IEEE Transactions on Multimedia* 23 (2021), 4014–4026. doi:10.1109/TMM.2020.3035277

[38] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 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*. 328–339.

[39] Yuhai Yu, Hongfei Lin, Jiana Meng, and Zhehuan Zhao. 2016. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms* 9, 2 (2016), 41.

[40] Yang Yu and Dong Zhang. 2022. Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling. In *2022 IEEE International Conference on Multimedia and Expo*. IEEE, 1–6.

[41] Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 189–198.

[42] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630* (2015).

[43] Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. 2023. Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2306.01004* (2023).

[44] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. 2022. Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia* 25 (2022), 3375–3385.