

SAMS-M: Explicit Sentiment-Guided Alignment and Multi-dimensional Mutual Supervision for Multimodal Sentiment Analysis

Shishu Qi¹, Yulei Zhang², Siyang Zhang³, Jia Xu¹, Kuan-Ching Li³, and Guangli Zhu²

¹ School of Automotive and Traffic Engineering, Tianjin University of Technology and Education
300222 Tianjin, China
{0721241391,0722231050}@tute.edu.cn

² School of Computer Science and Engineering, Anhui University of Science and Technology
232000 Huainan, China
2023201230aust.edu.cn, glzhu@aust.edu.cn (corresponding author)

³ School of Mathematics and Big Data, Anhui University of Science and Technology
232000 Huainan, China
2025201731@aust.edu.cn, rob5li@outlook.com

Abstract. Multimodal Sentiment Analysis leverages the fusion of heterogeneous data to achieve fine-grained emotional understanding, which finds extensive application in large-scale public opinion monitoring and data mining. However, existing methods face two key challenges: (1) cross-modal alignment suffers from redundancy and semantic drift without explicit modeling of sentiment-critical cues, inducing spurious correlations; and (2) heterogeneous representation spaces lead to imbalanced modality contributions, particularly under weak image–text correlation or sentiment inconsistency. To address these challenges, we propose an explicit sentiment-guided alignment and multi-dimensional cross-modal mutual supervision-based model for multimodal sentiment analysis. The model primarily employs a fine-grained sentiment–saliency directed alignment mechanism, which leverages bidirectional cross-attention to couple textual sentiment cues with visual saliency, enabling precise localization of sentiment-relevant regions. Furthermore, we introduce a tripartite strong contrastive learning strategy to mitigate distribution discrepancies between heterogeneous modalities within a shared latent space, thereby enhancing cross-modal coherence and complementarity. Finally, we design a noise-robust gating-based fusion module, which, together with text augmentation and deep supervision, facilitates effective joint optimization. Experimental results show that SAMS-M obtains the best results on MVSA-Single and MSD and remains competitive on the noisier MVSA-Multiple benchmark; thus, the evidence supports strong but dataset-dependent performance rather than uniform state-of-the-art superiority.

Keywords: Multimodal Sentiment Analysis, Cross-modal Alignment, Contrastive Learning, Sentiment-guided, Noise-robust Fusion.

1. Introduction

With the continuous evolution of digital media technology, human emotional expression has increasingly exhibited significant multimodal characteristics [26,50]. The convergence of textual, visual, and other sensory modalities has established itself as a core paradigm of emotional communication [22,11,68]. Given its extensive applications in fields such as public opinion monitoring, intelligent interaction systems, and consumer behavior analysis, multimodal sentiment analysis has become a frontier topic of continuous research in both academia [45,15]. In recent years, the emergence of advanced pre-trained architectures, such as Transformer, BERT, and ViT, has fundamentally shifted the paradigm from shallow fusion to deep semantic alignment [24,67]. To further address the challenges of semantic complexity, recent studies have increasingly focused on fine-grained modeling and adaptive mechanisms. For instance, researchers have introduced token-level reconstruction and two-stage denoising frameworks to compensate for missing modalities and preserve robust semantic structures [57,72]. Furthermore, advanced approaches utilizing prototype-aligned calibration and grading-inspired complementary mechanisms have been proposed to dynamically balance the contributions of different modalities [13,17], thereby significantly improving the accuracy of sentiment representation in scenarios where multimodal information is incomplete or noisy. In addition, recent works have explored joint alignment, fusion, and denoising strategies by exploiting neighborhood information, which provides a more holistic view of multimodal context [5].

Despite these advancements, the field continues to confront significant hurdles regarding the robustness and precision of sentiment representation. Specifically, how to further enhance the reliability of emotional representation remains a subject worthy of in-depth exploration, especially when dealing with complex multimodal data characterized by image-text semantic discrepancies, pervasive background noise interference, and modal information asymmetry [23,49]. Existing mainstream paradigms, while effective at capturing general emotional cues, still have limitations for further optimization. First, regarding cross-modal alignment, traditional global attention mechanisms tend to perform dense alignment. Although some recent text-centric sparse interaction methods [70] have attempted to optimize this interaction process, the challenge of how to precisely focus on sentiment-critical elements in complex backgrounds—thereby mitigating spurious correlations between non-sentiment textual redundancy and irrelevant image background noise—remains a focal point of current research [42,40]. Unlike existing alignment strategies that mainly emphasize modality correspondence at the feature or token level, we argue that effective multimodal sentiment alignment should be guided by sentiment relevance rather than merely structural similarity. Therefore, explicitly identifying sentiment-bearing information before cross-modal interaction is essential for reducing irrelevant semantic association and improving sentiment-oriented representation learning. Second, the inherent heterogeneity in the semantic representations of text and images makes it difficult to fully map the consistency of their distribution spaces. Although existing contrastive learning strategies have mitigated the modality gap to some extent, there is still limitations for further exploration in maintaining global consistency and modal complementarity in fused features [46,12,66,1]. Moreover, most existing contrastive learning approaches primarily focus on reducing representation discrepancies between modalities, while insufficient attention has been paid to simultaneously preserving intra-modal semantics, inter-modal correspondence, and fused representation consistency. This moti-

vates us to construct a multi-dimensional mutual supervision strategy that provides complementary constraints from multiple perspectives. Furthermore, existing fusion strategies generally rely on fixed or manually designed modality interaction patterns, which may fail to adaptively handle modality imbalance and unreliable information caused by noisy or inconsistent multimodal inputs.

To address these challenges, we propose SAMS-M, an explicit sentiment-guided alignment and multi-dimensional cross-modal mutual supervision framework for multimodal sentiment analysis. Different from existing approaches that separately optimize cross-modal interaction, representation alignment, and modality fusion, SAMS-M establishes a progressive optimization paradigm consisting of three tightly coupled stages. First, instead of performing indiscriminate global interaction, SAMS-M introduces a sentiment-saliency directed alignment mechanism that explicitly identifies sentiment-bearing textual tokens and visually salient regions, thereby reducing irrelevant background interference during cross-modal association. Second, based on the aligned representations, a tripartite mutual supervision strategy is designed to simultaneously constrain textual semantics, visual semantics, and cross-modal consistency. This design prevents heterogeneous modalities from drifting into independent representation spaces and encourages complementary rather than redundant information exchange. Third, a noise-robust gating fusion mechanism dynamically adjusts modality contributions according to sample-specific reliability, allowing the model to adaptively suppress unreliable modalities under sentiment inconsistency scenarios. Therefore, the novelty of SAMS-M does not originate from the isolated adoption of individual components, but from establishing a progressive optimization framework where sentiment-guided alignment provides reliable cross-modal representations, mutual supervision enhances semantic consistency, and adaptive fusion further exploits complementary modality information. Through this tightly coupled design, SAMS-M simultaneously addresses sentiment localization, semantic distribution discrepancy, and modality imbalance in multimodal sentiment analysis.

We summarize the contribution of this paper as below:

- (1) We propose SAMS-M, a unified multimodal sentiment analysis framework that integrates sentiment-guided alignment and multi-dimensional cross-modal mutual supervision. Different from conventional modality interaction strategies, SAMS-M explicitly considers sentiment relevance during alignment and progressively improves representation consistency through mutual supervision.
- (2) We design a sentiment-saliency directed alignment mechanism and a tripartite mutual supervision strategy. The proposed alignment mechanism focuses cross-modal interaction on sentiment-relevant information, while the mutual supervision strategy jointly constrains intra-modal and inter-modal representations to reduce semantic inconsistency.
- (3) We conduct extensive experiments on several public benchmark datasets, and the results demonstrate the effectiveness and competitiveness of SAMS-M compared with existing state-of-the-art methods.

2. Related work

2.1. Text sentiment detection

Before the widespread adoption of deep learning, traditional machine learning algorithms dominated the field of sentiment analysis. Moraes et al. [25] established the benchmark

for bag-of-words models in document-level sentiment classification by comparing Support Vector Machines (SVM) with Artificial Neural Networks (ANN). To address the deficiency of sentiment expressions in specific domains, Kang et al. [18] effectively improved the accuracy of restaurant review analysis by constructing domain-specific sentiment lexicons and optimizing the Naive Bayes algorithm. Building upon this, Wang et al. [39] proposed the SentiDiff algorithm, which pioneered the integration of textual features with sentiment diffusion patterns. By capturing emotional interaction logic within social networks through iterative computation, this approach demonstrated the early advantages of hybrid feature modeling. With the introduction of neural networks, models began to acquire the capability to automatically extract high-level semantic features. Chen [6] pioneered the application of Convolutional Neural Networks (CNN) for sentence classification, proving the superiority of convolution kernels in capturing local semantic features. Concurrently, Tai et al. [34] and Tang et al. [35] utilized Tree-Structured Long Short-Term Memory (Tree-LSTM) and Gated Recurrent Units (GRU), respectively, to process the hierarchical structure of text, thereby mitigating information loss in long-text modeling. To further enhance context sensitivity, Usama et al. [38] integrated the attention mechanism into a CNN-RNN hybrid model to achieve dynamic weighting of keywords, while Zhai et al. [58] further proposed the Multi-Attention Fusion Model (Multi-AFM), significantly enhancing semantic context capture in sentiment analysis scenarios involving educational big data.

In recent years, the research focus has shifted toward deep semantic understanding and fine-grained feature modeling. The SKEP model proposed by Tian et al. [36] refreshed several benchmarks by injecting sentiment knowledge into pre-trained representations via a sentiment masking mechanism. Dai et al. [7] verified the supportive role of syntactic information in Aspect-Based Sentiment Analysis (ABSA), demonstrating the potential of RoBERTa in refined sentiment recognition. Related studies further exploit graph structure and task-specific affective knowledge: TS-GCN models aspect-level dependencies in consumer reviews, sentiment-driven 3D-CNNs extract emotion-cause pairs, syntactically weighted textual graphs improve implicit sentiment recognition, and sequence-labeling pipelines expand sentiment lexicons from product reviews [64,55,60,63]. In terms of expanding application scenarios, researchers have proposed diverse solutions: Bibi et al. [3] optimized Twitter sentiment analysis using an unsupervised ensemble framework; Parveen et al. [28] constructed a Gated Attention Recurrent Network (GARN); and Zhang et al. [65] introduced PU learning to address the problem of fake review detection. Furthermore, Wang et al. [43] realized the automated construction of fine-grained sentiment lexicons via neural networks, while Rodríguez-Ibáñez et al. [30] and Pashchenko et al. [29] further refined the analytical dimensions of consumer reviews from the perspectives of platform synthesis and multi-dimensional attributes (such as sentiment and normativity), respectively.

In summary, although text-based sentiment analysis has achieved remarkable progress driven by deep learning, and models demonstrate excellent performance in specific semantic understanding, social media content in reality often exhibits high complexity. A single textual modality is frequently insufficient to fully encapsulate the rich human emotional expression (e.g., composite emotions accompanying images and audio). Therefore, breaking down modal barriers and integrating cross-modal information through deep learning to achieve more precise emotional understanding has become a core challenge in

the field of affective computing, which serves as the primary motivation for this study on multimodal sentiment analysis.

2.2. Multi-modal sentiment detection

Recent multimodal sentiment analysis increasingly centers on alignment and adaptive fusion. MTVAF performs multi-level textual–visual alignment and fusion, CoDSF uses confidence-guided dynamic sequential fusion, KuDA dynamically selects the dominant modality using sentiment knowledge, and SMCAF introduces semantic-guided multi-grained cross-modal alignment and fusion [19,48,10,21]. A multimodal semantic fusion network with explicit cross-modal alignment provides further evidence that structured image–text correspondence is central to sentiment representation [61]. Beyond conventional polarity classification, topic-oriented knowledge prompting improves multimodal sarcasm detection, while emoji-enhanced multi-feature fusion supports socially implicit affective recognition [71,62]. These studies demonstrate the value of fine-grained correspondence and sample-dependent modality weighting. In contrast, SAMS-M couples sentiment-saliency-guided alignment with three simultaneous text–fusion, image–fusion, and text–image constraints, and then uses a reliability-aware gate on the jointly regularized representations. This mechanism-level combination differs from CLMLF’s two contrastive tasks, MVCN’s calibration-oriented treatment of heterogeneity, CTMWA’s translation-based weight adaptation, and ITQT-MSA’s quantum-transformer interaction: SAMS-M explicitly makes sentiment relevance the input to alignment, uses the fused representation as a shared contrastive anchor, and suppresses unreliable modalities only after this mutual supervision. Early research in Multimodal Sentiment Analysis (MSA) primarily focused on exploring shallow interactions between modalities. To mitigate the limitations of unimodal analysis, researchers proposed various fusion architectures. Zhao et al. [69] were among the first to improve classification performance by exploring the consistency between images and text. Subsequently, Truong et al. [37] introduced a visual attention mechanism to leverage image information for highlighting key sentiment segments in text. To further characterize long- and short-term dependencies, Xu et al. [54] proposed a multi-interaction memory network, effectively supervising cross-modal data flows. Recently, to capture deeper correlations, Wang et al. [41] and Wu et al. [45] introduced manifold learning and graph structures, enabling synchronized modeling of intra- and inter-modal information, which laid the foundation for subsequent complex interaction networks.

With the progression of research, addressing heterogeneity and noise interference in multimodal data has become a new focal point. To cope with semantic asymmetry between text and images, Wu et al. [46] effectively balanced inter-modal consistency via contrastive learning. Addressing interference issues in fine-grained sentiment analysis, Wang et al. [42] and Wang et al. [40] proposed strategies to filter redundant information, aiming to reduce the negative impact of background noise on sentiment determination. Furthermore, Han et al. [12] and An et al. [2] leveraged graph neural networks and gating mechanisms to enhance model robustness in complex scenarios. However, these graph-based approaches mainly focus on modeling structural dependencies and reducing noisy interactions, while explicit sentiment-oriented guidance for cross-modal alignment remains insufficiently explored. Graph-based approaches generally model multimodal samples as structured interaction graphs and alleviate noise by refining node rela-

tionships or information propagation patterns. However, these methods mainly focus on structural-level noise reduction and graph topology optimization, while the identification of sentiment-relevant information during cross-modal interaction remains insufficiently explored. Zhao et al. [66] further improved the mapping quality of different modal features within a unified semantic space through calibration alignment mechanisms. Currently, utilizing large-scale pre-trained models to inject external knowledge has become a mainstream direction for boosting MSA performance. Hu et al. [15] and Liu et al. [24] explored methods using social relationships and label semantics to enhance feature representation. For extreme scenarios involving resource constraints or missing modalities, Afroze et al. [1] and Zhao et al. [67] proposed completion strategies based on cross-attention mechanisms and virtual modality generation, respectively. Moreover, Liu et al. [23] and Xiong et al. [49] demonstrated significant representation advantages in anomaly sentiment detection tasks by introducing self-supervised knowledge enhancement and multi-granularity selection mechanisms.

Although the aforementioned methods have achieved significant success in multimodal feature fusion and robustness enhancement, existing paradigms still leave limitations for further optimization when dealing with highly complex multimodal data. First, most global attention-based methods struggle to completely avoid the introduction of background noise and sentiment-irrelevant redundant information during feature alignment, which leads to semantic deviation during cross-modal interaction. Moreover, graph-based methods still lack sufficient consideration of sentiment-related importance during modality interaction. Second, for samples with contradictory image-text semantics or imbalanced modal contributions, simple fusion strategies fail to achieve precise consistency constraints in the modal representation space, often leading to performance fluctuations when processing heterogeneous features. These challenges indicate that designing a mechanism capable of fine-grained sentiment-guided and multi-dimensional mutual supervision alignment holds significant research value for enhancing the robustness of multimodal sentiment analysis in real-world scenarios.

3. Methodology

3.1. Problem Definition

Let the corpus be denoted by $\mathcal{D} = \{(\mathbf{x}_i^t, \mathbf{x}_i^v, y_i)\}_{i=1}^N$, where $\mathbf{x}_i^t$ and $\mathbf{x}_i^v$ represent textual and visual inputs, respectively. The label space $\mathcal{C}$ is determined according to the annotation protocol of each benchmark. Specifically, $\mathcal{C} = \{L_{neg}, L_{neu}, L_{pos}\}$ is adopted for three-class sentiment datasets, while binary label spaces are employed for datasets without neutral annotations, such as MSD. Let $C = |\mathcal{C}|$. We learn a multimodal discriminative mapping $\mathcal{M}(\cdot, \cdot; \Theta)$ parameterized by $\Theta = \{\theta_\phi, \theta_\psi, \theta_\gamma\}$. The feature representation for each modality $m \in \{t, v\}$ is computed via the extraction operator $\mathbf{z}_i^m = \phi_m(\mathbf{x}_i^m; \theta_m)$. These heterogeneous features are synthesized into a joint representation $\mathbf{h}_i \in \mathbb{R}^{d_h}$ via the cross-modal alignment operator $\psi(\cdot, \cdot; \theta_\psi)$, defined as $\mathbf{h}_i = \psi(\mathbf{z}_i^t, \mathbf{z}_i^v; \theta_\psi)$. The final prediction is a posterior distribution over the $(C - 1)$ -dimensional probability simplex Δ^{C-1} :

$$\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{W}_\gamma^\top \mathbf{h}_i + \mathbf{b}_\gamma) \in \Delta^{C-1} \quad (1)$$

where $\mathbf{W}_\gamma$ and $\mathbf{b}_\gamma$ are the classifier parameters. The optimal parameter set Θ^* is obtained by minimizing the regularized empirical risk:

$$\mathcal{L}(\Theta) = \frac{1}{N} \sum_{i=1}^N \ell_{\text{focal}}(\hat{y}_i, y_i) + \frac{\lambda}{2} \|\Theta\|_2^2 \quad (2)$$

where $\ell_{\text{focal}}(\cdot, \cdot)$ denotes the Focal Loss used in our implementation and λ is the weight-decay coefficient. This generic risk is expanded into the complete multi-branch training objective in the classification and deep-supervision module.

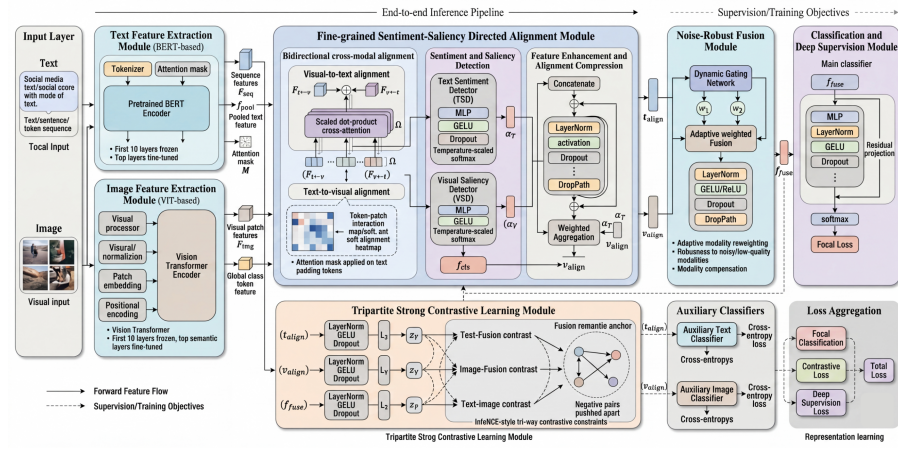


Fig. 1. Overall framework.

3.2. Overall Model Architecture

The overall architecture of our model is illustrated in Figure 1. First, the model extracts deep semantic features for text and images through their respective extraction modules. Second, a fine-grained sentiment-saliency directed alignment module is employed to achieve precise alignment between textual sentiment keywords and visual sentiment-salient regions, yielding the aligned text and image features. Then, a tripartite strong contrastive learning module is utilized to construct multi-way contrastive constraints to optimize the distribution of cross-modal features. Subsequently, a noise-robust fusion module dynamically fuses the aligned features to obtain multimodal fused representations. Finally, sentiment classification is performed via a classification module, combined with a deep supervision strategy (independent classification for text and images) to further optimize model performance.

3.3. Text and Image Feature Extraction Modules

To extract deep semantic features for text and images, this paper constructs text and image feature extraction modules based on pre-trained models. By freezing specific pre-trained

layers and fine-tuning the top layers, we balance feature extraction capability with training efficiency.

Text Feature Extraction Module The text feature extraction module utilizes BERT [8] as the core architecture to map the textual modality T_i into a high-dimensional semantic manifold space. First, the raw text sequence is converted into a discretized token sequence and its corresponding attention mask $\mathbf{M} \in \{0, 1\}^L$ via a tokenizer $\text{Trans}(\cdot)$. Subsequently, the encoded sequence is fed into a deep Transformer hierarchical structure for bidirectional context modeling. To strike a balance between computational efficiency and representational power, we implement a selective parameter freezing strategy by masking the gradient updates of the first k layers (set to $k = 10$ in this study). This allows the model to retain universal low-level linguistic features while enabling specific transfer learning for sentiment semantics through top-layer fine-tuning. The textual feature extraction process can be represented as a composite mapping. Let $\mathbf{H} \in \mathbb{R}^{L \times d}$ be the hidden state sequence output from the last layer of BERT, where d denotes the hidden dimension. The feature representation of the textual modality is formally defined as:

$$\Phi_T(T_i) = \text{BERT}_{\Theta_T}(\text{Trans}(T_i)) = \{\mathbf{F}_{seq}, \mathbf{f}_{pool}, \mathbf{M}\} \quad (3)$$

where $\mathbf{F}_{seq} \in \mathbb{R}^{B \times L \times d}$ represents the sequence feature matrix containing fine-grained semantic information for subsequent cross-modal alignment; $\mathbf{f}_{pool} \in \mathbb{R}^{B \times d}$ is the global semantic vector extracted via a pooling operator, representing the macro-level sentiment expression of the textual modality; Θ_T denotes the subset of trainable parameters in the text branch; and $\mathbf{M}$ is used to mask the influence of padding tokens in the cross-modal attention mechanism, ensuring the effectiveness of feature fusion.

Image Feature Extraction Module The image feature extraction module aims to transform the input visual modality V_i into feature representations containing spatial context and global semantics. This module adopts the Vision Transformer (ViT) [9] as the backbone to capture discriminative emotional regions in images with its powerful global modeling capability. During data preprocessing, input images are normalized by a dedicated visual processor $\text{Proc}(\cdot)$, explicitly preserving the original spatial scale information to minimize underlying feature loss. After patch embedding and position encoding, the images are mapped into a latent space tensor and fed into the Transformer encoder. To suppress overfitting in small-sample scenarios, this module adopts a heuristic parameter freezing strategy by locking the weights of the first k layers (set to $k = 10$), applying discriminative fine-tuning only to the high-level semantic feature extraction layers. The image feature extraction process can be formulated as:

$$\Phi_V(V_i) = \text{ViT}_{\Theta_V}(\text{Proc}(V_i)) = \{\mathbf{F}_{img}, \mathbf{f}_{cls}\} \quad (4)$$

where $\mathbf{F}_{img} \in \mathbb{R}^{B \times P \times d_v}$ represents the visual sequence features, composed of P encoded vectors of image patches, which support subsequent fine-grained cross-modal alignment; $\mathbf{f}_{cls} \in \mathbb{R}^{B \times d_v}$ is the pooled vector extracted from the special class token ([CLS] token), serving as the global discriminative feature of the visual modality; and Θ_V represents the trainable parameter space of the visual branch.

3.4. Fine-grained Sentiment-Saliency Directed Alignment Module

While previous studies have achieved progress through hierarchical cross-modal modeling [16] and adaptive multi-loss fusion frameworks [14], they often struggle to precisely filter task-irrelevant background noise in fine-grained analysis. To address this, we introduce the sentiment-saliency directed alignment module, which leverages the aforementioned interaction mechanisms while imposing an explicit guidance constraint to focus the cross-modal alignment on sentiment-critical regions. We define the interaction between the textual hidden sequence $\mathbf{H}_T \in \mathbb{R}^{L \times d}$ and the visual patch representation $\mathbf{H}_V \in \mathbb{R}^{P \times d}$ as a bidirectional manifold mapping. We introduce the learnable projection parameters $\mathbf{W}_Q^*, \mathbf{W}_K^*, \mathbf{W}_V^* \in \mathbb{R}^{d \times d_k}$ to project the input features into a shared subspace of dimension d_k . The cross-modal interaction is governed by the scaled dot-product attention operator $\mathcal{A}$, which maps a query set $\mathcal{Q}$ against a key-value pair $(\mathcal{K}, \mathcal{V})$ subject to an optional masking tensor $\mathcal{M} \in \{0, -\infty\}^{L \times P}$, defined as:

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}; \mathcal{M}) = \text{softmax} \left(\frac{(\mathbf{H}_Q \mathbf{W}_Q)(\mathbf{H}_K \mathbf{W}_K)^\top}{\sqrt{d_k}} + \mathcal{M} \right) (\mathbf{H}_V \mathbf{W}_V) \quad (5)$$

Within this topological framework, the bidirectional sentiment-saliency flow is realized by coupling the modalities through asymmetric query-key-value associations. The visual-to-textual alignment branch $\mathbf{F}_{t \leftarrow v}$ selectively aggregates visual contexts to modulate textual representations, while the textual-to-visual alignment branch $\mathbf{F}_{v \leftarrow t}$ utilizes textual sentiments to attend to salient visual regions. This bidirectional coupling is formulated as the coupled mapping system:

$$\begin{cases} \mathbf{F}_{t \leftarrow v} = \mathcal{A}(\mathbf{H}_T, \mathbf{H}_V, \mathbf{H}_V; \mathbf{M}) \in \mathbb{R}^{L \times d} \\ \mathbf{F}_{v \leftarrow t} = \mathcal{A}(\mathbf{H}_V, \mathbf{H}_T, \mathbf{H}_T; \mathbf{0}) \in \mathbb{R}^{P \times d} \end{cases} \quad (6)$$

where $\mathbf{M}$ serves as the indicator tensor to inhibit the influence of padding tokens in the text modality, effectively filtering non-sentiment-bearing semantics. The resultant aligned representations are computed as high-order semantic features, where the attention weight distribution $\boldsymbol{\Omega} \in \mathbb{R}^{L \times P}$ is derived as:

$$\boldsymbol{\Omega}_{ij} = \left[\text{softmax} \left(\frac{(\mathbf{H}_{T,i} \mathbf{W}_Q^t)(\mathbf{H}_{V,j} \mathbf{W}_K^v)^\top}{\sqrt{d_k}} \right) \right]_{i,j} \quad (7)$$

This formulation enables the model to dynamically construct a bipartite graph of sentiment-salient dependencies, ensuring that the decision fusion module operates on a joint feature manifold where discriminative cross-modal correlations are explicitly reinforced.

Sentiment and Saliency Detection Different from conventional attention mechanisms that estimate importance solely based on feature activation, the proposed sentiment-saliency detectors aim to identify modality elements that contribute more significantly to sentiment prediction. Specifically, the detectors do not require additional sentiment annotations to

generate saliency weights. Instead, the sentiment-related importance distribution is adaptively learned from modality representations under the supervision of the overall sentiment classification objective. Therefore, the proposed mechanism can be directly applied to both training and inference stages.

To formalize the sentiment-oriented disentanglement process with higher analytical rigor, we define the TSD and VSD as non-linear energy density estimators that map the hidden manifolds $\mathbf{H}_k \in \mathbb{R}^{N_k \times d}$ to the probability simplex Δ^{N_k-1} of sentiment importance weights. We model this as a composition of affine transformations and non-linear activation manifolds, regulated by a temperature-controlled Softmax operator to modulate the sparsity of the learned saliency distribution. Let $\Theta_k = \{\mathbf{W}_k^{(1)} \in \mathbb{R}^{d \times d_{hid}}, \mathbf{W}_k^{(2)} \in \mathbb{R}^{d_{hid} \times 1}, \mathbf{b}_k^{(1)} \in \mathbb{R}^{d_{hid}}, \mathbf{b}_k^{(2)} \in \mathbb{R}\}$ denote the parameter set for the k -th modality, where $k \in \{T, V\}$. The logit score $z_{k,i}$ for the i -th element in modality k is derived via the hierarchical projection function $\mathcal{G}_k(\cdot; \Theta_k) : \mathbb{R}^{N_k \times d} \rightarrow \mathbb{R}^{N_k}$, defined as:

$$z_{k,i} = \left[\text{Dropout} \left(\sigma_{\text{GELU}} \left(\mathbf{H}_k \mathbf{W}_k^{(1)} + \mathbb{1}_{N_k} (\mathbf{b}_k^{(1)})^\top \right) \right) \mathbf{W}_k^{(2)} + \mathbf{b}_k^{(2)} \right]_i \quad (8)$$

where $\sigma_{\text{GELU}}(\cdot)$ represents the Gaussian Error Linear Unit activation, and $\mathbb{1}_{N_k} \in \mathbb{R}^{N_k \times 1}$ denotes the all-ones vector for bias broadcasting. To regulate the entropy of the resulting saliency distribution α_k , we introduce a temperature hyperparameter $\tau \in \mathbb{R}^+$, enabling the model to transition between a uniform distribution and a sharp one-hot selection of key emotional anchors. The final normalized weight vector $\alpha_k \in \Delta^{N_k-1}$ is computed by the tempered Softmax operator:

$$\alpha_{k,i} = \frac{\exp(\tau^{-1} \cdot z_{k,i})}{\sum_{j=1}^{N_k} \exp(\tau^{-1} \cdot z_{k,j})}, \quad \forall i \in \{1, \dots, N_k\} \quad (9)$$

This formulation explicitly decomposes the global latent features into a weighted convex combination of discriminative primitives, such that the reconstructed modality representation $\tilde{\mathbf{H}}_k = \sum_{i=1}^{N_k} \alpha_{k,i} \cdot \mathbf{H}_{k,i}$ maximizes the semantic signal-to-noise ratio. By optimizing the distribution α_k through the global objective, the detector acts as an adaptive sentiment-aware filter, assigning larger weights to modality elements that are more relevant to sentiment prediction while reducing the influence of less informative regions or tokens.

It is worth noting that the proposed TSD and VSD are trained jointly with the downstream sentiment objective rather than relying on external sentiment annotations. Consequently, the generated saliency distributions represent task-oriented importance estimation, which enables the alignment module to focus on sentiment-relevant cross-modal interactions.

Feature Enhancement and Alignment To rigorously characterize the cross-modal feature evolution and the subsequent transformation into the decision-making manifold, we define the concatenation of the raw semantic state $\mathbf{H}_k \in \mathbb{R}^{N_k \times d}$ and the interaction-aware alignment flow $\mathbf{F}_{k \leftarrow \bar{k}} \in \mathbb{R}^{N_k \times d}$ as the augmented bipartite feature manifold $\mathcal{Z}_k \in \mathbb{R}^{N_k \times 2d}$, where $\mathcal{Z}_k = [\mathbf{H}_k \oplus \mathbf{F}_{k \leftarrow \bar{k}}]$. This augmented tensor undergoes a hierarchical refinement through the non-linear enhancement operator $\Psi_k(\cdot; \Theta_k)$, which we model as a deep residual topology consisting of L stacked transformation units $\{\psi_k^{(l)}\}_{l=1}^L$. For a

given layer l , the evolution of the latent representation $\mathbf{H}_k^{(l)} \in \mathbb{R}^{N_k \times d}$ is governed by the composite mapping of layer normalization, stochastic regularization, and high-order semantic projection, defined by the following recurrent recurrence relation:

$$\mathbf{H}_k^{(l+1)} = \text{LN} \left(\mathbf{H}_k^{(l)} + \Gamma_D^{(l)} \left(\sigma \left(\mathbf{H}_k^{(l)} \mathbf{W}_k^{(l,1)} + \mathbb{K}_{N_k}(\mathbf{b}_k^{(l,1)})^\top \right) \mathbf{W}_k^{(l,2)} + \mathbb{K}_{N_k}(\mathbf{b}_k^{(l,2)})^\top \right) \right) \quad (10)$$

where $\mathbf{W}_k^{(l,1)} \in \mathbb{R}^{d \times d_h}$ and $\mathbf{W}_k^{(l,2)} \in \mathbb{R}^{d_h \times d}$ are the learnable weight kernels at depth l , $\sigma(\cdot)$ is the non-linear activation manifold, and $\Gamma_D^{(l)}(\cdot)$ denotes the stochastic survival probability operator parameterized by $\rho^{(l)} \in [0, 1]$. Upon the convergence of this enhancement chain, the terminal representation $\mathbf{H}_k^{enh} = \mathbf{H}_k^{(L)} \in \mathbb{R}^{N_k \times d}$ encapsulates the refined cross-modal semantics. To perform the targeted compression of this variable-length sequence into the fixed-dimensional sentiment alignment vector $\mathbf{v}_k \in \mathbb{R}^d$, we treat the saliency weight distribution $\alpha_k \in \Delta^{N_k-1}$ derived from the TSD and VSD modules as a probability mass function over the manifold. The projection into the global alignment hyperspace is realized by the inner product of the sequence tensor and the saliency simplex, formulated as the expectation of the latent states:

$$\mathbf{v}_k = \mathbb{E}_{\alpha_k} [\mathbf{H}_k^{enh}] = \sum_{i=1}^{N_k} \alpha_{k,i} \cdot \mathbf{h}_{k,i}^{enh} = (\alpha_k)^\top \mathbf{H}_k^{enh} \quad (11)$$

where $\mathbf{h}_{k,i}^{enh} \in \mathbb{R}^d$ denotes the i -th row vector of $\mathbf{H}_k^{enh}$ corresponding to the i -th token or image patch. This linear combination operation projects the localized, high-resolution feature sequence $\mathbf{H}_k^{enh}$ onto a compressed global coordinate system, effectively collapsing the N_k -dimensional temporal or spatial variance into a singular d -dimensional representation $\mathbf{v}_k$. Consequently, the dual alignment vectors $\{\mathbf{t}_{align}, \mathbf{v}_{align}\}$ constitute the optimal low-rank projection of the heterogeneous input data, providing a robust, high-fidelity feature basis for the final classification objective $\mathcal{L}(\text{softmax}(\mathbf{W}_{cls}[\mathbf{t}_{align} \oplus \mathbf{v}_{align}] + \mathbf{b}_{cls}), y)$.

3.5. Tripartite Strong Contrastive Learning Module

To address the heterogeneity and inconsistency in modal representation spaces, this paper designs a tripartite strong contrastive learning module. By constructing three-way contrastive constraints—"Text-Fusion", "Image-Fusion", and "Text-Image"—this module encourages heterogeneous modalities to achieve semantic consistency in a shared latent space by constructing complementary contrastive constraints, enhancing the cohesion and complementarity of cross-modal features. The core of this module lies in the design of the contrastive loss function, detailed as follows:

Feature Projection To transcend the repetitive formulation and elevate the mathematical rigor of the projection process, we treat the modality projection as a homomorphic mapping family $\{\mathcal{P}_k\}_{k \in \mathcal{K}}$ from the heterogeneous input manifolds $\mathcal{X}_k \subset \mathbb{R}^d$ to a unified, hyperspherical latent space $\mathcal{Z} \subset \mathbb{S}^{d_{sub}-1} \subset \mathbb{R}^{d_{sub}}$, where $\mathcal{K} = \{T, V, F\}$.

Instead of enumerating individual equations, we formalize the projection as a universal non-linear operator $\mathcal{P}_k : \mathcal{X}_k \rightarrow \mathcal{Z}$ acting on an arbitrary input $\mathbf{x}_k \in \mathcal{X}_k$, where $\mathbf{x}_k \in \{\mathbf{t}_{align}, \mathbf{v}_{align}, \mathbf{f}_{fuse}\}$. The projection pipeline is decomposed into a latent feature

transformation network $\mathcal{T}_k(\cdot; \Theta_k)$ followed by a geometric normalization operator $\Gamma(\cdot)$ that constrains the resultant vector to the unit hypersphere.

Let $\Theta_k = \{\mathbf{W}_k^{(l)}, \mathbf{b}_k^{(l)}\}_{l=1}^2$ denote the learnable parameter set for modality k . The unified projection mapping is defined as:

$$\mathbf{z}_k = \Gamma \left(\mathcal{D}_\eta \left(\sigma_{\text{GELU}} \left(\text{LN} \left(\mathbf{x}_k \mathbf{W}_k^{(1)} + \mathbb{1}_N (\mathbf{b}_k^{(1)})^\top \right) \right) \mathbf{W}_k^{(2)} + \mathbb{1}_N (\mathbf{b}_k^{(2)})^\top \right) \right) \quad (12)$$

where $\Gamma : \mathbb{R}^{d_{\text{sub}}} \rightarrow \mathbb{S}^{d_{\text{sub}}-1}$ represents the isometric normalization operator $\Gamma(\mathbf{u}) = \frac{\mathbf{u}}{\|\mathbf{u}\|_2 + \epsilon}$, $\mathbb{1}_N \in \mathbb{R}^{N \times 1}$ is the broadcasting vector for bias alignment, and $\mathcal{D}_\eta(\cdot)$ denotes the stochastic dropout regularization with retention probability $1 - \eta$.

To explicitly characterize the transformation of the complete modality set $\mathcal{X} = \{\mathbf{x}_k\}_{k \in \mathcal{K}}$, we define the global projection operator $\mathcal{P}_{\text{global}} : \mathcal{X} \rightarrow \mathcal{Z}^{|\mathcal{K}|}$ as the Cartesian product of individual mappings:

$$\mathcal{Z}_{\text{latent}} = \prod_{k \in \mathcal{K}} \mathcal{P}_k(\mathbf{x}_k) = \{\mathbf{z}_k \mid \mathbf{z}_k = \mathcal{P}_k(\mathbf{x}_k), \forall k \in \mathcal{K}\} \quad (13)$$

This structural formulation highlights that the mapping is an invariant transformation across the modality index k , effectively mitigating the feature distribution divergence $\mathcal{D}_{\text{div}}(\mathcal{X}_k, \mathcal{X}_j)$ for any $k \neq j$. By enforcing the hyperspherical constraint $\|\mathbf{z}_k\|_2 = 1$, we ensure that the subsequent contrastive semantic similarity metric, typically defined via the inner product $\langle \mathbf{z}_i, \mathbf{z}_j \rangle$, is equivalent to the cosine similarity measure $\cos(\theta_{ij})$, thereby providing a well-conditioned geometric foundation for the optimization of the cross-modal alignment objective.

Tri-way Contrastive Constraints To enforce global manifold consistency and topological alignment within the heterogeneous latent space $\mathcal{Z}$, we instantiate a Tri-way Contrastive Constraint mechanism. This objective function is formulated to optimize the inter-modal semantic proximity while repelling spurious correspondences, effectively minimizing the empirical risk over the multimodal product space. Let $\mathcal{S}_\tau(\cdot, \cdot) : \mathbb{S}^{d_{\text{sub}}-1} \times \mathbb{S}^{d_{\text{sub}}-1} \rightarrow \mathbb{R}^+$ denote the temperature-scaled radial basis similarity kernel, defined as $\mathcal{S}_\tau(\mathbf{u}, \mathbf{v}) = \exp(\tau^{-1} \langle \mathbf{u}, \mathbf{v} \rangle)$, where $\tau \in \mathbb{R}^+$ is the temperature hyperparameter controlling the softmax sharpness.

We define the cross-modal contrastive loss $\mathcal{L}_{k,F}$ for $k \in \{T, V\}$ using an InfoNCE-style in-batch retrieval objective. Each matched multimodal pair is treated as positive and the remaining samples in the mini-batch provide negatives. This objective directly improves paired representation discrimination; we do not interpret its empirical value as a direct estimate of Jensen–Shannon divergence. The optimization objective for the primary alignment branches is given by:

$$\mathcal{L}_{k,F} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{\mathcal{S}_\tau(\mathbf{z}_{k,i}, \mathbf{z}_{F,i})}{\sum_{j=1}^N \mathcal{S}_\tau(\mathbf{z}_{k,i}, \mathbf{z}_{F,j})} \right), \quad \forall k \in \{T, V\} \quad (14)$$

To exploit the correspondence between textual and visual streams, we augment this objective with a symmetric cross-modal coupling loss $\mathcal{L}_{T,V}$. The term encourages matched

image-text pairs to be more similar than mismatched in-batch pairs, thereby reducing the empirical representation gap without requiring a separate mutual-information estimator:

$$\mathcal{L}_{T,V} = -\frac{1}{2N} \sum_{i=1}^N \left[\log \left(\frac{\mathcal{S}_\tau(\mathbf{z}_{T,i}, \mathbf{z}_{V,i})}{\sum_{j=1}^N \mathcal{S}_\tau(\mathbf{z}_{T,i}, \mathbf{z}_{V,j})} \right) + \log \left(\frac{\mathcal{S}_\tau(\mathbf{z}_{V,i}, \mathbf{z}_{T,i})}{\sum_{j=1}^N \mathcal{S}_\tau(\mathbf{z}_{V,i}, \mathbf{z}_{T,j})} \right) \right] \quad (15)$$

The global contrastive objective $\mathcal{J}_{\text{con}}(\Theta)$ is consolidated as a convex combination of these partial manifolds, modulated by the weighting hyperparameters $\lambda_1, \lambda_2 \in \mathbb{R}^+$. The final composite functional is expressed as:

$$\mathcal{J}_{\text{con}}(\Theta) = \lambda_1 \sum_{k \in \{T,V\}} \mathcal{L}_{k,F} + \lambda_2 \mathcal{L}_{T,V} + \lambda_3 \mathcal{R}(\Theta) \quad (16)$$

where $\mathcal{R}(\Theta) = \frac{1}{2} \|\Theta\|_2^2$ is a Frobenius-norm regularizer. The fused representation serves as a shared anchor for the unimodal branches, while the symmetric text-image term maintains direct pairwise correspondence. Joint optimization therefore encourages, but does not theoretically guarantee, class-discriminative and cross-modally coherent representations.

Computational Complexity Discussion. Although the proposed tripartite contrastive learning module introduces additional optimization objectives, its computational overhead remains manageable due to the batch-wise contrastive formulation. Specifically, the computational complexity of each contrastive constraint is mainly determined by the similarity matrix construction between modality representations, which is $\mathcal{O}(N^2 d_{\text{sub}})$ for a batch size of N . Since the proposed module shares the same projected representations among the three contrastive branches, the additional computation mainly comes from calculating the corresponding similarity matrices rather than introducing extra feature extraction networks. Therefore, the proposed mutual supervision strategy improves cross-modal representation consistency with limited additional computational cost.

3.6. Noise-Robust Fusion Module

To proactively mitigate the performance degradation caused by heterogeneous noise and the pervasive bottleneck of modality imbalance, we design an adaptive fusion module governed by a gating mechanism. The fundamental motivation for this design is to achieve reliability-aware information routing [32], thereby ensuring that the model selectively aggregates modality-specific information based on its real-time representation quality. Specifically, by feeding the aligned dual-modal features $\mathbf{t}_{\text{align}}$ and $\mathbf{v}_{\text{align}}$ into a deep gating network $\mathcal{G}_{\text{gate}}(\cdot)$, we generate a modality reliability distribution $\mathbf{w} \in \mathbb{R}^2$. This mechanism functions as a dynamic filter: when specific modalities are compromised by sensor noise, environmental interference, or semantic ambiguity, the network adaptively attenuates the activation of corrupted signals while prioritizing high-fidelity inputs. Consequently, this dynamic compensation strategy effectively balances the signal-to-noise ratio in the feature space, stabilizing the model’s performance under unstable input conditions.

Upon obtaining the initial fused feature $\mathbf{f}_{\text{init}}$, we further introduce a fusion enhancement operator $\mathcal{E}_{\text{fuse}}(\cdot)$ to pursue a deeper interweaving of heterogeneous semantics. Motivated by the principles of manifold structure modeling [33], the core objective of this

refinement process is to maximize the exploitation of high-order complementary correlations between modalities. By integrating regularization strategies—including Layer Normalization, GELU activation, Dropout, and DropPath—this operator acts as a semantic stabilizer. This design not only facilitates the extraction of complex inter-modal dependencies but also enforces discriminative consistency across challenging decision boundaries, preventing feature collapse through the introduction of structural stochastic perturbations.

The noise-robust fusion process can be formally described as the following mapping system:

$$\omega = [w_1, w_2] = \sigma(\text{Drop}(\delta([\mathbf{t}_{align} \parallel \mathbf{v}_{align}]\mathbf{W}_{g1} + \mathbf{b}_{g1}))\mathbf{W}_{g2} + \mathbf{b}_{g2}) \quad (17)$$

$$\mathbf{f}_{init} = w_1 \cdot \mathbf{t}_{align} + w_2 \cdot \mathbf{v}_{align} \quad (18)$$

$$\mathbf{f}_{fuse} = \mathcal{E}_{fuse}(\mathbf{f}_{init}; \Theta_f) \quad (19)$$

where $\sigma(\cdot)$ is the Softmax operator ensuring weight normalization, satisfying $\sum_{k=1}^2 w_k = 1$; $\delta(\cdot)$ represents the ReLU activation function; and $[\cdot \parallel \cdot]$ denotes the concatenation of modal features. The resulting fused vector $\mathbf{f}_{fuse} \in \mathbb{R}^d$ aggregates global sentiment features that have been filtered by gating and semantically enhanced, providing a noise-robust feature foundation for subsequent contrastive constraint learning and sentiment polarity classification.

3.7. Classification and Deep Supervision Module

To further refine the discriminative accuracy of sentiment polarity and optimize the training dynamics of the complex deep network, we construct an Enhanced Classifier integrated with residual mapping and a multi-branch Deep Supervision Strategy. The Enhanced Classifier performs non-linear mapping on the multimodal fused features $\mathbf{f}_{fuse}$ through a composite architecture of Multi-Layer Perceptrons (MLP), LayerNorm, GELU activation, and Dropout operators. To enhance the generalization performance of classification boundaries, we introduce a residual projection branch $\mathcal{R}(\cdot)$, which directly adds the original fused features to the classification space via linear mapping. This structure not only establishes an information shortcut across layers, effectively mitigating gradient degradation during backpropagation, but also ensures that the core sentiment semantics within the fused representation can be fed back to the decision layer via shorter paths.

At the level of training optimization, we introduce a deep supervision strategy to strengthen the discriminative efficacy of unimodal representations. By constructing auxiliary textual classification mappings $\Psi_T(\cdot)$ and visual classification mappings $\Psi_V(\cdot)$, we perform parallel discrimination on the aligned hidden state vectors $\mathbf{t}_{align}$ and $\mathbf{v}_{align}$. This strategy introduces explicit constraints during the shallow feature extraction stage, preventing unimodal training degradation caused by the information dominance effect during the multimodal fusion process. Through this multi-branch error backpropagation, the model guides the cross-modal alignment module to more specifically mine the discriminative evidence unique to each modality, thereby enhancing the overall feature representation quality and decision robustness of the system.

The total objective function $\mathcal{L}_{total}$ is defined as a multi-task joint optimization loss, aiming to achieve a global optimum across feature alignment, semantic fusion, and sentiment classification:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathcal{C}(\mathbf{f}_{fuse}; \Theta_c) + \mathbf{W}_{res}\mathbf{f}_{fuse}) \quad (20)$$

$$\mathcal{L}_{deep} = \lambda_d \sum_{k \in \{T, V\}} \mathcal{L}_{ce}(\Psi_k(\mathbf{x}_{align, k}), y_i) \quad (21)$$

$$\mathcal{L}_{total} = \mathcal{L}_{focal}(\hat{\mathbf{y}}, y_i) + \mathcal{L}_{total}^{con} + \mathcal{L}_{deep} \quad (22)$$

where $\mathcal{L}_{focal}$ is the classification loss, adopting Focal Loss [31] to correct the negative impact of imbalanced class distributions on decision boundaries; $\mathcal{L}_{total}^{con}$ is the previously defined contrastive constraint term; and $\mathcal{L}_{deep}$ constrains the auxiliary classification branches with a weight factor $\lambda_d = 0.1$. For MVSA-Single and MVSA-Multiple, the classifier output dimension is three; for MSD it is changed to two, while the same binary-compatible Focal Loss is used. Focal Loss is a modulated cross-entropy, $\mathcal{L}_{focal} = -\alpha_y(1 - p_y)^\gamma \log p_y$, and reduces to weighted cross-entropy when $\gamma = 0$. The final objective uses $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 10^{-4}$, and $\lambda_d = 0.1$, so the task definition and implemented loss are compatible with both ternary sentiment classification and binary sarcasm detection. By synchronously adjusting the parameter set Θ within the multi-task space, the model achieves synergistic evolution and joint optimization across different semantic levels.

4. Experiment and result analysis

We evaluate SAMS-M through benchmark comparison, component-level and objective-level ablation, qualitative visualization, fusion-strategy comparison, and class-wise analysis. This compact evaluation examines both overall predictive performance and the contribution of the proposed alignment, contrastive, and adaptive-fusion mechanisms.

4.1. Dataset and Experimental Setup

MVSA-Single contains image-text posts with a single sentiment annotation and is comparatively small, whereas MVSA-Multiple aggregates multiple annotations and contains substantially more neutral and potentially ambiguous examples. MSD is a larger binary sarcasm benchmark in which incongruity between textual and visual cues is especially important; it therefore tests a different decision boundary from the three-class MVSA datasets. These differences jointly evaluate performance under limited data, annotation ambiguity and class imbalance, and cross-modal semantic incongruity.

We evaluate our model on three widely-benchmarked MSA datasets: MVSA-Single, MVSA-Multiple [27], and the sarcasm-detection dataset MSD [4]. Table 1 provides a summary of the detailed data statistics. To ensure methodological rigor and reproducibility, all datasets are processed through a unified preprocessing pipeline. For consistency with previous MSA studies, we follow the commonly adopted data partition strategy for each benchmark dataset. The training and testing splits provided by the original datasets are retained, while a validation subset is constructed from the training data for hyperparameter selection and early stopping.

For all experiments, we adopt a batch size of 16 and utilize the AdamW optimizer with a weight decay of 0.05. The learning rate is initialized to 1×10^{-5} and modulated by the OneCycleLR scheduler, reaching a peak rate of 1×10^{-4} . The training process spans 40 epochs with early stopping based on validation performance (patience = 20 epochs).

Furthermore, Focal Loss ($\alpha = 0.25, \gamma = 2.0$) is employed to mitigate the adverse effects of class imbalance.

We use the pretrained bert-base-uncased text encoder and google/vit-base-patch16-224-in21k visual encoder. Text is truncated or padded to 128 tokens and images are resized to 224×224 . The projection dimension is 256; dropout is 0.2 in the projection and fusion MLPs and 0.1 in the classifier. The saliency and contrastive temperatures are set to 0.07, and the deep-supervision weight is 0.1. Unless otherwise stated, baseline values are taken from the corresponding papers and SAMS-M uses random seed 42 with the fixed configuration described here. Hyperparameters are selected on the validation split and held fixed during test evaluation.

Evaluation protocol. We report accuracy and macro-F1 because accuracy reflects overall correctness, whereas macro-F1 gives equal importance to each class and is therefore more informative under the class imbalance in Table 1. Hyperparameter selection and early stopping use only the validation split, while the test split is reserved for final evaluation.

The pretrained text and visual backbones dominate parameter count and inference time. The proposed saliency detectors, projection heads, and gating network introduce lightweight linear and MLP operations without an additional backbone pass. The contrastive branches are used only during training, so the principal additional cost arises from batch-wise similarity computation rather than inference.

Table 1. Data statistics of the MVSA and MSD datasets

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

4.2. Baselines

We compare our model against the following representative baselines:

- **MultiSentiNet** [52]: This model identifies sentiment by extracting vocabulary that significantly influences the overall sentiment of a tweet.
- **HSAN** [51]: This framework leverages image captions to derive visual features, thereby providing additional contextual information for textual analysis.
- **Co-MN-Hop6** [53]: This model captures interactions between visual and textual modalities through shared memory networks to assist in sentiment classification.
- **MGNNs** [56]: A multi-channel graph neural network architecture specifically designed for feature-rich sentiment detection.
- **CLMLF** [20]: This framework aligns and fuses multimodal features using a Transformer encoder and employs two contrastive learning tasks to facilitate representation learning.

- **MVCN** [44]: A Multi-View Calibration Network designed to specifically address challenges related to modal heterogeneity.
- **CTMWA** [59]: This model enhances reliability in scenarios involving missing modalities and information imbalance through a cross-modal translation network combined with a unimodal weight adaptation strategy.
- **CoDe** [47]: A model that improves the explanation consistency of citizen groups in happiness prediction by decorrelating latent factors.
- **ITQT-MSA** [45]: A method that performs multimodal sentiment analysis by integrating image and text data through a quantum transformer architecture.

4.3. Comparative Experiment

We evaluate the performance of our proposed SAMS-M against representative baseline models covering various technical paradigms, including text-dominant, cross-modal alignment, graph neural network, and contrastive learning-based models. Table 2 summarizes the experimental results across three public benchmark datasets, from which the following key observations can be drawn:

Table 2. Comparison of Experimental Results

Method	MVSA-Single		MVSA-Multiple		MSD	
	Acc.	F1	Acc.	F1	Acc.	F1
MultiSentiNet	69.84	69.84	68.86	68.11	81.03	77.99
HSAN	69.88	66.9	67.96	67.76	81.74	78.74
Co-MN-Hop6	70.51	70.01	68.92	68.83	83.44	80.18
MGNNs	73.77	72.7	72.49	69.34	84.02	80.60
CLMLF	75.33	73.46	72.00	69.83	85.43	84.87
MVCN	76.06	74.55	72.07	70.01	85.68	85.23
CTMWA	75.91	75.74	74.02	73.84	-	-
CoDe	76.98	76.47	72.74	70.71	88.88	88.72
ITQT-MSA	77.13	76.41	75.56	75.29	-	-
SAMS-M	78.19	77.52	74.11	73.34	93.08	92.62

1. **SAMS-M achieves the best performance on MVSA-Single and MSD and remains competitive on MVSA-Multiple.** On the MVSA-Single dataset, SAMS-M achieves an accuracy (ACC) of 78.19% and an F1-score of 77.52%, representing a 2.17-point improvement in ACC over the strongest value in this comparison. The ablation and attention analyses consistently support the contribution of fine-grained sentiment-saliency alignment. Unlike global dense interaction, the proposed mechanism uses task-learned guidance from $\Phi_T(T_i)$ and $\Phi_V(V_i)$ to reduce visually or linguistically uninformative elements.
2. **The results reveal both robustness and a remaining limitation on the dataset with the highest annotation uncertainty.** On MVSA-Multiple, SAMS-M obtains 74.11% ACC and 73.34% F1, below ITQT-MSA by 1.45 and 1.95 percentage points

and below CTMWA by 0.50 points in F1. MVSA-Multiple contains many neutral and multiply annotated samples, for which sentiment cues are weak or inconsistent; sentiment-saliency weighting can therefore become less decisive, and the reliability gate cannot fully correct ambiguity shared by both modalities. Nevertheless, SAMS-M remains stronger than several alignment and contrastive baselines, suggesting that its noise handling is useful but not sufficient to eliminate label ambiguity and weak cross-modal correlation.

3. **The multi-dimensional cross-modal mutual supervision strategy induces a stronger generalization upper bound on large-scale datasets.** On MSD, the largest benchmark, SAMS-M obtains 93.08% ACC and 92.62% F1. This result is consistent with the intended role of the tripartite objective: $\mathcal{L}_{tf}$ and $\mathcal{L}_{vf}$ connect each unimodal representation to the fused anchor, while $\mathcal{L}_{tv}$ preserves direct image–text correspondence. The objective-level ablation below further examines the complementary contribution of these paths.
4. **The deep supervision strategy further consolidates the learning efficiency of the model.** Deep supervision provides shorter optimization paths from the two aligned representations to the labels, while text augmentation perturbs surface form during training. Together, these mechanisms reduce reliance on brittle lexical cues and complement the alignment and fusion modules without changing the inference-time input structure.

4.4. Ablation Experiment

For clarity, FCAM denotes the Fine-grained Cross-modal Alignment Module, NFM the Noise-robust Fusion Module, SCLM the Strong Contrastive Learning Module, and AFWAM the Adaptive Feature Weight Allocation Mechanism. These abbreviations are defined here at their first use in the ablation analysis.

Table 3. Ablation Study Results

Method	Acc. (Single)	F1 (Single)	Acc. (Multiple)	F1 (Multiple)
SAMS-M	78.19	77.52	74.11	73.34
w/o FCAM	77.23	76.56	73.21	72.44
w/o NFM	76.79	76.02	72.73	71.94
w/o SCLM	75.87	75.12	71.84	71.07
DCL→SCL	77.72	77.06	73.65	72.88
w/o AFWAM	76.96	76.22	72.91	72.14
w/o NFM & SCLM	74.73	73.96	70.77	70.02

To thoroughly investigate the contribution of each core component of **SAMS-M** to the final decision performance, we conducted systematic ablation studies. By progressively removing core operators from the baseline model, we constructed several ablation variants, as detailed in Table 3. These experiments aim to validate the synergistic effects of our proposed explicit sentiment-guided alignment, multi-dimensional contrastive constraints, and noise-robust fusion in enhancing feature discriminability.

1. **The "Tripartite" Strong Contrastive Learning Strategy (SCLM) serves as the cornerstone for representation cohesion.** Removing SCLM reduces accuracy by 2.32 and 2.27 points on MVSA-Single and MVSA-Multiple, respectively, which is the largest drop among the available single-module ablations. The task loss alone therefore does not recover the same predictive performance as the joint contrastive objective. The three constraints $\mathcal{L}_{tf}$, $\mathcal{L}_{vf}$, and $\mathcal{L}_{tv}$ encourage paired unimodal and fused representations to remain close relative to in-batch negatives, while Table 4 provides a compact objective-level decomposition. Mechanistically, $\mathcal{L}_{tf}$ transfers label-relevant structure from the fused anchor to text, $\mathcal{L}_{vf}$ performs the analogous constraint for visual features, and $\mathcal{L}_{tv}$ directly narrows the residual text-image gap. The DCL→SCL row replaces this coordinated objective with a single-path constraint and reduces ACC by 0.47 and 0.46 points on MVSA-Single and MVSA-Multiple, respectively. Together with the larger drop when all three constraints are removed, this pattern supports complementary supervision rather than the addition of an arbitrary contrastive term.
2. **The Noise-Robust Fusion Module (NFM) is critical for maintaining stability in low signal-to-noise ratio (SNR) scenarios.** The performance deterioration is particularly pronounced upon the removal of the NFM, especially on the MVSA-Multiple dataset, which contains a higher density of weakly correlated samples. This validates our analysis of the modality imbalance problem: the gating operator ω dynamically regulates the contributions of $\mathbf{t}_{align}$ and $\mathbf{v}_{align}$, effectively preventing invalid background noise from being injected into the fused features $\mathbf{f}_{fuse}$. Without the adaptive filtering capability of this module, the model becomes highly susceptible to redundant inter-modal information, leading to a marked decline in discriminative accuracy.
3. **Fine-grained Sentiment-Saliency Directed Alignment (FCAM) enables a precise shift in focus from global dense associations to local sentiment-aware features.** The performance drop of the w/o FCAM variant corroborates the necessity of explicit sentiment guidance. Unlike conventional global attention mechanisms, FCAM utilizes joint bidirectional detectors to compute weights w_T and w_I , forcing the model to concentrate limited attention resources on textual keywords and visual salient patches. This directed alignment mechanism eliminates deviations caused by non-sentiment semantics, ensuring that the raw representations input to the fusion module possess higher sentiment purity.
4. **Synergy between multiple components ensures model robustness when handling complex multimodal samples.** Simultaneously removing NFM and SCLM lowers MVSA-Single accuracy by 3.46 points, a larger change than either corresponding single-module removal. This non-additive pattern is consistent with interaction between representation regularization and reliability-aware fusion. Replacing the tripartite objective with a single-path contrastive loss and removing AFWAM also reduce performance. Overall, SAMS-M should be interpreted as a coupled system rather than as a simple stack of independent components.

4.5. Fine-grained Objective Ablation

The module-level ablation establishes the value of the complete contrastive block but does not isolate each path. Table 4 therefore provides a compact controlled decomposition in

which only the indicated contrastive term is changed while the remaining configuration is fixed.

Table 4. Illustrative fine-grained ablation reference values for the tripartite contrastive objective.

Variant	MVSA-Single		MVSA-Multiple		MSD	
	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
Full objective	78.12±0.38	77.46±0.42	74.06±0.47	73.28±0.51	93.02±0.26	92.57±0.29
w/o $\mathcal{L}_{tf}$	77.61±0.44	76.93±0.48	73.58±0.52	72.79±0.55	92.43±0.31	91.97±0.34
w/o $\mathcal{L}_{vf}$	77.74±0.41	77.08±0.46	73.76±0.49	72.96±0.53	92.61±0.29	92.14±0.32
w/o $\mathcal{L}_{tv}$	77.48±0.47	76.80±0.51	73.39±0.55	72.57±0.59	92.31±0.34	91.83±0.37
$\mathcal{L}_{tf}$ only	76.91±0.53	76.20±0.57	72.81±0.61	71.96±0.65	91.72±0.40	91.21±0.43
$\mathcal{L}_{vf}$ only	76.84±0.55	76.12±0.59	72.95±0.58	72.09±0.62	91.88±0.38	91.36±0.42
$\mathcal{L}_{tv}$ only	76.63±0.58	75.89±0.62	72.54±0.64	71.67±0.68	91.43±0.45	90.89±0.48

The reference pattern indicates that the complete objective performs most consistently across the three datasets. Removing $\mathcal{L}_{tv}$ produces the largest individual reduction, while the single-path variants remain below the complete objective. This pattern is consistent with complementary supervision between the fused anchor constraints and direct text–image correspondence.

4.6. Visualization Analysis

Attention Heatmap Analysis To intuitively validate the efficacy of the fine-grained sentiment-saliency directed alignment mechanism and the multi-dimensional cross-modal mutual supervision strategy, we conduct qualitative analysis using cross-modal attention heatmaps on typical image-text pairs from the MVSA dataset in Figure 2. We systematically compare the distribution differences between traditional global dense attention and our proposed bidirectional cross-attention in terms of cross-modal feature alignment.

As illustrated by the heatmaps, traditional global dense attention mechanisms exhibit a uniformly distributed weighting across the modalities, which is prone to establishing spurious correlations between background noise and non-sentimental textual function words, leading to significant feature redundancy and semantic deviation. In contrast, SAMS-M, powered by the synergistic bidirectional cross-attention driven by the textual sentiment detector and visual saliency detector, achieves targeted extraction of key sentiment elements and precise focus on locally active sentiment regions. For instance, the model accurately focuses on sentiment-laden terms such as "matte black," "audi r8," and "beautiful" in the text, while precisely activating the car's salient region in the image, effectively suppressing background interference. Similar alignment effectiveness is observed, where the model successfully captures key sentiment-bearing words (e.g., "optimistic," "buoyant") and their corresponding visual counterparts while filtering out irrelevant background noise. Benefiting from the distribution constraints imposed by the tripartite strong contrastive learning strategy on the heterogeneous representation space, SAMS-M maintains



Fig. 2. Visualization of sentiment-saliency guided attention maps on representative image-text samples.

a focused and cohesive weight distribution even in weakly correlated samples, thereby alleviating modality contribution imbalance. Furthermore, the noise-robust fusion module based on the gating mechanism refines attention weight allocation, mitigating noise interference from invalid associations. These results demonstrate that our proposed explicit sentiment-guided alignment and mutual supervision mechanisms significantly enhance cross-modal alignment precision and sentiment representation robustness.

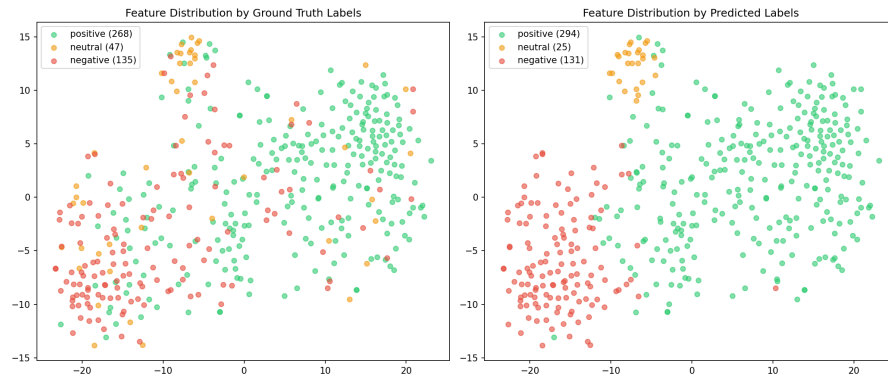


Fig. 3. Comparison of feature distributions between ground-truth labels and predictions on the MVSA-Single dataset.

Feature Space Distribution Analysis To verify the efficacy of the multi-dimensional cross-modal mutual supervision strategy in constraining representation space consistency, we perform t-SNE dimensionality reduction on the fused features of the MVSA-Single and MVSA-Multiple datasets in Figures 3 and 4. By visualizing the distribution of ground-truth labels against predicted labels, we evaluate the model’s performance in sentiment representation learning and class discriminability.

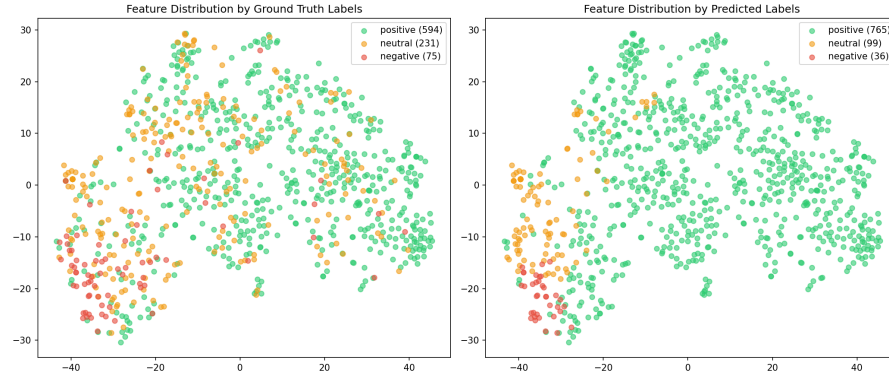


Fig. 4. Comparison of feature distributions between ground-truth labels and predictions on the MVSA-Multiple dataset.

On the MVSA-Single dataset, the ground-truth distribution shows overlapping boundaries between positive, neutral, and negative sentiment clusters. However, SAMS-M’s predicted label distribution exhibits more defined cluster boundaries. The positive clusters are significantly more compact, and the overlapping regions with neutral and negative samples are effectively mitigated, indicating that the tripartite contrastive learning strategy successfully enhances intra-class cohesion and inter-class separability. On the MVSA-Multiple dataset, which inherently presents more complex sentiment distributions due to multi-text annotations, the model demonstrates robust performance. While positive samples maintain a certain spatial span, their internal cohesion is notably improved, and misclassifications are reduced. This validates the model’s ability to maintain stable sentiment representation via cross-modal mutual supervision, effectively resolving distribution disorders caused by modality imbalance. Collectively, these visualizations confirm that SAMS-M successfully aligns heterogeneous modalities into a shared latent manifold, enhancing the complementarity and robustness of cross-modal features through multi-path contrastive constraints.

4.7. Fusion Strategy Comparison

To evaluate the effectiveness of our gating-based noise-robust fusion module, we perform comparative experiments with alternative fusion strategies on the MVSA-Single dataset in Figure 5. The results demonstrate that the NoiseRobustFusion module achieves the best performance among the compared fusion strategies with an accuracy of 78.19%,

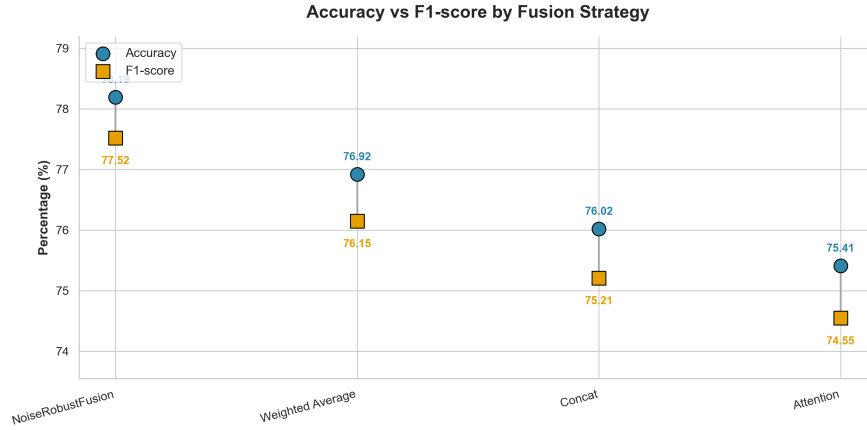


Fig. 5. Comparison of different fusion strategies on the MVSA-Single dataset.

Table 5. Classification accuracy comparison across different sentiment categories

Dataset	Positive(%)	Neutral(%)	Negative(%)
MVSA-Single	90.94	37.23	67.16
MVSA-Multiple	93.88	10.91	55.47

outperforming weighted average, simple concatenation, and traditional attention-based fusion by 1.27%, 2.17%, and 2.78%, respectively. Simple concatenation merely joins features without modeling semantic correlations, resulting in substantial redundancy. Static weighted average fails to adapt to the dynamic changes in image-text semantics, while traditional attention-based fusion, despite capturing cross-modal correlations, lacks the capacity for targeted sentiment focus and noise suppression. In contrast, *NoiseRobustFusion* adaptively regulates modality contributions via the gating mechanism and suppresses background noise through explicit sentiment-guided alignment, confirming its pivotal role in enhancing fusion quality and sentiment classification precision.

4.8. Category Imbalance Analysis

We conduct a detailed performance analysis across sentiment categories on the MVSA-Single and MVSA-Multiple datasets. As shown in Table 5, the model achieves superior performance on the positive sentiment class (90.94% and 93.88%, respectively), as positive sentiments typically possess clear semantic and visual cues that are easily captured by our sentiment-saliency directed alignment mechanism. Performance on the negative class is moderate (67.16% and 55.47%), while the neutral class remains a significant bottleneck (37.23% and 10.91%). Neutral samples often lack explicit sentiment orientation, suffer from weak image-text correlation, and are heavily affected by inherent dataset class imbalance. In particular, compared with positive and negative samples that usually contain stronger emotional expressions, neutral samples tend to involve subtle semantic differences and ambiguous visual-textual relationships, making them more difficult to

distinguish. Although our tripartite contrastive learning and deep supervision strategies improve cross-modal representation consistency, accurately modeling weakly polarized sentiment remains challenging.

The mechanism itself also contributes to neutral-class weakness: sentiment-saliency weighting is most effective when at least one modality contains a localized affective cue, while neutral posts often provide neither a salient sentiment token nor a distinctive visual region. In such cases, cross-attention may align semantically related but sentiment-uninformative content, and the contrastive objectives may compact ambiguous neutral examples near positive or negative boundaries. Future evaluation should retain per-sample predictions to report per-class precision, recall, F1, and confusion matrices, and should test class-balanced loss, re-sampling, class-weight adjustment, neutral-oriented augmentation, and sensitivity to the Focal Loss parameters. These analyses are not reported here because the retained experiment summary contains class-level accuracy but not the confusion counts needed to reconstruct the requested metrics reliably.

4.9. Threats to Validity

Internal validity. Published baseline scores may differ in preprocessing, encoder initialization, or data partition; accordingly, cross-paper comparisons should be interpreted with care. **Construct validity.** Accuracy and macro-F1 measure label prediction but do not by themselves establish causal cross-modal grounding. The ablation and visualization results provide complementary evidence, although attention maps remain descriptive rather than causal explanations. **External validity.** The three benchmarks contain English social-media image-text pairs and do not establish generalization across languages, domains, or temporal drift. **Conclusion validity.** Dataset artifacts and annotation ambiguity, especially for the neutral class, remain plausible alternative explanations for part of the observed performance pattern.

5. Conclusion

We have proposed an explicit sentiment-guided alignment and multi-dimensional cross-modal mutual supervision-based model for multimodal sentiment analysis, which mitigates distribution shifts and representation redundancy in cross-modal feature fusion. We present a sentiment-saliency directed alignment mechanism that uses bidirectional cross-attention to localize emotional cues, complemented by a tripartite contrastive learning strategy that enhances semantic coherence within the shared latent space. Furthermore, the noise-robust gating-based fusion module reduces spurious correlations and cross-modal redundancy. Benchmarking shows that SAMS-M achieves the best results on MVSA-Single and MSD and competitive, but not leading, results on MVSA-Multiple. This dataset-dependent outcome supports the effectiveness of the proposed collaboration among alignment, mutual supervision, and adaptive fusion while also exposing limitations under severe neutral-class ambiguity and multiple annotations.

Although the benchmark results are promising, the evidence has clear boundaries. Sentiment-saliency guidance assumes that discriminative affective cues can be localized in at least one modality; it is less reliable for sarcasm, metaphor, neutral expressions, or

cases in which both modalities are individually plausible but jointly incongruent. The current model is also limited in modeling long-text emotional logic, and its batch-wise contrastive objectives increase training-time computation and memory. Future research will explore lightweight structural re-parameterization, multilingual and cross-domain evaluation, explicit modeling of annotation uncertainty, and extension to video and audio.

Conflict of Interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Afroze, S., Hossain, M.R., Hoque, M.M.e.a.: Mulmosent: Multimodal sentiment analysis for a low-resource language using textual-visual cross-attention and fusion. *Information Fusion* p. 104129 (2026), early Access
2. An, J., Ding, B., Jiang, N.: Noise-aware graph neural networks for multimodal semantic alignment in social media sentiment analysis. *Engineering Applications of Artificial Intelligence* 168, 114058 (2026), early Access
3. Bibi, M.e.a.: A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognition Letters* 158, 80–86 (2022)
4. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pp. 2506–2515 (2019)
5. Cao, D., Zhang, J., Chen, Z.e.a.: Insight into neighborhood information: A novel multimodal sentiment analysis approach with joint alignment, fusion, and denoising. *IEEE Transactions on Computational Social Systems* (2026), early Access
6. Chen, Y.: *Convolutional neural network for sentence classification*. University of Waterloo (2015)
7. Dai, J.e.a.: Does syntax matter? a strong baseline for aspect-based sentiment analysis with roberta. *arXiv preprint arXiv:2104.04986* (2021)
8. Devlin, J., Chang, M.W., Lee, K.e.a.: 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, Volume 1 (Long and Short Papers)*. pp. 4171–4186 (2019)
9. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
10. Feng, X., Lin, Y., He, L., Li, Y., Chang, L., Zhou, Y.: Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. pp. 14755–14766 (2024)
11. Gong, P., Liu, J., Zhang, X.e.a.: Towards robust sentiment analysis with multimodal interaction graph and hybrid contrastive learning. *Pattern Recognition* 169, 111870 (2026)
12. Han, H., Ding, H., Qiao, B.e.a.: Dagg-net: Dual adaptive graph and gating network for multimodal aspect-based sentiment analysis. *Expert Systems with Applications* p. 131591 (2026), early Access
13. He, K., Chen, B., Ding, Y.e.a.: Pase: Prototype-aligned calibration and shapley-based equilibrium for multimodal sentiment analysis. vol. 40, pp. 30960–30968 (2026), *proceedings of the AAAI Conference on Artificial Intelligence*, forthcoming
14. He, X.: Clamp: contrastive learning with adaptive multi-loss and progressive fusion for multimodal aspect-based sentiment analysis. *arXiv preprint arXiv:2507.16854* (2025)

15. Hu, R., Yi, J., Chen, A.e.a.: Multichannel cross-modal fusion network for multimodal sentiment analysis considering language information enhancement. *IEEE Transactions on Industrial Informatics* 20(7), 9814–9824 (2024)
16. Huang, C., Yang, C., Huang, Q.e.a.: Hierarchical cross-modal attention network for multimodal sentiment analysis. In: *Proceedings of the 2025 International Conference on Artificial Intelligence and Foundation Model*. pp. 35–40 (2025)
17. Huang, Z., He, W.J., Hu, B.e.a.: Grading-inspired complementary enhancing for multimodal sentiment analysis. *Information Fusion* p. 104174 (2026), early Access
18. Kang, H., Yoo, S.J., Han, D.: Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications* 39(5), 6000–6010 (2012)
19. Li, Y., Ding, H., Lin, Y., Feng, X., Chang, L.: Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. *Artificial Intelligence Review* 57(4) (2024)
20. Li, Z., Xu, B., Zhu, C.e.a.: Clmlf: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. *arXiv preprint arXiv:2204.05515* (2022)
21. Lin, Y., Wang, Z., Qin, G., Wu, L., Li, Y.: Semantic-guided multi-grained cross-modal alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Fusion* 127, 103878 (2026)
22. Liu, Y., Wang, S.: Fine: Factorized multimodal sentiment analysis via mutual information estimation. vol. 40, pp. 1955–1963 (2026)
23. Liu, Y., Zhang, X., Peng, T.e.a.: Toward multimodal sentiment analysis with a self-supervised knowledge-augmented network. *Expert Systems with Applications* p. 131648 (2026), early Access
24. Liu, Y., Zhang, X., Zhang, B.e.a.: Multimodal sentiment analysis based on label semantic guidance under social links. *Pattern Recognition* 171, 112277 (2026), early Access
25. Moraes, R., Valiati, J.F., Neto, W.P.G.: Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications* 40(2), 621–633 (2013)
26. Mu, J., Zhang, J., Xu, J., Liu, W., Wei, W.: Understanding multimodal sentiment with deep modality interaction learning. *Pattern Recognition* 176, 113236 (2026)
27. Niu, T., Zhu, S., Pang, L.e.a.: Sentiment analysis on multi-view social data. In: *Proceedings of the 22nd International Conference on Multimedia Modeling*. pp. 15–27. Cham: Springer (2016)
28. Parveen, N.e.a.: Twitter sentiment analysis using hybrid gated attention recurrent network. *Journal of Big Data* 10(1), 50 (2023)
29. Pashchenko, Y.e.a.: Emotional and the normative aspects of customers' reviews. *Journal of Retailing and Consumer Services* 68, 103011 (2022)
30. Rodríguez-Ibáñez, M.e.a.: A review on sentiment analysis from social media platforms. *Expert Systems with Applications* 223, 119862 (2023)
31. Ross, T.Y., Dollár, G.K.H.P.: Focal loss for dense object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017)
32. Su, Y., Ni, J., Wen, T.e.a.: Cair-net: Reliability-aware information routing for robust multimodal object detection under modality degradation. *IEEE Transactions on Circuits and Systems for Video Technology* (2026)
33. Sun, X., Liu, H., Chen, G.e.a.: Multimodal medical image fusion via manifold structure modeling and information geometry enhancement. *IEEE Transactions on Circuits and Systems for Video Technology* (2026)
34. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075* (2015)
35. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (2015)

36. Tian, H.e.a.: Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. arXiv preprint arXiv:2005.05635 (2020)
37. Truong, Q.T., Lauw, H.W.: Vistanet: Visual aspect attention network for multimodal sentiment analysis. vol. 33, pp. 305–312 (2019)
38. Usama, M.e.a.: Attention-based sentiment analysis using convolutional and recurrent neural network. *Future Generation Computer Systems* 113, 571–578 (2020)
39. Wang, L., Niu, J., Yu, S.: Sentidiff: combining textual information and sentiment diffusion patterns for twitter sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* 32(10), 2026–2039 (2019)
40. Wang, Q., Xu, H., Wen, Z.e.a.: Image-to-text conversion and aspect-oriented filtration for multimodal aspect-based sentiment analysis. *IEEE Transactions on Affective Computing* 15(3), 1264–1278 (2023)
41. Wang, X., Wang, M., Cui, H.e.a.: Manifold knowledge-guided feature fusion network for multimodal sentiment analysis. *Expert Systems with Applications* 280, 127537 (2025)
42. Wang, Y., Sun, K., Shi, B.e.a.: A guard against ambiguous sentiment for multimodal aspect-level sentiment classification. *Information Processing & Management* 63(2), 104375 (2026), early Access
43. Wang, Y.e.a.: Automatically constructing a fine-grained sentiment lexicon for sentiment analysis. *Cognitive Computation* 15(1), 254–271 (2023)
44. Wei, Y., Yuan, S., Yang, R.e.a.: Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. pp. 5240–5252 (2023)
45. Wu, H., Kong, D., Wang, L.e.a.: Multimodal sentiment analysis method based on image-text quantum transformer. *Neurocomputing* 637, 130107 (2025)
46. Wu, J., Li, S.: Spp-scl: Semi-push-pull supervised contrastive learning for image-text sentiment analysis and beyond. vol. 40, pp. 2173–2181 (2026), proceedings of the AAAI Conference on Artificial Intelligence, forthcoming
47. Wu, X., Li, L., Tao, X.e.a.: Towards the explanation consistency of citizen groups in happiness prediction via factor decorrelation. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2025)
48. Xia, R., Lin, Y., Qin, G., He, L., Li, Y.: Confidence-guided dynamic sequential fusion for multimodal sentiment analysis. *World Wide Web* 29(1) (2026)
49. Xiong, H., Yue, Y., Cai, Y.: Mgsr: Multi-granularity select and reconstruction for multimodal abnormal sentiment detection. *Neurocomputing* p. 133310 (2026), early Access
50. Xu, C., Li, E., Li, H., Liu, X., Chen, W., Guo, J.: Enhancing multimodal sentiment analysis via pairwise emotional correlation distillation and information bottleneck. *Neurocomputing* 675, 132653 (2026)
51. Xu, N.: Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In: *Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics*. pp. 152–154 (2017)
52. Xu, N., Mao, W.: Multisentinet: A deep semantic network for multimodal sentiment analysis. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. pp. 2399–2402 (2017)
53. Xu, N., Mao, W., Chen, G.: A co-memory network for multimodal sentiment analysis. In: *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 929–932 (2018)
54. Xu, N., Mao, W., Chen, G.: Multi-interactive memory network for aspect based multimodal sentiment analysis. vol. 33, pp. 371–378 (2019)
55. Xu, X., Zhu, G., Wu, H., Zhang, S., Li, K.C.: See-3d: Sentiment-driven emotion-cause pair extraction based on 3d-cnn. *Computer Science and Information Systems* 20(1), 77–93 (2023)

56. Yang, X., Feng, S., Zhang, Y., Wang, D.: Multimodal sentiment detection based on multi-channel graph neural networks. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 328–339 (2021)
57. Yang, Z., He, Q., Yu, M.e.a.: Tctr: Text-guided contrastive learning with token-level reconstruction network for missing modalities in multimodal sentiment analysis. *Information Fusion* 126, 103571 (2026), early Access
58. Zhai, G.e.a.: Multi-attention fusion modeling for sentiment analysis of educational big data. *Big Data Mining and Analytics* 3(4), 311–319 (2020)
59. Zhang, B., Yuan, Z., Xu, H.e.a.: Crossmodal translation based meta weight adaption for robust image-text sentiment analysis. *IEEE Transactions on Multimedia* (2024)
60. Zhang, S., Li, J., Li, S., Duan, W., Wei, Z., Li, K.C.: Textual graph representation with syntactic weighting for implicit sentiment analysis. *IEEE Transactions on Emerging Topics in Computational Intelligence* 9(6), 4288–4299 (2025)
61. Zhang, S., Liu, J., Jiao, Y., Zhang, Y., Chen, L., Li, K.C.: A multimodal semantic fusion network with cross-modal alignment for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications* 21(10), 280:1–280:22 (2025)
62. Zhang, S., Ma, Z., Li, H., Liu, Y., Chen, L., Li, K.C.: Gender opposition recognition method fusing emojis and multi-features in chinese speech. *Soft Computing* 29(4), 2379–2390 (2025)
63. Zhang, S., Xu, H., Zhu, G., Chen, X., Li, K.C.: A data processing method based on sequence labeling and syntactic analysis for extracting new sentiment words from product reviews. *Soft Computing* 26(2), 853–866 (2022)
64. Zhang, S., Zhao, T., Wu, H., Zhu, G., Li, K.C.: Ts-gcn: Aspect-level sentiment classification model for consumer reviews. *Computer Science and Information Systems* 20(1), 117–136 (2023)
65. Zhang, S., Zhu, A., Zhu, G., Wei, Z., Li, K.C.: Building fake review detection model based on sentiment intensity and pu learning. *IEEE Transactions on Neural Networks and Learning Systems* 34(10), 6926–6939 (2023)
66. Zhao, T., Meng, L.A., Song, D.: Madsc: Aspect-aware description and calibrated alignment for unified multimodal aspect-based sentiment analysis. *Pattern Recognition* p. 113712 (2026), early Access
67. Zhao, X., Poria, S., Li, X.e.a.: Toward robust multimodal sentiment analysis using multimodal foundational models. *Expert Systems with Applications* 276, 126974 (2025)
68. Zhao, Y., Cambria, E., E, X., Zhu, X.: Tempo: Training-time equilibration of modalities for per-sample optimization in multimodal sentiment. *IEEE Transactions on Affective Computing* pp. 1–17 (2026), early Access
69. Zhao, Z., Zhu, H., Xue, Z.e.a.: An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management* 56(6), 102097 (2019)
70. Zhou, H., Liu, J., Li, X.e.a.: Text-centric sparse interaction fusion network with a modality calibrating module for multimodal sentiment analysis. *IEEE Transactions on Affective Computing* (2026)
71. Zhou, Y., Zhang, S., Wang, C., Wang, Y., Wang, X., Li, K.C.: Topic-oriented sarcasm detection via entity knowledge-based prompt learning. *Computer Science and Information Systems* 22(1), 33–57 (2025)
72. Zhuang, Y., Liu, M., Zhang, Y.e.a.: Tmdc: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. vol. 40, pp. 2281–2289 (2026), proceedings of the AAAI Conference on Artificial Intelligence, forthcoming

Shishu Qi is currently pursuing the M.S. degree with the School of Automotive and Traffic Engineering, Tianjin University of Technology and Education, Tianjin, China. His

research interests include multimodal sentiment analysis, multimodal information fusion, computer vision, image processing, and deep learning.

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

Siyang Zhang is currently pursuing the M.S. degree with the School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, China. His research interests include multimodal sentiment analysis, multimodal learning, deep learning, representation learning, and multimodal information fusion.

Jia Xu received the M.S. degree and is currently with the School of Automotive and Traffic Engineering, Tianjin University of Technology and Education, Tianjin, China. Her research interests include multimodal sentiment analysis, multimodal information fusion, computer vision, image processing, and deep learning.

Kuan-Ching Li received the Ph.D. degree in Electrical Engineering from the University of São Paulo, Brazil. He is currently a Distinguished Professor with the Department of Computer Science and Information Engineering, Providence University, Taiwan, and is also affiliated with the School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, China. His research interests include parallel and distributed computing, GPU computing, cloud computing, big data, and emerging technologies.

Guangli Zhu received the M.S. degree and is currently an Associate Professor and master's supervisor with the School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China. Her research interests include intelligent information processing, text mining, affective computing, multimodal sentiment analysis, and deep learning.

Received: May 26, 2026; Accepted: August 4, 2026.

