Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dominant-Aware Dynamic Routing and Multi-Center Enhancement Network for Multimodal Sentiment Analysis

Lei Ma
Kunming University of Science and Technology
Yajie Guo
1149106839@qq.com
Kunming University of Science and Technology
Sanli Yi
Kunming University of Science and Technology
Dangguo Shao
Kunming University of Science and Technology

Research Article

Keywords:

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11028525/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Dominant-Aware Dynamic Routing and Multi-Center Enhancement Network for Multimodal Sentiment Analysis

Lei Ma[1], Yajie Guo[1], Sanli Yi[1], Dangguo Shao[1*]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, 727 Jingming South Road, Chenggong District, Kunming, 650500, Yunnan, China.

*Corresponding author(s). E-mail(s): shaodg@kust.edu.cn;
Contributing authors: lei_ma@kust.edu.cn; 1149106839@qq.com; 152514845@qq.com;

Abstract

Purpose Multimodal sentiment analysis (MSA) aims to integrate language, vision, and audio for precise emotion understanding. However, modality contribution varies dynamically across samples, and existing methods either predefine a fixed dominant modality or adopt static fusion rules, failing to reconstruct the fusion architecture according to sample-specific characteristics.

Methods We propose the DAMME network (Dominant-Aware Multi-center Modality Enhancement Network), comprising four components: (1) an emotion-ratio-based dominant modality identification mechanism; (2) a feature disentanglement module decomposing multimodal information into shared and specific subspaces; (3) multi-center enhancers TFM, VFM, and AFM for text-, vision-, and audio-dominated scenarios; (4) a dynamic routing module that activates the corresponding enhancer according to the identified dominant modality during training.

Results Experiments on CMU-MOSI, CMU-MOSEI, CH-SIMS, and CH-SIMSv2 demonstrate that DAMME achieves state-of-the-art performance, outperforming GRCF on 7-class classification (50.05% vs. 49.3% on MOSI) and surpassing MSLMamba across all metrics, while maintaining a lightweight architecture without external LLM supervision.

Conclusion The proposed dynamic architecture reconstruction paradigm offers a new research direction for multimodal sentiment analysis.

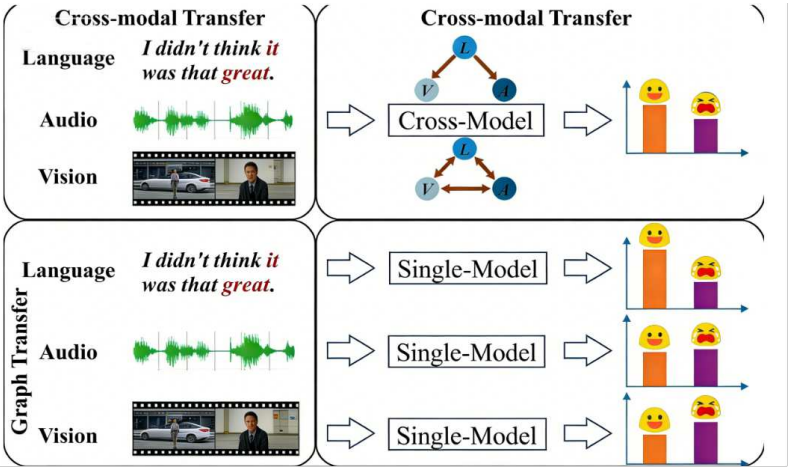


Fig. 1 Schematic illustration of the importance of primary modality identification.

Keywords: Multimodal sentiment analysis, dynamic dominant modality identification, multi-center enhancement, dynamic routing, feature disentanglement

1 Introduction

With the rapid development of social media and multimedia platforms, human emotional expression has become increasingly diverse. Relying solely on a single modality, such as text or speech alone, is no longer sufficient to capture the subtle variations of complex emotions [1, 2]. Multimodal sentiment analysis (MSA) has emerged as a solution that integrates heterogeneous modalities—language, vision, and audio—to achieve a more accurate understanding of user emotions [3, 4]. Anwar et al. [5] systematically reviewed the progress of MSA, covering data modalities, fusion methods, technical challenges, and future prospects, and also provided a systematic overview of statistical analyses of datasets and evaluation metrics. This technology holds broad application prospects in areas such as mental health monitoring, human-computer interaction, and public opinion analysis [6]. However, as illustrated in Figure 1, the contribution of each modality to sentiment prediction is not constant: a large body of research indicates that language typically carries the densest semantic information and plays a dominant role in multimodal fusion [7, 8]; yet in silent videos or non-verbal communication scenarios, visual expressions or vocal prosody may become the primary carriers of sentiment. This dynamic variation in modality dominance poses a fundamental challenge to the design of multimodal fusion strategies.

Existing methods can be grouped into three categories based on how they handle dominant modality:

Text-dominant methods:

These assume that language is central to sentiment expression, typically employing pre-trained language models such as BERT [9] to extract deep semantic features, and then incorporating visual and audio features as auxiliary information through attention mechanisms or feature concatenation [10, 11]. Representative works include MAG-BERT [12] and AOBERT [13], which perform well in language-rich scenarios but tend to lose crucial emotional cues when vision or audio dominates.

Non-text-dominant methods:

Designed for cases where text is missing or unreliable, these approaches focus on mining sentiment information from visual and audio modalities, e.g., facial action unit (AU) based visual sentiment analysis [14] and acoustic feature based speech emotion recognition [15]. While they can function in specific contexts, they underutilize linguistic information and struggle with samples rich in textual content [16].

Fixed fusion rule methods:

This is the current mainstream approach, which aims to learn an optimal fusion strategy without presupposing a dominant modality. Techniques include feature-level concatenation [17], decision-level voting [18], and more complex tensor fusion (TFN) [19], low-rank fusion (LMF) [20], and multi-head attention (MulT) [21]. These methods learn modality weights in a global statistical sense, but once trained, they process all test samples through a fixed forward path, failing to adjust modality contributions dynamically on a per-sample basis.

The common limitation of the above approaches is that they handle modality dominance statically. Whether by defaulting to text dominance or by learning global fusion weights, the model employs identical computational logic during inference, which cannot accommodate the dynamic shifts in modality dominance across samples [22]. When non-dominant modalities contain noise or express sentiment that conflicts with the dominant modality, forced fusion introduces redundancy or even contradictory information, degrading overall performance [23].

To alleviate this issue, several recent methods have introduced dynamic modality selection mechanisms:

GRCF [24]:

Utilizes a large language model fine-tuned with low-rank adapters and Group Relative Policy Optimization (GRPO) to generate chain-of-thought reasoning that explicitly decides which modalities to use for the current sample.

MSLMamba [25]:

Based on the state-space model (Mamba), it builds cross-modal interactions with linear complexity, dynamically modeling long-range dependencies between modalities.

Unified fusion framework [26]:

Integrates cross-modal attention interactions with a neural variational mixture of experts (NV-MoE) to adaptively combine expert networks according to the input context.

MODS [27]:

Proposes a sample-adaptive dominant modality selector (MSelector) that explicitly determines the dominant modality for each sample, and designs a dominant-centered cross-attention (PCCA) module.

KuDA [28]:

Through emotional knowledge injection and dynamic attention fusion, it dynamically selects the dominant modality based on sentiment ratios and adjusts modality contributions, achieving state-of-the-art performance on several datasets.

AGFN [29]:

Dynamically adjusts fusion weights of each modality based on information entropy estimation, effectively suppressing interference from noisy modalities.

Despite their progress in dynamic weighting and modality selection, these methods still suffer from a fundamental limitation: the dynamic selection mechanism is deeply decoupled from the subsequent fusion architecture. Specifically, the existing approaches either use the selection result merely to discard low-quality modalities (e.g., GRPO) or to adjust weights within a fixed fusion network (e.g., MSLMamba, MODS). However, the fusion architecture itself does not restructure when the dominant modality switches. When vision is identified as dominant, the ideal strategy should treat vision as the query center and align text and audio features toward it, rather than preserving a language-centric interaction pattern. This "dynamic decision + static fusion" paradigm restricts the potential of dynamic selection.

On the other hand, dominant-modality-centric feature enhancement has proven to be an effective means of improving multimodal representations. The DLF framework [30] proposes a language-focused attractor that enhances multimodal representations in language-dominant scenarios through feature disentanglement and language-as-query cross-attention. However, DLF presumes that language is the sole dominant modality and cannot handle vision- or audio-dominated samples, and its enhancement paradigm has not been extended to other modalities.

To address these issues, we propose the DAMME network (Dominant-Aware Multicenter Modality Enhancement Network), which deeply integrates dynamic dominant modality identification with dominant-modality-centric feature enhancement, constructing a unified model that can dynamically select the dominant modality and perform dedicated enhancement for each dominant modality (text/vision/audio). The framework consists of four core components:

1. **Emotion-Ratio-Based Dominant Modality Identification**: Leverages emotional knowledge injection and unimodal sentiment predictions to compute a

sentiment ratio for each modality, which explicitly identifies the dominant modality during training and guides the dynamic fusion.
2. **Feature Disentanglement Module**: Decomposes each modality's features into modality-shared and modality-specific subspaces to reduce redundancy.
3. **Multi-Center Enhancer Set**: Constructs a text-focused fusion module (TFM), a vision-focused fusion module (VFM), and an audio-focused fusion module (AFM) for text-, vision-, and audio-dominated scenarios, respectively, performing cross-attention enhancement centered on that modality.
4. **Dynamic Routing Module**: Based on the sentiment ratios, dynamically activates the corresponding center enhancer and reconstructs the fusion path—activating TFM when text dominates, VFM when vision dominates, and AFM when audio dominates—achieving sample-level adaptation of the fusion architecture during training.

This framework fundamentally addresses two core limitations of existing methods: **(1) The disconnect between dynamic selection and fusion**—the selection result directly drives the architectural reconfiguration rather than merely adjusting weights.

(2) The single-enhancement paradigm—dedicated enhancers for each dominant modality ensure that feature enhancement matches modality characteristics.

Experiments on four benchmark datasets—CMU-MOSI [31], CMU-MOSEI [32], CH-SIMS [33], and CH-SIMSv2 [34]—demonstrate that DAMME achieves competitive performance on datasets with relatively balanced modality importance, particularly exhibiting superior adaptability on samples where vision or audio is dominant compared to existing methods. Ablation studies validate the effectiveness of each component, and visualizations showcase the model's dynamic adaptability to dominant modalities.

The main contributions of this paper are summarized as follows:

- We propose the DAMME network, which combines dynamic dominant modality identification with modality-centric dedicated enhancement, enabling architecture-level dynamic reconstruction for multimodal sentiment analysis.
- We design multi-center enhancers (TFM/VFM/AFM), extending the text-centric enhancement paradigm to vision and audio modalities, and verify their effectiveness.
- We introduce a dynamic routing mechanism that reconstructs the fusion architecture based on the sentiment ratios, resolving the disconnect between dynamic selection and fusion.
- We achieve competitive performance on four datasets, with superior adaptability on vision- or audio-dominated samples, demonstrating the model's capability across diverse dominant-modality scenarios.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details the DAMME network; Section 4 reports experimental setup and results; Section 5 concludes and discusses future directions.

2 Related Work

Multimodal sentiment analysis aims to integrate heterogeneous information from language, vision, and audio to improve the accuracy of sentiment prediction, and has thus attracted widespread attention in recent years [2, 3]. Early studies primarily explored symmetric fusion architectures, which treat all modalities equally and model bidirectional interactions between every pair of modalities. The Tensor Fusion Network (TFN) [19] captures high-order interactions among modalities via a three-dimensional outer product, albeit at a high computational cost; the Low-rank Multimodal Fusion (LMF) [20] introduces tensor decomposition to reduce this computational overhead. With the growing popularity of attention mechanisms, MulT [21] employs cross-modal attention to interact with unaligned multimodal sequences, serving as the foundation for many subsequent works. On this basis, graph neural networks [16] and contrastive learning [35] have been introduced to enhance cross-modal associations. Although symmetric approaches can capture rich cross-modal information, they neglect the varying importance of different modalities in emotional expression. When certain modalities contain noise or ambiguous sentimental cues, forced fusion tends to introduce redundancy or conflicting information [22].

In contrast to symmetric approaches, another mainstream line of work recognizes that language typically plays a dominant role in sentiment analysis, and therefore adopts a text-centric fusion strategy. MAG-BERT [12] incorporates visual and audio features into a pre-trained language model through adapters to enhance textual representations; TETFN [11] designs a text-enhanced Transformer fusion network that uses text as the query for cross-modal attention. AOBERT [13] compresses visual and audio information into special tokens embedded within BERT, achieving an end-to-end text-centric fusion. These methods perform well in text-dominated scenarios, but their fusion architecture remains fixed as text-centric. When vision or audio becomes dominant, the model's attention becomes dispersed and fails to effectively utilize the emotional cues from non-linguistic modalities [28].

Dynamic modality selection and feature disentanglement.

Recent studies have explored dynamic modality selection to address the varying importance of modalities. Gated multimodal units [36] adjust modality weights based on input features. DAMFT [37], DDMM [38], DAF [39], and AMTE [40] attempt to select the primary modality or adapt fusion weights dynamically. However, these methods mainly adjust weights or attention directions within a fixed fusion architecture, lacking architecture-level dynamic adaptation. Feature disentanglement methods such as DMD [41] separate shared and specific information to reduce redundancy, but still rely on fixed fusion schemes or unidirectional enhancement. This motivates DAMME, which integrates dynamic dominant modality identification with multi-center enhancement to reconstruct the fusion architecture at the sample level.

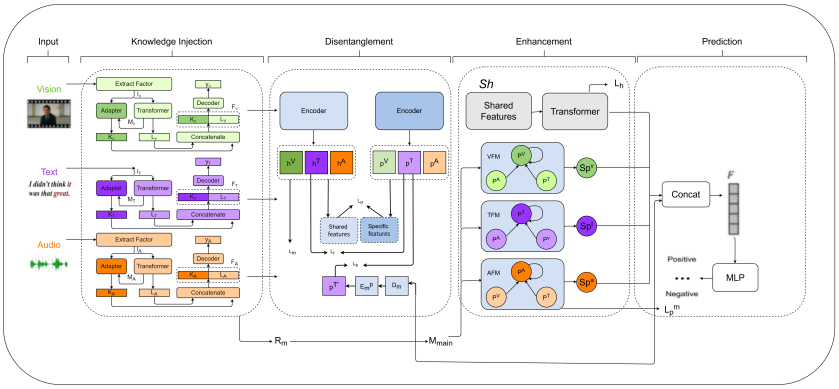


Fig. 2 Architecture of the DAMME network.

3 DAMME Network

3.1 Problem Definition and Notation

In the multimodal sentiment analysis task, the input data consists of three modalities: text (t), vision (v), and audio (a). The raw input for each modality can be represented in sequence form:

$$I_t \in \mathbb{R}^{T_t \times d_t}, \quad I_v \in \mathbb{R}^{T_v \times d_v}, \quad I_a \in \mathbb{R}^{T_a \times d_a} \tag{1}$$

where T_m denotes the sequence length, d_m represents the feature dimension, and $m \in \{t, v, a\}$. The objective of the model is to predict a sentiment score y, typically a continuous value (e.g., $[-3, 3]$ or $[-1, 1]$), with positive values indicating positive sentiment and negative values indicating negative sentiment.

3.2 Overall Architecture

The proposed DAMME network is illustrated in Figure 2. Its core philosophy is to deeply integrate the capability of dynamic dominant modality identification with a modality-centric feature enhancement paradigm. The framework consists of five main stages: knowledge-injected encoding, feature disentanglement, dominant modality identification, multi-center enhancement, and fusion with hierarchical prediction. First, pre-trained encoders and adapters extract global semantic representations and knowledge-aware emotional representations from each modality, and unimodal sentiment scores are predicted to compute sentiment ratios. Second, the representations of each modality are fed into a disentanglement module and decomposed into modality-shared and modality-specific features. Third, during training, the dominant modality for the current sample is identified based on sentiment ratios, and the model learns to route each sample to the corresponding enhancement path. Fourth, at inference, since the sentiment ratio is fixed, the learned attention parameters determine the contribution of each enhancement path. Finally, the enhanced features are fused and passed through a hierarchical prediction module to output the final sentiment score.

3.3 Knowledge-Injected Encoding

We first encode each modality and inject emotional knowledge. For the text modality, we employ BERT [9] to extract a global semantic representation:

$$L_t, M_t = \text{BERT}(I_t) \tag{2}$$

where $L_t \in \mathbb{R}^{T_t \times d_t}$ denotes the last-layer hidden states, and M_t denotes the intermediate-layer outputs, with I_t being the raw textual input. For the vision and audio modalities, we adopt stacked Transformer encoder layers [42] to extract their representations:

$$L_m, M_m = \text{Encoder}_m(I_m), \quad m \in \{v, a\} \tag{3}$$

To inject modality-specific emotional knowledge, we attach an adapter after each encoder layer. The adapter takes the raw input I_m and the intermediate output M_m as inputs, and outputs a knowledge-aware emotional representation $K_m \in \mathbb{R}^{T_m \times d_m}$:

$$K_m = \text{Adapter}_m(I_m, M_m), \quad m \in \{t, v, a\} \tag{4}$$

Subsequently, we concatenate the global semantic representation with the knowledge-aware emotional representation to obtain the emotion-fused modality representation F_m:

$$F_m = [K_m; L_m] \in \mathbb{R}^{T_m \times 2d_m}, \quad m \in \{t, v, a\} \tag{5}$$

where $[\cdot;\cdot]$ denotes the concatenation operation. To obtain unimodal sentiment cues, we use an MLP-based decoder to predict from F_m:

$$\widehat{y}_m = \text{Decoder}_m(F_m), \quad m \in \{t, v, a\} \tag{6}$$

where $\widehat{y}_m$ is the predicted sentiment score for the m-th modality. These unimodal predictions will be used in the subsequent computation of sentiment ratios.

3.4 Feature Disentanglement

To reduce redundancy and conflicting information across modalities, we employ a shared encoder E_m^h and a specific encoder E_m^p to decompose each modality's representation into a modality-shared feature h^m and a modality-specific feature p^m:

$$h^m = E_m^h(F_m), \quad p^m = E_m^p(F_m), \quad m \in \{t, v, a\} \tag{7}$$

where both E_m^h and E_m^p are implemented as multi-layer Transformers. To ensure the effectiveness of disentanglement, we introduce four geometric regularization losses:

Reconstruction loss L_r***:***

Reconstruct F_m from $[h^m \oplus p^m]$ via a decoder D_m, minimizing the reconstruction error. This loss ensures complete information preservation during the encoding-decoding process, serving as the foundation of disentanglement.

The generation process of modality features can be formulated as:

$$F'_m = D_m([h^m \oplus p^m]) \tag{8}$$

where D_m is a 1D convolutional decoder, and $\oplus$ denotes concatenation.

The discrepancy between the low-level feature F_m and the generated feature F'_m, termed the reconstruction loss, serves as a regularization term for the feature disentanglement module:

$$L_r = \|F_m - F'_m\|^2 \tag{9}$$

Specific loss L_s:

Extract specific features again from the reconstructed features and compute the error against the original p^m. This loss ensures that the private feature encoder can independently capture and reconstruct modality-exclusive information.

The reconstruction process for modality-specific features can be expressed as:

$$p^{m'} = E^p_m(F'_m) \tag{10}$$

where $p^{m'}$ is the estimated modality-specific feature derived from the modality-specific reconstruction process. The difference between the original specific feature p^m and the estimated one $p^{m'}$ can be regarded as the specific loss L_s:

$$L_s = \|p^m - p^{m'}\|^2 \tag{11}$$

Shared loss L_m:

In the shared space, make cross-modal samples with the same sentiment close to each other, while pushing those with different sentiments apart. This loss aims to regularize the shared space and enhance the representational consistency of shared features across modalities for the same sentiment.

$$L_m = \frac{1}{|TP|} \max\left(0,\ d(S, PS) - d(S, NS) + \mu\right) \tag{12}$$

where S denotes a sample drawn from the modality-shared space, PS denotes a positive sample from a different modality with the same sentiment label, NS denotes a negative sample from the same modality with a different sentiment label, TP is the total number of positive and negative sample pairs, $d(\cdot, \cdot)$ computes the cosine similarity between two feature vectors, and μ denotes the margin.

Soft orthogonal loss L_o:

This loss enforces orthogonality between the shared and specific features to achieve physical separation of information, thereby minimizing redundancy and conflict between the shared and specific multimodal features.

$$L_o = O(h^m, p^m) \tag{13}$$

where $O(\cdot, \cdot)$ denotes the non-negative version of cosine similarity to promote orthogonality between the two feature spaces.

These losses are combined into the disentanglement loss:

$$L_d = \sum_{k \in \{r,s,m,o\}} \lambda_k L_k \tag{14}$$

3.5 Dominant Modality Identification

We exploit the discrepancy between the unimodal predictions $\widehat{y}_m$ and the ground-truth label y to assess each modality's contribution to the final sentiment. During training, we compute the sentiment ratio R_m for each modality:

$$D_m = \exp\left(-k|\widehat{y}_m - y|^2\right), \quad R_m = \frac{D_m}{D_t + D_v + D_a} \tag{15}$$

where k is a scaling coefficient. R_m reflects how close the m-th modality's sentiment prediction is to the ground truth; a larger value indicates higher reliability of that modality.

We designate the modality with the largest sentiment ratio as the dominant modality M_{main} for the current sample:

$$M_{main} = \arg \max_{m \in \{t,v,a\}} R_m \tag{16}$$

During training, the sentiment ratio R_m is computed using the ground-truth label y as in Eq. 15, and is used to guide the dynamic routing and fusion. Since the dominant modality M_{main} is defined through R_m, it is also only meaningful during training. At the test stage, because y is unavailable, we follow the practice of KuDA [28] and fix the sentiment ratio to $R_m = 1$ for all modalities. In this way, the model relies on the attention parameters learned during training to differentiate modality contributions, rather than on the test-time sentiment ratio.

3.6 Multi-Center Enhancement

To fully leverage the emotional cues from the dominant modality, we extend the text-centric fusion module (TFM) of DLF [30] and construct three enhancers centered on different modalities: the Text-centric Fusion Module (TFM), the Vision-centric Fusion Module (VFM), and the Audio-centric Fusion Module (AFM). These enhancers all operate on the disentangled modality-specific spaces p^m, with multi-head cross-attention as their core mechanism.

First, projection layers map the specific features to a unified dimension d_f:

$$\overline{p^m} = \text{Projector}_m(p^m) \in \mathbb{R}^{T_f \times d_f}, \quad m \in \{t, v, a\} \tag{17}$$

where T_f is the unified sequence length for the fusion stage. The shared features h^m are similarly projected to $\overline{h^m}$ and subsequently aggregated through a shared Transformer layer to produce a global shared representation $Sh \in \mathbb{R}^{T_f \times d_f}$.

During training, the dominant modality M_{main} determines which enhancer is emphasized, while at the test stage all three enhancers participate through soft routing, as detailed later.

Text-dominant scenario ***($M_{main} = t$):***

TFM takes p^t as the query Q_t, and p^v, p^a as key-value pairs (K_m, V_m), performing two streams of cross-attention: $V \to T$ and $A \to T$ for targeted enhancement. A self-attention path $T \to T$ is also retained. The enhanced text-specific feature is denoted as Sp^t. This process can be formalized as:

$$K_t = V_t = p^t, \quad K_a = V_a = p^a, \quad K_v = V_v = p^v \tag{18}$$

$$Sp^t = \text{Fusion}\left(\text{MultiHead}(Q_t, K_m, V_m)\right), \quad m \in \{t, v, a\} \tag{19}$$

Vision-dominant scenario ***($M_{main} = v$):***

VFM is designed to address the critical role of the visual modality in emotional expression, along with its issues of sequence redundancy and noise. VFM introduces a Graph-based Dynamic Compressor (GDC), which employs capsule networks and graph convolution operations to compress and filter redundant frames and noisy units in the visual features p^v along both the temporal and channel axes, producing a visual query representation with higher information density. Subsequently, VFM uses Q_v as the query and p^t, p^a as key-value pairs, performing two cross-attention streams to complete the targeted interactive enhancement from text and audio toward the visual features. Additionally, VFM retains a self-attention path $V \to V$ to preserve the original dynamic emotional characteristics of the visual modality. This process is formalized as:

$$Q_v = \text{GDC}(p^v) \tag{20}$$

$$Sp^v = \text{Fusion}\left(\text{MultiHead}(Q_v, K_m, V_m)\right), \quad m \in \{t, v, a\} \tag{21}$$

where MultiHead$(\cdot)$ denotes the multi-head cross-attention operation. This design effectively suppresses sequential redundancy and noise interference in the visual modality via GDC, while ensuring that cross-modal interactions are oriented toward the visual modality through the PCCA core, thereby enhancing the contribution efficiency of visual features in sentiment analysis.

Audio-dominant scenario ***($M_{main} = a$):***

AFM is built upon a dual-audio parallel channel architecture, further incorporating an audio-visual polarity vector auxiliary task. Specifically, AFM adopts the core philosophy of Hierarchical Cross-Modal Attention and Dual Audio Pathways [43]. It first performs hierarchical enhancement on the audio features: the primary channel employs a pre-trained Wav2Vec 2.0 model to extract deep acoustic embeddings, capturing high-order emotional information such as intonation and fundamental frequency; the auxiliary channel adopts the COVAREP framework to capture shallow waveforms and foundational acoustic features. The outputs of the two channels are adaptively weighted and fused through a dynamic gating module:

$$F_a = \text{Gate}(F_{deep}) \odot F_{deep} + \text{Gate}(F_{shallow}) \odot F_{shallow} \tag{22}$$

where $\odot$ denotes element-wise multiplication. On this basis, AFM introduces an Audio-Visual Polarity Vector Auxiliary task [44], which explicitly reinforces the dimensions strongly correlated with emotional polarity discrimination in the audio features and suppresses ambiguous noise components through a contrastive learning strategy, thereby enhancing the discriminative capability of audio sentiment polarity. Subsequently, AFM uses the enhanced audio feature as the query Q_a, and takes the text-specific feature p^t and vision-specific feature p^v as keys and values, respectively, performing two streams of hierarchical cross-attention ($T \rightarrow A$ and $V \rightarrow A$) for targeted cross-modal interaction. This process is formalized as:

$$Q_a = \text{PolarityEnhance}(F_a) \tag{23}$$

$$Sp^a = \text{Fusion}\left(\text{MultiHead}(Q_a, K_m, V_m)\right), \quad m \in \{t, v, a\} \tag{24}$$

Taking TFM as an example, its cross-attention operation can be formalized as:

$$\text{MultiHead}(Q_t, K_m, V_m) = \text{softmax}\left(\frac{Q_t K_m^T}{\sqrt{d_f}}\right) V_m \tag{25}$$

After passing through multiple Transformer blocks, we obtain the enhanced specific features $Sp^m \in \mathbb{R}^{T_f \times d_f}$ for each modality. This design enables the model to dynamically reconstruct the interaction direction based on the dominant modality, rather than merely adjusting weights within a fixed structure.

To avoid the information loss caused by hard switching and preserve the contribution of auxiliary modalities, we further introduce a soft routing mechanism. Instead of activating only the enhancer corresponding to the dominant modality, we perform a weighted fusion of all three enhancer outputs, where the weights are derived from the normalized sentiment ratios R_m defined in Eq. 15. This soft routing strategy ensures that non-dominant modalities still contribute to the final representation when their sentiment predictions are relatively reliable. At the test stage, since R_m is fixed to 1, the soft routing weights become uniform, and the model relies on the learned cross-attention parameters to differentiate modality contributions.

3.7 Fusion and Hierarchical Prediction

After obtaining the enhanced specific features Sp^m ($m \in \{t, v, a\}$) and the shared feature Sh, we first perform a weighted fusion of the three enhanced features using the sentiment ratios R_m as weights, then concatenate the fused feature with the shared feature to form the final multimodal representation:

$$F = \text{Concat}(R_t \cdot Sp^t + R_v \cdot Sp^v + R_a \cdot Sp^a, Sh) \in \mathbb{R}^{T_f \times 2d_f} \tag{26}$$

Subsequently, F is fed into an MLP to produce the final sentiment prediction:

$$\widehat{y} = \text{MLP}(\text{Mean}(F)) \tag{27}$$

We also adopt a hierarchical prediction strategy, leveraging the pre-fusion features for auxiliary predictions to further boost model performance. Specifically, we pass the

shared feature Sh and each modality's enhanced specific feature Sp^m through individual classifiers to obtain predictions y_{Sh} and y_{Sp^m}, and compute their corresponding losses L_h and L_{p^m}. The final multimodal sentiment analysis loss is formulated as:

$$L_{MSA} = \beta_d L_d + \beta_h L_h + \sum_{m \in \{t,v,a\}} \beta_{p^m} L_{p^m} \tag{28}$$

where L_d denotes the disentanglement loss, L_h and L_{p^m} represent the shared and specific prediction losses, respectively, and the β terms are balancing coefficients.

Additionally, to further reinforce the contribution of the dominant modality, we introduce a contrastive learning loss. We compute a noise-contrastive estimation (NCE) loss between the final multimodal representation F and the enhanced specific feature $Sp^{M_{main}}$ of the dominant modality:

$$L_{cor} = \mathcal{L}_{\text{NCE}}(F, Sp^{M_{main}}) \tag{29}$$

This loss encourages the multimodal representation to be closer to the features of the dominant modality, thereby improving the utilization efficiency of dominant cues.

3.8 Overall Training Objective

The total loss function of the integrated framework combines the disentanglement loss, the multimodal sentiment analysis loss, and the contrastive loss:

$$L_{total} = L_{MSA} + \alpha L_{cor} \tag{30}$$

where α is a balancing parameter. The entire model is trained in an end-to-end manner. However, the adapters and decoders in the knowledge-injection stage can be pre-trained separately using unimodal sentiment prediction tasks, and then fine-tuned jointly on the downstream task. In this way, DAMME learns to identify the dominant modality during training and reconstruct the fusion architecture accordingly, achieving self-adaptation to scenarios with different dominant modalities.

4 Experiments and Analysis

4.1 Experimental Setup

4.1.1 Datasets

To comprehensively evaluate the effectiveness of the DAMME network, we conduct experiments on four publicly available multimodal sentiment analysis benchmark datasets, including two English datasets (CMU-MOSI, CMU-MOSEI) and two Chinese datasets (CH-SIMS, CH-SIMSv2).

CMU-MOSI [31] is the first large-scale multimodal sentiment analysis dataset, comprising 2,199 monologue video clips sourced from YouTube. Each clip is annotated with a sentiment score ranging from -3 (strongly negative) to $+3$ (strongly positive). The dataset is partitioned into 1,284 training samples, 229 validation samples, and 686 test samples.

CMU-MOSEI [32] is an extended version of MOSI, containing 22,856 movie-review video clips, also annotated on a -3 to $+3$ scale. The partition includes 16,326 training samples, 1,871 validation samples, and 4,659 test samples.

CH-SIMS [33] is a Chinese multimodal sentiment analysis dataset that includes 2,281 video clips. Each clip provides not only a holistic multimodal sentiment annotation (from -1 to 1), but also independent unimodal annotations for vision, text, and audio. This characteristic makes it an ideal benchmark for validating dominant modality selection. The dataset is divided into 1,368 training samples, 456 validation samples, and 457 test samples.

CH-SIMSv2 [34] is an extension of CH-SIMS, comprising 4,403 video clips with the same fine-grained unimodal and multimodal annotations. The split consists of 2,722 training samples, 647 validation samples, and 1,034 test samples.

4.1.2 Evaluation Metrics

We adopt multiple evaluation metrics to comprehensively assess model performance:

- **Mean Absolute Error (MAE)**: The average absolute error between predictions and ground truths; lower values indicate better performance.
- **Pearson Correlation Coefficient (Corr)**: The correlation between predictions and ground truths; higher values indicate better performance.
- **Classification Accuracy**: We use different granularities for different datasets. For CH-SIMS and CH-SIMSv2, we report five-class accuracy (Acc-5) and three-class accuracy (Acc-3); for MOSI and MOSEI, we report seven-class accuracy (Acc-7) and binary accuracy (Acc-2).
- **F1 Score**: A comprehensive evaluation metric that balances precision and recall.

4.1.3 Baseline Models

To verify the superiority of DAMME, we compare it against multiple mainstream approaches:

Symmetric fusion methods:

- **TFN** [19]: Captures high-order interactions across modalities via a three-dimensional outer product.
- **LMF** [20]: A low-rank multimodal fusion approach based on tensor decomposition.
- **MulT** [21]: A Transformer-based fusion model leveraging cross-modal attention.
- **MISA** [22]: Projects each modality into modality-invariant and modality-specific subspaces.
- **Self-MM** [45]: A self-supervised multi-task learning framework.
- **CubeMLP** [46]: An MLP-based multimodal fusion model.
- **TMBL** [47]: A Transformer-based multimodal binding learning model.

Text-centric methods:

- **MAG-BERT** [12]: Integrates multimodal information into a pre-trained BERT model.
- **BBFN** [48]: Controls fusion via bimodal correlation.

- **MMIM** [49]: Maximizes hierarchical mutual information.
- **CENet** [50]: A cross-modal enhancement network.
- **TETFN** [11]: A text-enhanced Transformer fusion network.
- **ALMT** [51]: A language-guided adaptive hyper-modality learning method.

Dynamic selection methods:

- **GRCF** [24]: A large-language-model-based modality selection approach using GRPO.
- **MSLMamba** [25]: A state-space-model-based dynamic fusion framework.
- **MODS** [27]: Dynamic dominant modality selection and optimization.
- **KuDA** [28]: A knowledge-guided dynamic attention fusion method.

4.1.4 Implementation Details

Feature extraction: For the text modality, we employ BERT [9] to extract features. Specifically, we use `bert-base-uncased` for English datasets and `bert-base-chinese` for Chinese datasets. For the visual modality, we use OpenFace [14] to extract facial action unit features with a dimensionality of 35. For the audio modality, following the core idea of hierarchical cross-modal attention with dual audio pathways, the primary channel employs Wav2Vec 2.0 to extract deep acoustic embeddings, while the auxiliary channel uses the traditional acoustic feature set COVAREP [15] to extract features with a dimensionality of 74. For the CH-SIMS and CH-SIMSv2 datasets, we directly use their provided pre-processed features.

Model parameters: The number of Transformer layers in each modality encoder is set to 3, with a hidden dimension of 128. The adapter dimension is set to 64. In the feature disentanglement module, the number of layers is 4 for the shared encoder and 2 for the specific encoder. The number of Transformer blocks in the center fusion modules is set to 3.

Training settings: We adopt a two-stage training strategy.

In the first stage, we pre-train the knowledge-injection modules (adapters and decoders) on external data, with a learning rate of 2×10^{-5}, a batch size of 16, and 30 training epochs.

In the second stage, we load the pre-trained modules, freeze their parameters, and jointly train the remaining parts of the model with an initial learning rate of 5×10^{-5} using a cosine annealing scheduler, a batch size of 32, using the AdamW optimizer and a weight decay of 1×10^{-4}. The loss balancing coefficients are set as $\lambda_r = 0.1$, $\lambda_s = 0.1$, $\lambda_m = 0.5$, $\lambda_o = 0.3$, $\beta_d = 0.2$, $\beta_h = 0.3$, $\beta_{p^t} = 0.3$, $\beta_{p^a} = 0.15$, $\beta_{p^v} = 0.15$, $\alpha = 0.3$. The sentiment ratio scaling coefficient k is set to 0.5, and the number of training epochs is 80. We adopt an early stopping mechanism with a patience of 10 epochs on the validation loss to prevent overfitting and save the best-performing model for testing.

To evaluate the stability and statistical significance of the model, all DAMME experiments are run with 5 different random seeds independently. The final reported results are the mean and standard deviation over these 5 independent runs.

4.2 Overall Performance Comparison

4.2.1 Results on CH-SIMS and CH-SIMSv2

Tables 1 and 2 present the performance comparison of various methods on the CH-SIMS and CH-SIMSv2 datasets, respectively. A notable characteristic of these two datasets is that the modality importance distribution is relatively balanced, with vision, text, and audio all potentially serving as the dominant modality [28], making them particularly suitable for evaluating a model's dynamic adaptation capability.

On CH-SIMSv2 (Table 2), DAMME achieves the best performance among all compared methods. Specifically, the average MAE of DAMME is 0.263 ± 0.005, lower than KuDA's 0.271; the average Acc-5 reaches $64.18 \pm 0.38\%$, surpassing KuDA's 61.22% and the text-centric method ALMT's 52.90%.

On CH-SIMS (Table 1), DAMME achieves a MAE of 0.392 ± 0.007, outperforming KuDA (0.408), MODS (0.407), and MSLMamba (0.417). More importantly, its Acc-5 reaches $50.35 \pm 0.44\%$, surpassing MSLMamba's 49.02% and all other compared methods. These results demonstrate that DAMME's dynamic architecture reconstruction effectively handles datasets with balanced modality importance distributions, while exhibiting superior adaptability on samples where vision or audio is dominant.

Table 1 Performance comparison on CH-SIMS.

Method	**Type**	MAE↓	Corr↑	Acc-5↑	Acc-3↑	Acc-2↑	F1↑
TFN	Symmetric	0.432	0.591	39.30	65.12	78.38	78.62
LMF	Symmetric	0.441	0.576	40.53	64.68	77.77	77.88
MulT	Symmetric	0.453	0.564	37.94	64.77	78.56	79.66
Self-MM	Symmetric	0.425	0.595	41.53	65.47	80.04	80.44
CubeMLP	Symmetric	0.419	0.593	41.79	65.86	77.68	77.59
TMBL	Symmetric	0.429	0.592	41.58	65.43	79.12	78.75
BBFN	Text-centric	0.430	0.564	40.92	61.05	78.12	77.88
CENet	Text-centric	0.471	0.534	33.92	62.58	77.90	77.53
TETFN	Text-centric	0.420	0.577	41.79	63.24	81.18	80.24
ALMT	Text-centric	0.408	0.594	43.11	65.86	78.77	78.71
MSLMamba	Dynamic sel.	0.417	0.642	49.02	71.90	82.96	82.38
MODS	Dynamic sel.	0.407	0.605	45.51	66.74	80.96	80.94
KuDA	Dynamic sel.	0.408	0.613	43.54	66.52	80.74	80.71
DAMME	**Integrated**	**0.392 ± 0.007**	**0.635 ± 0.009**	**50.35 ± 0.44**	**68.55 ± 0.38**	**83.05 ± 0.35**	**82.93 ± 0.38**

Table 2 Performance comparison on CH-SIMSv2.

Method	**Type**	MAE↓	Corr↑	Acc-5↑	Acc-3↑	Acc-2↑	F1↑
TFN	Symmetric	0.303	0.707	52.55	72.21	80.14	80.14
LMF	Symmetric	0.367	0.557	47.79	64.90	74.18	73.88
MulT	Symmetric	0.291	0.738	54.81	73.19	80.68	80.73
Self-MM	Symmetric	0.311	0.695	52.77	72.61	79.69	79.76
CubeMLP	Symmetric	0.334	0.648	52.90	71.95	78.53	78.53
TMBL	Symmetric	0.313	0.706	52.03	73.02	80.46	80.36
BBFN	Text-centric	0.300	0.708	53.29	71.47	78.53	78.41
CENet	Text-centric	0.310	0.699	53.04	73.10	79.56	79.63
TETFN	Text-centric	0.310	0.695	54.47	73.65	79.73	79.81
ALMT	Text-centric	0.308	0.700	52.90	71.86	79.59	79.51
GRCF	Dynamic sel.	0.305	0.691	54.4	74.3	81.7	81.6
MODS	Dynamic sel.	0.297	0.712	55.51	73.69	79.59	79.53
KuDA	Dynamic sel.	0.271	0.759	61.22	76.21	82.11	82.04
DAMME	**Integrated**	**0.263 ± 0.005**	**0.768 ± 0.007**	**64.18 ± 0.38**	**78.65 ± 0.32**	**84.25 ± 0.29**	**84.05 ± 0.32**

4.2.2 Results on MOSI and MOSEI

Tables 3 and 4 present the performance comparison of various methods on the MOSI and MOSEI datasets, respectively. These two datasets are predominantly text-dominant [7], yet there still exist some samples where vision or audio plays a dominant role [28].

On the text-dominant MOSI and MOSEI, DAMME achieves competitive performance with GRCF while surpassing it on the more challenging fine-grained classification task. Specifically, on MOSI (Table 3), DAMME attains an MAE of 0.597 ± 0.007 (compared to GRCF's 0.581) and an Acc-7 of $50.05 \pm 0.40\%$ (surpassing GRCF's 49.3% and MSLMamba's 49.52%). On MOSEI (Table 4), DAMME achieves an MAE of 0.498 ± 0.007 and an Acc-7 of $56.13 \pm 0.38\%$, significantly outperforming the recent Neurocomputing publication MSLMamba (55.16%). It is worth noting that GRCF relies on a large language model and reinforcement learning with substantial computational overhead, whereas DAMME achieves these results with a lightweight architecture and no external LLM supervision.

Additionally, DAMME substantially outperforms other dynamic selection methods on both datasets: on MOSI, its Acc-7 ($50.05 \pm 0.40\%$) exceeds KuDA's 47.08% and MODS's 49.27%; on MOSEI, its Acc-7 ($56.13 \pm 0.38\%$) surpasses KuDA's 52.89% and MODS's 54.32%. These results confirm that even on text-dominant datasets, DAMME's dynamic architecture reconstruction effectively leverages complementary information from non-textual modalities.

Table 3 Performance comparison on MOSI.

Method	**Type**	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑
TFN	Symmetric	0.947	0.673	34.46	80.08	80.07
LMF	Symmetric	0.950	0.651	33.82	82.50	82.40
MulT	Symmetric	0.879	0.702	36.91	83.00	82.00
MISA	Symmetric	0.776	0.778	41.37	83.54	83.58
Self-MM	Symmetric	0.708	0.796	46.67	83.44	85.43
CubeMLP	Symmetric	0.755	0.772	43.44	80.76	84.23
MAG-BERT	Text-centric	0.727	0.781	43.62	84.43	84.58
BBFN	Text-centric	0.796	0.744	43.88	80.32	82.44
MMIM	Text-centric	0.744	0.780	44.75	82.51	84.23
ALMT	Text-centric	0.712	0.793	46.79	83.97	85.86
GRCF	Dynamic sel.	0.581	0.866	49.3	90.3	88.5
MSLMamba	Dynamic sel.	0.606	0.855	49.52	87.30	87.13
MODS	Dynamic sel.	0.688	0.798	49.27	85.83	85.96
KuDA	Dynamic sel.	0.705	0.795	47.08	86.43	86.46
DAMME	**Integrated**	**0.597 ± 0.007**	**0.837 ± 0.008**	**50.05 ± 0.40**	**88.63 ± 0.28**	**88.75 ± 0.30**

4.2.3 Contribution of Core Components

To verify the effectiveness of each core module, we conduct ablation experiments on CH-SIMSv2 and MOSI, with results reported in Table 5.

w/o Knowledge Injection: Removing the knowledge-injection pre-training (i.e., omitting the adapters and using only the base encoders) leads to a noticeable performance drop. On CH-SIMSv2, the MAE rises to 0.277 ± 0.006 and Acc-5 declines to

Table 4 Performance comparison on MOSEI.

Method	**Type**	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑
TFN	Symmetric	0.572	0.714	51.60	82.50	82.10
LMF	Symmetric	0.576	0.717	51.59	82.00	82.10
MulT	Symmetric	0.559	0.733	52.84	82.50	82.30
MISA	Symmetric	0.557	0.751	52.05	84.67	84.66
Self-MM	Symmetric	0.531	0.764	53.87	85.15	84.90
CubeMLP	Symmetric	0.537	0.761	53.35	85.23	85.04
MAG-BERT	Text-centric	0.543	0.755	52.67	84.71	84.65
BBFN	Text-centric	0.545	0.760	52.88	85.73	85.56
MMIM	Text-centric	0.550	0.761	51.88	85.42	85.26
ALMT	Text-centric	0.530	0.774	53.62	85.99	86.05
GRCF	Dynamic sel.	0.46	0.835	58.7	89.1	91.5
MSLMamba	Dynamic sel.	0.516	0.791	55.16	86.89	86.72
MODS	Dynamic sel.	0.527	0.772	54.32	85.88	86.14
KuDA	Dynamic sel.	0.529	0.776	52.89	86.46	86.59
DAMME	**Integrated**	**0.498 ± 0.007**	**0.798 ± 0.009**	**56.13 ± 0.38**	**88.35 ± 0.30**	**88.53 ± 0.32**

$62.03 \pm 0.42\%$. This indicates that emotional knowledge injection can provide more reliable unimodal emotional cues for dynamic selection.

w/o Feature Disentanglement: Removing the feature disentanglement module and directly using raw representations for enhancement results in performance degradation. On CH-SIMSv2, Acc-5 drops to $61.43 \pm 0.45\%$, confirming the critical role of the disentanglement module in mitigating redundancy and conflicts.

w/o Dynamic Selection: Removing the dominant modality selection mechanism and fixing sentiment ratios to equal values causes a severe performance drop. On CH-SIMSv2, Acc-5 falls to $59.83 \pm 0.50\%$, demonstrating the importance of dynamic selection in handling variations in modality importance.

w/o Multi-Center Enhancement: Removing the text-, vision-, and audio-centric fusion modules (TFM, VFM, AFM) reduces the model to a fixed text-centric enhancement. On CH-SIMSv2, Acc-5 drops to $58.23 \pm 0.54\%$, indicating that the lack of dedicated enhancers for vision/audio-dominated samples leads to information loss.

w/o Dynamic Routing: Removing the dynamic routing module and forcing simultaneous activation of all three enhancers with simple concatenation of their outputs causes performance degradation. Acc-5 drops to $62.43 \pm 0.40\%$, suggesting that targeted invocation of a single enhancer is more effective than redundant multi-enhancer fusion.

w/o Contrastive Loss: Removing the correlation evaluation loss L_{cor} leads to a slight performance decline. On CH-SIMSv2, Acc-5 falls to $63.35 \pm 0.36\%$, indicating that this loss helps further reinforce the contribution of the dominant modality.

4.2.4 Importance of Different Modalities

To validate the model's robustness under various modality missing conditions, we separately remove the text, vision, and audio modalities, and compare the performance of DAMME against baseline methods (ALMT as a representative of text-centric methods, and CubeMLP as a representative of symmetric methods) on CH-SIMSv2 and MOSI. Results are shown in Table 6.

Table 5 Ablation results of core components.

CH-SIMSv2			
Method	**MAE↓**	**Corr↑**	**Acc-5↑**
DAMME (Full)	0.263 ± 0.005	0.768 ± 0.007	64.18 ± 0.38
w/o Knowledge Injection	0.277 ± 0.006	0.747 ± 0.009	62.03 ± 0.42
w/o Feature Disentanglement	0.281 ± 0.007	0.743 ± 0.010	61.43 ± 0.45
w/o Dynamic Selection	0.294 ± 0.008	0.723 ± 0.012	59.83 ± 0.50
w/o Multi-Center Enhancement	0.304 ± 0.009	0.707 ± 0.013	58.23 ± 0.54
w/o Dynamic Routing	0.274 ± 0.006	0.754 ± 0.008	62.43 ± 0.40
w/o Contrastive Loss	0.269 ± 0.005	0.763 ± 0.007	63.35 ± 0.36
MOSI			
Method	**MAE↓**	**Corr↑**	**Acc-7↑**
DAMME (Full)	0.597 ± 0.007	0.837 ± 0.008	50.05 ± 0.40
w/o Knowledge Injection	0.632 ± 0.009	0.815 ± 0.010	48.23 ± 0.46
w/o Feature Disentanglement	0.638 ± 0.010	0.803 ± 0.011	47.53 ± 0.50
w/o Dynamic Selection	0.654 ± 0.011	0.793 ± 0.012	46.93 ± 0.52
w/o Multi-Center Enhancement	0.668 ± 0.012	0.775 ± 0.013	46.03 ± 0.56
w/o Dynamic Routing	0.621 ± 0.008	0.825 ± 0.009	48.83 ± 0.43
w/o Contrastive Loss	0.613 ± 0.007	0.831 ± 0.008	49.45 ± 0.38

On CH-SIMSv2, after removing any single modality, DAMME's performance degradation is smaller than that of the compared methods. Notably, when the text modality is removed (retaining only V+A), our model still achieves a MAE of 0.406 ± 0.012 and an Acc-5 of $46.83 \pm 0.58\%$, which remains acceptable; in contrast, ALMT and CubeMLP suffer from severe performance deterioration. This demonstrates that DAMME can effectively utilize emotional cues from the remaining modalities, even when the non-dominant modalities can still play a significant role.

On MOSI, removing the text modality causes a substantial performance decline across all methods, which is consistent with the text-dominant nature of this dataset. However, DAMME's degradation is relatively smaller, and its performance with T+V (removing audio) and T+A (removing vision) both exceed those of the compared methods, indicating that even on a text-dominant dataset, our model can more effectively exploit auxiliary modality information.

4.2.5 Training-Stage Dominant Modality Identification Analysis

Since the sentiment ratio is fixed to 1 at the test stage, the dominant modality identification is only meaningful during training. To verify the accuracy of this identification, we calculate the consistency between the dominant modality identified by the sentiment ratio and the unimodal ground-truth labels provided in the CH-SIMSv2 dataset. CH-SIMSv2 supplies independent sentiment labels for vision, text, and audio for each sample. We define the ground-truth dominant modality as the one whose unimodal ground-truth label y_m is closest to the multimodal label y, i.e., the modality with the smallest value of $|y_m - y|$, where $m \in \{t, v, a\}$.

Table 6 Analysis of different modality importance (T: Text, V: Vision, A: Audio).

Method	Modality Combo	CH-SIMSv2 MAE↓	CH-SIMSv2 Acc-5↑	MOSI MAE↓	MOSI Acc-7↑
DAMME	T+V+A	0.263 ± 0.005	64.18 ± 0.38	0.597 ± 0.007	50.05 ± 0.40
	V+A	0.406 ± 0.012	46.83 ± 0.58	1.173 ± 0.038	22.83 ± 0.75
	T+V	0.283 ± 0.006	60.15 ± 0.44	0.645 ± 0.009	47.35 ± 0.45
	T+A	0.286 ± 0.006	59.43 ± 0.46	0.636 ± 0.009	47.83 ± 0.43
ALMT	T+V+A	0.308	52.90	0.712	46.79
	V+A	0.451	39.94	1.437	15.74
	T+V	0.346	48.26	0.772	43.15
	T+A	0.368	46.03	0.736	43.88
CubeMLP	T+V+A	0.334	52.90	0.755	43.44
	V+A	0.452	38.97	1.453	15.45
	T+V	0.354	48.94	0.796	41.98
	T+A	0.362	47.87	0.816	41.54

The statistical results show that the dominant modality identification accuracy of DAMME on the training set reaches 79.2%, which is higher than random selection (33.3%). This confirms that knowledge injection and dynamic attention fusion can effectively capture the reliability of each modality, providing a reliable basis for subsequent architectural reconstruction.

4.2.6 Comparison of Different Fusion Strategies

To validate the advantage of dynamic architectural reconstruction, we compare multiple fusion strategies on CH-SIMSv2:

- **Simple Concatenation**: Directly concatenates the features from all modalities.
- **Additive Fusion**: Sums the features of each modality.
- **Tensor Fusion (TFN)** [19]: Captures high-order interactions via outer product.
- **Low-Rank Fusion (LMF)** [20]: A low-rank version based on tensor decomposition.
- **Cross-Modal Transformer (CMT)** [48]: Employs text-centric or symmetric cross-attention.
- **Dynamic Weighting (KuDA)** [28]: Dynamically adjusts attention weights.
- **Dynamic Architecture Reconstruction (DAMME)**: Invokes dedicated enhancers based on the dominant modality.

The results are presented in Table 7. The dynamic architecture reconstruction strategy achieves strong performance, surpassing dynamic weighting and other static fusion methods, thereby validating the core idea of reconstructing the fusion architecture according to the dominant modality.

Table 7 Comparison of different fusion strategies on CH-SIMSv2.

Fusion Strategy	MAE↓	Corr↑	Acc-5↑	Acc-3↑	Acc-2↑	F1↑
Simple Concatenation	0.341	0.693	50.87	70.32	78.95	78.82
Additive Fusion	0.326	0.708	52.64	71.58	79.73	79.61
Tensor Fusion (TFN)	0.282	0.754	55.80	75.05	80.75	80.64
Low-Rank Fusion (LMF)	0.321	0.711	53.09	72.24	79.50	79.48
Cross-Modal Transformer	0.282	0.745	57.16	74.08	81.43	81.33
Dynamic Weighting (KuDA)	0.271	0.759	61.22	76.21	82.11	82.84
Dynamic Arch. Recon. (DAMME)	**0.263 ± 0.005**	**0.768 ± 0.007**	**64.18 ± 0.38**	**78.65 ± 0.32**	**84.25 ± 0.29**	**84.05 ± 0.32**

5 Conclusion

5.1 Summary of This Work

In this paper, we address the core challenge of dynamic variation in modality dominance within multimodal sentiment analysis by proposing a unified network—DAMME—that integrates dynamic dominant modality identification with multi-center feature enhancement. By integrating dynamic dominant modality identification with multi-center enhancement, our method enables the fusion architecture to adapt to different dominant modalities during training, while the learned attention parameters continue to differentiate modality contributions at inference. Meanwhile, through multi-center enhancement, DAMME achieves architecture-level dynamic reconstruction centered on the dominant modality.

Specifically, DAMME first employs knowledge-injected encoders to extract both global semantic representations and knowledge-aware emotional representations from each modality, and explicitly identifies the dominant modality for the current sample via a sentiment ratio mechanism. To address the issue that ground-truth labels are unavailable during testing, we follow the practice of KuDA and fix the sentiment ratio to 1 at the test stage, allowing the learned attention parameters to determine modality contributions. Subsequently, a feature disentanglement module decomposes the multimodal information into modality-shared and modality-specific subspaces, effectively suppressing redundancy and conflicts. On this basis, during training, according to the identified dominant modality, the corresponding center fusion module (text-centric TFM, vision-centric VFM, or audio-centric AFM) is emphasized to perform targeted feature enhancement with that modality as the query; at inference, the learned attention parameters determine the effective contribution of each enhancer. Finally, hierarchical prediction and a contrastive learning loss further reinforce the contribution of the dominant modality. The entire framework achieves deep integration at both the dynamic selection and architectural reconstruction levels, overcoming the fundamental limitations of existing methods, namely the disconnect between dynamic weighting and static fusion, as well as the singular enhancement paradigm.

Extensive experiments on four public benchmark datasets (CMU-MOSI, CMU-MOSEI, CH-SIMS, CH-SIMSv2) fully validate the effectiveness of DAMME. The experimental results demonstrate that: (1) On CH-SIMS and CH-SIMSv2, where modality importance distributions are relatively balanced, DAMME surpasses various baselines including symmetric fusion, text-centric fusion, and existing dynamic selection methods, proving the advantage of combining dynamic selection with multi-center enhancement;(2) On text-dominated MOSI and MOSEI, DAMME achieves competitive regression accuracy with GRCF while outperforming it on the more challenging fine-grained 7-class classification (50.05% vs. 49.3% on MOSI), and significantly surpasses the recent Neurocomputing publication MSLMamba on all metrics, confirming the effectiveness of our dynamic architecture reconstruction paradigm across diverse dominant-modality scenarios.(3) Ablation studies systematically verify the indispensability of each core module (knowledge injection, feature disentanglement, dynamic selection, multi-center enhancement, dynamic routing, and contrastive loss); (4) Further analyses reveal the model's performance in terms of dominant modality

identification accuracy and comparison of different fusion strategies. Collectively, these results demonstrate that DAMME, through dynamic identification of the dominant modality and reconstruction of the fusion architecture, can more fully utilize multimodal emotional cues, thereby enhancing the robustness and accuracy of sentiment analysis.

5.2 Future Work Outlook

Robustness Extension for Modality-Missing Scenarios:

The current framework assumes that all modalities are completely available during both training and testing phases. However, in real-world applications, modality missing is a common occurrence due to network transmission delays, sensor failures, or privacy protection concerns. In future work, we plan to draw upon ideas from self-distillation, generative adversarial networks, or masked autoencoders to introduce missing-modality imputation or modality-agnostic representation learning mechanisms, enabling the model to maintain stable sentiment prediction capability under arbitrary modality missing conditions.

In recent years, a series of robust methods for modality-missing scenarios have been proposed, providing valuable references for improvements in this direction. Shi et al. [52] proposed a language-dominated fusion and self-distillation framework, which treats language as the dominant modality, leverages its rich emotional cues to integrate auxiliary modalities such as audio and vision, and reconstructs missing information through a unified knowledge distillation approach, thereby significantly improving model performance under both random missing and complete modality conditions. Building on this, Huang et al. [53] proposed SG-DTAM, which combines staged generation with multi-head dynamic time alignment to simultaneously address modality missing and temporal misalignment, demonstrating robust performance in complex scenarios where both issues coexist. Liu et al. [54] proposed DCGRM-Net, which takes a different technical path by leveraging the linear-time modeling capability of the Mamba architecture to construct a lightweight dual-channel guided reconstruction network, effectively overcoming the high computational cost limitations of Transformer-based reconstruction methods. Li et al. [44] further extended the application of Mamba in missing-modality scenarios with TF-Mamba, which aligns and enriches non-textual modalities through a text-aware modality enhancement module while reconstructing missing textual semantics, achieving robust fusion under missing modalities with text as the dominant guide.

The above methods provide solutions to the modality-missing problem from different perspectives, including knowledge distillation, temporal alignment, lightweight reconstruction, and text enhancement. In future work, we plan to consider combining the sentiment-ratio-based dominant modality identification mechanism proposed in this paper with missing-modality imputation mechanisms—when a modality is missing, the remaining modalities' features are used through the predictor to dynamically adjust fusion weights, while drawing upon the self-distillation or generative imputation ideas from the aforementioned methods to generate sentiment-discriminative pseudo-features for the missing modalities, thereby maintaining the advantage of architecture-level dynamic reconstruction even under modality-missing scenarios.

Statements and Declarations

Competing Interests

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.

Funding

[Not applicable]

Data Availability

The datasets used in this study are publicly available (CMU-MOSI, CMU-MOSEI, CH-SIMS, CH-SIMSv2). The code will be made available upon reasonable request.

Author Contribution

Lei Ma: Resources, Supervision, Project administration, Funding acquisition. Yajie Guo: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – original draft, Visualization. Sanli Yi: Supervision. Dangguo Shao: Supervision, Corresponding author.

References

[1] Ali, K., Hughes, C.E.: A Unified Transformer-based Network for multimodal Emotion Recognition (2023). https://arxiv.org/abs/2308.14160

[2] Ezzameli, K., Mahersia, H.: Emotion recognition from unimodal to multimodal analysis: A review. Information Fusion **99**, 101847 (2023) https://doi.org/10.1016/j.inffus.2023.101847 . Accessed 2026-07-23

[3] Xu, P., Zhu, X., Clifton, D.A.: Multimodal Learning With Transformers: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12113–12132 (2023) https://doi.org/10.1109/TPAMI.2023.3275156 . Accessed 2026-07-23

[4] A.v., G., T., M., D., P., E., U.: Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions. Information Fusion **105**, 102218 (2024) https://doi.org/10.1016/j.inffus.2023.102218 . Accessed 2026-07-23

[5] Anwar, K., Shreya, Sharma, M., Saanvi, K.: A review of multimodal sentiment analysis: Taxonomy, issues, challenges, and future perspectives. Computers and Electrical Engineering **131**, 110959 (2026) https://doi.org/10.1016/j.compeleceng.2026.110959 . Accessed 2026-07-23

[6] Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.-S.: NExT-GPT: Any-to-Any Multimodal LLM (2023). https://arxiv.org/abs/2309.05519v3 Accessed 2026-07-23

[7] Pham, H., Liang, P.P., Manzini, T., Morency, L.-P., Póczos, B.: Found in translation: learning robust joint representations by cyclic translations between modalities. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'19/IAAI'19/EAAI'19, vol. 33, pp. 6892–6899. AAAI Press, Honolulu, Hawaii, USA (2019). https://doi.org/10.1609/aaai.v33i01.33016892 . https://dl.acm.org/doi/10.1609/aaai.v33i01.33016892 Accessed 2026-07-23

[8] Li, M., Yang, D., Lei, Y., Wang, S., Wang, S., Su, L., Yang, K., Wang, Y., Sun, M., Zhang, L.: A Unified Self-Distillation Framework for Multimodal Sentiment Analysis with Uncertain Missing Modalities. Proceedings of the AAAI Conference on Artificial Intelligence **38**(9), 10074–10082 (2024) https://doi.org/10.1609/aaai.v38i9.28871 . Accessed 2026-07-23

[9] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv (2019). https://doi.org/10.48550/arXiv.1810.04805 . http://arxiv.org/abs/1810.04805 Accessed 2026-07-23

[10] Guo, J., Tang, J., Dai, W., Ding, Y., Kong, W.: Dynamically Adjust Word Representations Using Unaligned Multimodal Information. In: Proceedings of the 30th ACM International Conference on Multimedia. MM '22, pp. 3394–3402. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3548137 . https://dl.acm.org/doi/10.1145/3503161.3548137 Accessed 2026-07-23

[11] Sun, L., Lian, Z., Liu, B., Tao, J.: Efficient Multimodal Transformer With Dual-Level Feature Restoration for Robust Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing **15**(1), 309–325 (2024) https://doi.org/10.1109/TAFFC.2023.3274829 . Accessed 2026-07-23

[12] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, Ehsan Hoque: Integrating Multimodal Information in Large Pretrained Transformers - ACL Anthology (2020). https://doi.org/10.18653/v1/2020.acl-main.214 . https://aclanthology.org/2020.acl-main.214/ Accessed 2026-07-23

[13] Kim, K., Park, S.: AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis. Information Fusion **92**, 37–45 (2023) https://doi.org/10.1016/j.inffus.2022.11.022 . Accessed 2026-07-23

[14] Baltrušaitis, T., Robinson, P., Morency, L.-P.: OpenFace: An open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of

Computer Vision (WACV), pp. 1–10 (2016). https://doi.org/10.1109/WACV.2016.7477553 . https://ieeexplore.ieee.org/document/7477553 Accessed 2026-07-23

[15] Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP — A collaborative voice analysis repository for speech technologies. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 960–964 (2014). https://doi.org/10.1109/ICASSP.2014.6853739 . https://ieeexplore.ieee.org/document/6853739 Accessed 2026-07-23

[16] Lv, F., Chen, X., Huang, Y., Duan, L., Lin, G.: Progressive Modality Reinforcement for Human Multimodal Emotion Recognition from Unaligned Multimodal Sequences. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2554–2562 (2021). https://doi.org/10.1109/CVPR46437.2021.00258 . https://ieeexplore.ieee.org/document/9578811 Accessed 2026-07-23

[17] Williams, J., Kleinegesse, S., Comanescu, R., Radu, O.: Recognizing Emotions in Video Using Multimodal DNN Feature Fusion. In: Zadeh, A., Liang, P.P., Morency, L.-P., Poria, S., Cambria, E., Scherer, S. (eds.) Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), pp. 11–19. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/W18-3302 . https://aclanthology.org/W18-3302/ Accessed 2026-07-23

[18] Williams, J., Comanescu, R., Radu, O., Tian, L.: DNN Multimodal Fusion Techniques for Predicting Video Sentiment. In: Zadeh, A., Liang, P.P., Morency, L.-P., Poria, S., Cambria, E., Scherer, S. (eds.) Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), pp. 64–72. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/W18-3309 . https://aclanthology.org/W18-3309/ Accessed 2026-07-23

[19] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P.: Tensor Fusion Network for Multimodal Sentiment Analysis. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1103–1114. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/D17-1115 . https://aclanthology.org/D17-1115/ Accessed 2026-07-23

[20] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Bagher Zadeh, A., Morency, L.-P.: Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2247–2256. Association for Computational Linguistics, Melbourne, Australia

(2018). https://doi.org/10.18653/v1/P18-1209 . https://aclanthology.org/P18-1209/ Accessed 2026-07-23

[21] Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R.: Multimodal Transformer for Unaligned Multimodal Language Sequences. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6558–6569. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1656 . https://aclanthology.org/P19-1656/ Accessed 2026-07-23

[22] Hazarika, D., Zimmermann, R., Poria, S.: MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, pp. 1122–1131. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3394171.3413678 . https://doi.org/10.1145/3394171.3413678

[23] Tsai, Y.-H.H., Liang, P.P., Zadeh, A., Morency, L.-P., Salakhutdinov, R.: Learning Factorized Multimodal Representations. arXiv (2019). https://doi.org/10.48550/arXiv.1806.06176 . http://arxiv.org/abs/1806.06176 Accessed 2026-07-23

[24] Gao, M., Zhang, L., Han, S., Hu, H., Jiang, Y., Mai, S.: GRCF: Two-Stage Group-wise Ranking and Calibration Framework for Multimodal Sentiment Analysis. arXiv (2026). https://doi.org/10.48550/arXiv.2601.09606 . http://arxiv.org/abs/2601.09606 Accessed 2026-07-23

[25] Ma, F., Li, S., Yang, F., Xu, G.: MSLMamba: Mamba-driven cross-modal fusion network with modality-specific learning for multimodal sentiment analysis. Neurocomputing **665**, 132164 (2026) https://doi.org/10.1016/j.neucom.2025.132164 . Accessed 2026-07-23

[26] Wafa, A.A., Farhan, M.S., Eldefrawi, M.M.: A unified framework for multimodal emotion recognition across homogeneous and heterogeneous modalities with adaptive fusion. Information Fusion **129**, 104072 (2026) https://doi.org/10.1016/j.inffus.2025.104072 . Accessed 2026-07-23

[27] Yang, D., Li, M., Wu, X., Chen, Z., Jiang, K., Liu, K., Zhai, P., Zhang, L.: Improving Multimodal Sentiment Analysis via Modality Optimization and Dynamic Primary Modality Selection. arXiv (2026). https://doi.org/10.48550/arXiv.2511.06328 . http://arxiv.org/abs/2511.06328 Accessed 2026-07-23

[28] Feng, X., Lin, Y., He, L., Li, Y., Chang, L., Zhou, Y.: Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 14755–14766. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.865 . https://aclanthology.org/2024.findings-emnlp.865/ Accessed 2026-07-23

[29] Wu, H., Sun, Y., Yang, Y., Wong, D.F.: Beyond Simple Fusion: Adaptive Gated Fusion for Robust Multimodal Sentiment Analysis. arXiv (2025). https://doi.org/10.48550/arXiv.2510.01677 . http://arxiv.org/abs/2510.01677 Accessed 2026-07-23

[30] Wang, P., Zhou, Q., Wu, Y., Chen, T., Hu, J.: DLF: Disentangled-Language-Focused Multimodal Sentiment Analysis. Proceedings of the AAAI Conference on Artificial Intelligence **39**(20), 21180–21188 (2025) https://doi.org/10.1609/aaai.v39i20.35416 . Accessed 2026-07-23

[31] Zadeh, A., Zellers, R., Pincus, E., Morency, L.-P.: MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. arXiv (2016). https://doi.org/10.48550/arXiv.1606.06259 . http://arxiv.org/abs/1606.06259 Accessed 2026-07-23

[32] Bagher Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.-P.: Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2236–2246. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1208 . https://aclanthology.org/P18-1208/ Accessed 2026-07-23

[33] Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., Yang, K.: CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3718–3727. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.343 . https://aclanthology.org/2020.acl-main.343/ Accessed 2026-07-23

[34] Liu, Y., Yuan, Z., Mao, H., Liang, Z., Yang, W., Qiu, Y., Cheng, T., Li, X., Xu, H., Gao, K.: Make Acoustic and Visual Cues Matter: CH-SIMS v2.0 Dataset and AV-Mixup Consistent Module. arXiv (2022). https://doi.org/10.48550/arXiv.2209.02604 . http://arxiv.org/abs/2209.02604 Accessed 2026-07-23

[35] Mai, S., Zeng, Y., Zheng, S., Hu, H.: Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing **14**(3), 2276–2289 (2023) https://doi.org/10.1109/TAFFC.2022.3172360 . Accessed 2026-07-23

[36] Arevalo, J., Solorio, T., Montes-y-Gómez, M., González, F.A.: Gated Multimodal Units for Information Fusion. arXiv (2017). https://doi.org/10.48550/arXiv.1702.01992 . http://arxiv.org/abs/1702.01992 Accessed 2026-07-23

[37] Wenyan Wu, Qintai Hu, Xingbo Shen, Zhenwei Zhang: Dynamically Adaptive Multimodal Fusion Transformer for Multimodal Sentiment Analysis

\textbackslashtextbar IEEE Journals & Magazine \textbackslashtextbar IEEE Xplore. https://doi.org/10.1109/ACCESS.2025.3628494 . https://ieeexplore.ieee.org/document/11224493 Accessed 2026-07-23

[38] Jianing Zhao & Qun Jin: Dynamic Dominate-Modal-Aware Model for Multimodal Sentiment Analysis \textbackslashtextbar Springer Nature Link. https://doi.org/10.1007/978-3-032-12657-3_7 . https://link.springer.com/chapter/10.1007/978-3-032-12657-3_7 Accessed 2026-07-23

[39] Abdulhalim, S., Albaghdadi, M., Farazi, M.: Multi-Modal Sentiment Analysis with Dynamic Attention Fusion. arXiv (2025). https://doi.org/10.48550/arXiv.2509.22729 . http://arxiv.org/abs/2509.22729 Accessed 2026-07-23

[40] Tuerhong, G., Fu, F., Wushouer, M.: Adaptive multimodal transformer based on exchanging for multimodal sentiment analysis. Scientific Reports **15**(1), 27265 (2025) https://doi.org/10.1038/s41598-025-11848-4 . Accessed 2026-07-23

[41] Li, Y., Wang, Y., Cui, Z.: Decoupled Multimodal Distilling for Emotion Recognition. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6631–6640 (2023). https://doi.org/10.1109/CVPR52729.2023.00641 . https://ieeexplore.ieee.org/document/10203191 Accessed 2026-07-23

[42] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv (2023). https://doi.org/10.48550/arXiv.1706.03762 . http://arxiv.org/abs/1706.03762 Accessed 2026-07-23

[43] Vamsidhar, D., Desai, P., Shahade, A.K., Patil, S., Deshmukh, P.V.: Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis. Scientific Reports **15**(1), 25440 (2025) https://doi.org/10.1038/s41598-025-09000-3 . Accessed 2026-07-23

[44] Li, X., Cheng, X., Miao, D., Zhang, X., Li, Z.: TF-Mamba: Text-enhanced Fusion Mamba with Missing Modalities for Robust Multimodal Sentiment Analysis. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 11252–11267. Association for Computational Linguistics, Suzhou, China (2025). https://doi.org/10.18653/v1/2025.findings-emnlp.602 . https://aclanthology.org/2025.findings-emnlp.602/ Accessed 2026-07-23

[45] Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. Proceedings of the AAAI Conference on Artificial Intelligence **35**(12), 10790–10797 (2021) https://doi.org/10.1609/aaai.v35i12.17289 . Accessed 2026-07-23

[46] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, Lanfen Lin: CubeMLP: An MLP-based Model for Multimodal Sentiment Analysis and Depression Estimation

\textbackslashtextbar Proceedings of the 30th ACM International Conference on Multimedia (2022). https://doi.org/10.1145/3503161.3548025 . https://dl.acm.org/doi/10.1145/3503161.3548025 Accessed 2026-07-23

[47] Huang, J., Zhou, J., Tang, Z., Lin, J., Chen, C.Y.-C.: TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis. Knowledge-Based Systems **285**, 111346 (2024) https://doi.org/10.1016/j.knosys.2023.111346 . Accessed 2026-07-23

[48] Han, W., Chen, H., Gelbukh, A., Zadeh, A., Morency, L.-p., Poria, S.: Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In: Proceedings of the 2021 International Conference on Multimodal Interaction. ICMI '21, pp. 6–15. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3462244.3479919 . https://dl.acm.org/doi/10.1145/3462244.3479919 Accessed 2026-07-23

[49] Han, W., Chen, H., Poria, S.: Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In: Moens, M.-F., Huang, X., Specia, L., Yih, S.W.-t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 9180–9192. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.723 . https://aclanthology.org/2021.emnlp-main.723/ Accessed 2026-07-23

[50] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He and X. Gao: Cross-Modal Enhancement Network for Multimodal Sentiment Analysis \textbackslashtextbar IEEE Journals & Magazine \textbackslashtextbar IEEE Xplore (2023). https://doi.org/10.1109/TMM.2022.3183830 . https://ieeexplore.ieee.org/document/9797846 Accessed 2026-07-23

[51] Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., Yu, T.: Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 756–767. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.emnlp-main.49 . https://aclanthology.org/2023.emnlp-main.49/ Accessed 2026-07-23

[52] Shi, P., Wang, J., Zhang, X.: Language-dominated fusion and self-distillation for multimodal sentiment analysis with incomplete modalities. Pattern Recognition **172**, 112644 (2026) https://doi.org/10.1016/j.patcog.2025.112644 . Accessed 2026-07-23

[53] Huang, D., Gao, R., Zhang, G., Yu, J.: SG-DTAM: Joint staged generation and dynamic time alignment for missing and unaligned modalities in sentiment analysis. Expert Systems with Applications **299**, 129750 (2026) https://doi.org/10.1016/j.eswa.2025.129750 . Accessed 2026-07-23

[54] Liu, S., Zhao, H., Chen, Y., Kong, F., Li, K.: DCGRM-Net: Dual-Channel Guided Reconstruction Mamba Network for robust multimodal sentiment analysis. Information Processing & Management **63**(2, Part B), 104491 (2026) https://doi.org/10.1016/j.ipm.2025.104491 . Accessed 2026-07-23