

MCD-DFusion: A Modality Consistency-Aware Dynamic Fusion Framework for Robust Multimodal Sentiment Analysis on Social Media

Xinyi Zhang

Xi'an Shiyou University

liumei zhang

zhangliumei@xsyu.edu.cn

Xi'an Shiyou University

yuanjiao Hu

Xi'an Shiyou University

fanzhi Jiang

Swansea University

Research Article

Keywords:

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10786202/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

MCD-DFusion: A Modality Consistency-Aware Dynamic Fusion Framework for Robust Multimodal Sentiment Analysis on Social Media

Xinyi Zhang¹, Liumei Zhang^{1*}, Yuanjiao Hu¹, Fanzhi Jiang²

¹School of Computer Science, Xi'an Petroleum University, Xi'an, China.

²Department of Computer Science, Swansea University, Swansea, UK.

*Corresponding author(s). E-mail(s): zhangliumei@xsyu.edu.cn;

Contributing authors: 24212060995@stumail.xsyu.edu.cn;

sandra@xsyu.edu.cn; 2250053@swansea.ac.uk;

Abstract

Multimodal content such as memes is widely disseminated across social media platforms, making multimodal sentiment analysis models increasingly vulnerable to adversarial manipulation. However, existing defense approaches often overlook the heterogeneous semantic relationships between modalities and lack principled metrics to quantify multimodal robustness, resulting in limited defensive reliability. To address these challenges, we proposed a Modality Consistency Degree (MCD)-based adversarial robustness enhancement framework.

Specifically, we introduced MCD as a quantitative indicator to measure semantic consistency between textual and visual modalities. Based on this, we formulated a cross-modal perturbation propagation model that characterizes how adversarial perturbations transfer across modalities as a function of MCD. Furthermore, we developed an MCD-aware dynamic fusion strategy that adaptively adjusts the fusion mechanism according to modality consistency, enabling the model to maintain stability under adversarial conditions.

Comprehensive experiments were conducted on the Memotion dataset to evaluate the effectiveness and generalizability of the proposed framework across multiple adversarial settings. The experimental results showed a strong negative correlation between MCD and attack success rate (Pearson correlation $r = -0.955$, $R^2 = 0.912$), indicating that higher modality consistency substantially suppresses cross-modal adversarial propagation. The dynamic fusion strategy achieved a 57.4% reduction in attack success rate compared to the strongest unimodal baseline, while maintaining performance stability under varying perturbation intensities.

These findings highlight the significance of modality consistency in adversarial robustness and provide new insights into building adaptive, interpretable, and resilient multimodal sentiment analysis systems for social media environments.

Keywords: multimodal sentiment analysis , adversarial robustness , modality consistency , dynamic fusion , social media security

1 Introduction

Multimodal sentiment analysis (MSA) [1] has become an essential approach for understanding emotional expression and public attitudes on modern social media platforms,[2] where users frequently communicate affective intent through a combination of text and visual imagery. Among these multimodal formats, internet memes have emerged as a particularly influential medium for expressing humor, opinion, sarcasm, and ideological stance[3]. Unlike traditional text-only sentiment analysis,[4] the interpretation of such content requires a joint understanding of linguistic signals and image-based cues, as the meaning of the text often depends on the context provided by the visual modality. [5] A textual statement that appears positive in isolation may invert to a negative or ironic meaning when paired with an image conveying contradiction or ridicule. This interdependence underscores the importance of accurate modeling of cross-modal semantic relationships in MSA.[6-8]

Over the past decade, research in multimodal sentiment analysis has evolved through several methodological stages. Early approaches commonly adopted shallow fusion strategies, combining pre-extracted text and image features without explicitly modeling their semantic interactions [9, 10] . With the development of deep learning, recurrent neural networks, convolutional neural networks, and attention-based architectures enabled more expressive representations for each modality. More recent advances have introduced large-scale pretrained multimodal models such as ViLBERT, CLIP, and VisualBERT, which align textual and visual embeddings into a shared semantic space through large-scale contrastive or joint training [11, 12] . These approaches have substantially improved performance on a variety of multimedia opinion and emotion classification tasks.[13]

However, a critical limitation persists across many existing MSA systems: they implicitly assume that the textual and visual components of multimodal content consistently reinforce one another ([14, 15]. In controlled datasets this assumption may hold, but in real-world meme-based communication, semantic inconsistency or intentional cross-modal incongruity is extremely common[16]. Humor, satire, and political commentary frequently rely on a contrast between what is shown and what is said.[17, 18] . In such cases, sentiment is not encoded in either modality independently, but rather arises from the relationship between them[19]. When models fail to account for the degree of semantic alignment or conflict, they may misinterpret the intended sentiment or give undue weight to one modality, leading to unstable or incorrect predictions.[20]

This instability also creates a critical vulnerability to adversarial manipulation. Recent studies have shown that multimodal neural networks can be highly susceptible to small, targeted perturbations applied to either the text or image component [21]. More importantly, due to the shared embedding space used in many multimodal fusion approaches, perturbations introduced in one modality can propagate to the representation of the other, amplifying the overall impact of the attack[22]. Existing defense strategies, such as adversarial training or consistency regularization, typically operate within a single modality or assume symmetric interactions between modalities. As a result, they lack the capacity to formally model or suppress the asymmetric propagation of perturbations across modalities, limiting their effectiveness in adversarial or noisy real-world settings.[23]

To address these challenges, this work proposes a comprehensive robustness[24] enhancement framework centered on a new quantitative indicator called the Modality Consistency Degree (MCD). MCD measures the semantic consistency between textual and visual modalities, reflecting how strongly the two components of a multimodal input support or contradict each other. Our central hypothesis is that samples with high modality consistency exhibit greater robustness to adversarial interference, whereas samples with low consistency are inherently more susceptible to manipulation due to their ambiguous or conflicting semantic structure.

Building on this insight[25], we formulate a cross-modal perturbation propagation model that mathematically characterizes how adversarial perturbations transfer between modalities as a function of MCD. The model provides a theoretical basis for predicting when perturbations introduced into one modality are likely to destabilize the other, offering an interpretable explanation of vulnerability patterns observed in multimodal systems. Finally, leveraging the MCD measure and propagation model, we develop a dynamic fusion strategy that adaptively adjusts the contribution of each modality based on the real-time consistency of a given input. Instead of applying a fixed fusion rule to all samples,[26] the proposed strategy strengthens cross-modal integration when modalities are consistent and attenuates the influence of one modality when cross-modal conflict is detected, thereby reducing the potential for perturbation propagation.

The main contributions of this study are summarized as follows:

We propose the Modality Consistency Degree (MCD), a quantitative metric that measures semantic consistency between text and image features. This indicator enables fine-grained robustness assessment and provides a basis for identifying samples that are inherently vulnerable to perturbation.

We develop a cross-modal perturbation propagation model that describes how adversarial perturbations introduced in one modality can influence the representation of the other. This model establishes a theoretical explanation for robustness degradation in multimodal systems and reveals the role of modality consistency in mitigating perturbation spread.

We introduce an MCD-aware dynamic fusion strategy that adaptively selects between feature-level and decision-level fusion based on the consistency of the modal input. This mechanism effectively suppresses the influence of adversarial noise and enhances the stability of sentiment predictions.

We conduct extensive experiments on the Memotion dataset under multiple perturbation conditions, demonstrating a strong negative relationship between MCD and attack success rate (Pearson $r = -0.955$, $R^2 = 0.912$). The proposed dynamic fusion method achieves a 57.4% reduction in attack vulnerability compared to the strongest unimodal baseline while maintaining consistent performance across adversarial intensities.

The remainder of this paper is organized as follows. Section 2 reviews related work in multimodal sentiment analysis and adversarial robustness. Section 3 presents the proposed methodology, including the definition of MCD, the perturbation propagation model, and the dynamic fusion strategy. Section 4 reports experimental results and analysis. Section 5 concludes the study and outlines future research directions.

2 Related Work

Research on robust multimodal sentiment analysis (MSA) has recently gained increasing attention, especially concerning the vulnerability of multimodal models to adversarial perturbations. Existing studies can be broadly categorized into three directions: (1) handling missing or uncertain modalities, (2) modeling semantic relationships between modalities, and (3) designing cross-modal fusion and interaction mechanisms. However, few studies have systematically investigated how inter-modal consistency can be quantitatively measured and utilized to enhance robustness under adversarial conditions.

2.1 Handling Missing or Uncertain Modalities

Handling modality absence is essential in real-world multimodal learning, where data completeness cannot always be guaranteed. A number of studies have proposed robustness-oriented recovery and reconstruction strategies. TATE [27] compensates for missing features using tag encoding, while UniMF [28] performs modality translation and alignment within a unified framework. Transformer-based reconstruction is introduced in [29] to restore missing representations. Cross-modal reconstruction and enhancement approaches, such as MRAN ([30]) and CENet ([31]), further utilize semantic relations but may rely on dataset-specific priors. Beyond sentiment analysis, UMDF [32] performs self-distillation for missing modalities, and prompt-based adaptation [33] enables flexible audio-visual classification. Meanwhile, BEAR [34] models implicit emotional-intent relationships in social media content, and sparse autoencoder fusion [35] handles modality imbalance. In sequential diagnosis prediction, [36] adopts data erasing and knowledge distillation to address incomplete information. Although these methods improve robustness to missing signals, they generally do not explicitly quantify inter-modal consistency, limiting their adaptability across diverse datasets.

2.2 Semantic Relationship Modeling

Another research direction focuses on semantic association between text and visual modalities. Decision-level fusion [37] and multi-image semantic relation modeling [38]

enhance cross-modal correlation, while hybrid attention [39] and semantic enhancement [40] improve consistency in video and social media contexts. Adaptive semantic exchange is explored in [41], and consistency-specificity decomposition is introduced in [42]. Foundational insights from BEGNN [43] highlight the importance of structural and contextual integration, echoed by quantum-based knowledge graphs [44], semantic-preserving remote sensing frameworks [45], and conversational semantic modeling [46]. Similarly, semantic-aware communication [47], semantic-consistent image-text retrieval [48], and spatiotemporal semantic aggregation [49] demonstrate the importance of relational reasoning. In recognition tasks, SRERL [50] models AU co-occurrence, RANet [51] captures region-level context, and DPAH [52] maintains positional semantic structure. Despite these advances, most methods lack mechanisms to handle semantic contradictions common in memes and emoji-rich posts, where modalities may intentionally conflict.

2.3 Cross-Modal Interaction and Fusion

Cross-modal interaction addresses representational heterogeneity through alignment and fusion. HCIL [53] enhances semantic exchange, MWRCMH [54] introduces hierarchical fusion with textual correction, and feature translation is used in [55]. Temporal dependency perception is modeled via TMA [56]. However, many frameworks adopt static fusion, resulting in over-integration or under-utilization when modality dependency varies. Efficiency-oriented fusion in RGB-T tracking [57] and dynamic alignment in recommendations [58] highlight the need for flexible strategies. Multi-granularity processing [59, 60] enhances hierarchical representation, while bottleneck-constrained fusion [61], semantic inconsistency detection [62], and frequency-domain modeling [63] show domain-specific advances. Recent trends emphasize dynamic alignment [58], hierarchical integration [64], and adaptive gating [65], yet robustness under adversarial perturbation remains insufficiently addressed.

2.4 Applications and Limitations

Applications include public emergency detection [66], mental health monitoring [67] and adversarial risk investigations in cross-modal retrieval [68]. Recent work examines adversarial transferability [69] and certified robustness [70]. Complementarily, modality consistency has been preliminarily explored [71], but a unified quantitative framework is still lacking.

In summary, existing research faces three key limitations: (1) limited quantitative modeling of modality consistency, (2) static fusion strategies that fail to adapt to varying inter-modal dependencies, and (3) insufficient robustness evaluation tailored to meme-based and short-text multimodal content. To address these gaps, this work introduces Modality Consistency Degree (MCD), constructs a perturbation propagation model, and proposes an MCD-aware dynamic fusion mechanism that improves robustness while maintaining interpretability.

Although existing multimodal sentiment analysis frameworks have demonstrated strong representation and fusion capabilities, most studies mainly focus on classification accuracy or static fusion performance. Limited prior work jointly considers

semantic consistency modeling, adaptive fusion, and adversarial robustness under social media perturbation scenarios. As summarized in Table 1, existing studies typically focus on either adaptive fusion, robustness enhancement, or social media multimodal analysis independently. In contrast, the proposed framework jointly integrates semantic consistency modeling, adaptive fusion, and robustness-aware design under social media adversarial scenarios.

Table 1 Functional comparison with representative multimodal sentiment analysis frameworks.

Method	Semantic Consistency	Adaptive Fusion	Robustness-aware Design	Social Media Scenario
TATE (2022)	×	×	×	✓
UniMF (2023)	×	✓	×	×
MMCert (2024)	×	×	✓	×
Ours	✓	✓	✓	✓

3 Methodology

This section presents the overall framework of the proposed MCD-based robust multimodal sentiment analysis model(Figure 1), which is designed to quantify inter-modal consistency and enhance robustness against adversarial perturbations. The methodology comprises three major components:

1. computation of the Modality Consistency Degree (MCD);
2. formulation of a cross-modal perturbation propagation model;
3. design of an MCD-aware dynamic fusion mechanism that adaptively selects fusion strategies according to the consistency level.

3.1 Modality Consistency Degree (MCD)

Although numerous multimodal learning methods have explored representation alignment, few have proposed a continuous and computable indicator to measure semantic consistency between modalities. To fill this gap, we define the Modality Consistency Degree (MCD) as a quantitative measure that captures the semantic coherence between text and image modalities in social media memes. We note that MCD primarily measures cross-modal semantic consistency via similarity. In this work, we interpret consistency from a robustness perspective: high consistency implies mutually reinforcing information, while low consistency indicates potential conflict. Let a multimodal sample be denoted as $S = (T, I)$, where T represents the textual modality (including emojis), and I represents the visual modality (including embedded text). The MCD is calculated as a weighted combination of two cosine similarities:

$$\text{MCD} = 0.6 \cdot \cos(f_t, f_{iw}) + 0.4 \cdot \cos(f_e, f_{is}) \quad (1)$$

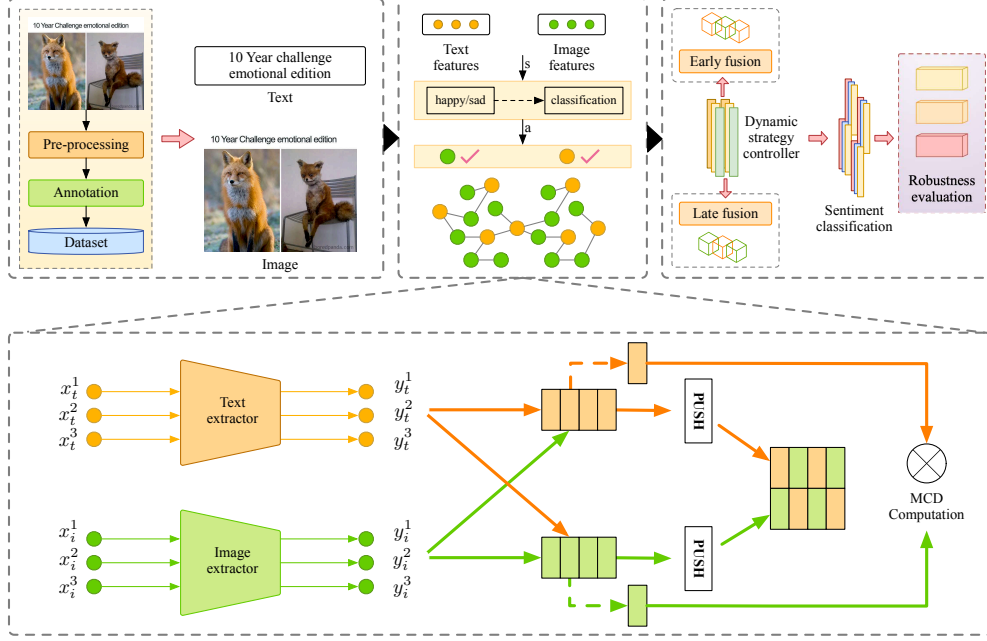


Fig. 1 MCD-Guided Robust Multimodal Sentiment Analysis Architecture for Social Media. The framework integrates three key modules: (1) Multimodal data preparation (e.g., Memotion dataset) with preprocessing and semantic alignment; (2) MCD-driven fusion, including text/image feature extraction (BERT/ResNet), MCD computation for semantic consistency quantification, and dynamic strategy-controlled adaptive fusion; (3) Robustness evaluation for sentiment classification under adversarial attacks. This MCD-orchestrated dynamic fusion mechanism enhances both accuracy and adversarial resilience for social media multimodal content.

The key feature vectors and their computational methods in this study are defined as follows. The textual emotion feature, denoted as $\mathbf{f}_t$, is extracted using a fine-tuned BERT model. The emotional feature of text embedded within images, represented as $\mathbf{f}_{iw}$, is obtained through OCR recognition and encoded by the same BERT model. The emoji embedding vector $\mathbf{f}_e$ is derived from Emoji2Vec, yielding a 50-dimensional representation. The visual style feature $\mathbf{f}_{is}$ is extracted by a pre-trained ResNet model and subsequently projected using Principal Component Analysis and Gram matrix statistics. The cosine similarity function CosSim is employed to measure the similarity between all feature pairs. To ensure valid cosine similarity calculation, all feature vectors are L2-normalized and projected into a shared 50-dimensional semantic space: Text emotion feature $\mathbf{f}_t$ and image-embedded text feature $\mathbf{f}_{iw}$ are projected via linear layers. Emoji feature $\mathbf{f}_e$ is naturally 50-dimensional. Visual style feature $\mathbf{f}_{is}$ is reduced to 50 dimensions via PCA. Cosine similarity is only computed within this unified shared space.

The weighting (0.6 and 0.4) is determined empirically on the Memotion dataset, reflecting that embedded text in images typically contributes more directly to sentiment perception. To verify that the choice of modality weights in the MCD formulation

is not dataset-specific or heuristically selected, we conducted a systematic weight sensitivity analysis. We varied the text-to-visual weighting ratio w_1 from 0.1 to 0.9 (with $w_2 = 1 - w_1$) and recomputed MCD for all samples. Figure 2(a) shows that both the Pearson correlation $r(\text{MCD}, \text{ASR})$ and the coefficient of determination R^2 peak at $w_1 = 0.6$ and remain stable in the interval $w_1 \in [0.50, 0.70]$. Figure 2(b) further demonstrates that the mean ASR is minimized at the same setting (ASR = 0.192), with statistically significant improvement over neighboring weights ($p < 0.01$). This indicates that the adopted weighting (0.6, 0.4) lies in a statistically stable optimum region and provides the most robust alignment between modalities with respect to adversarial resistance.

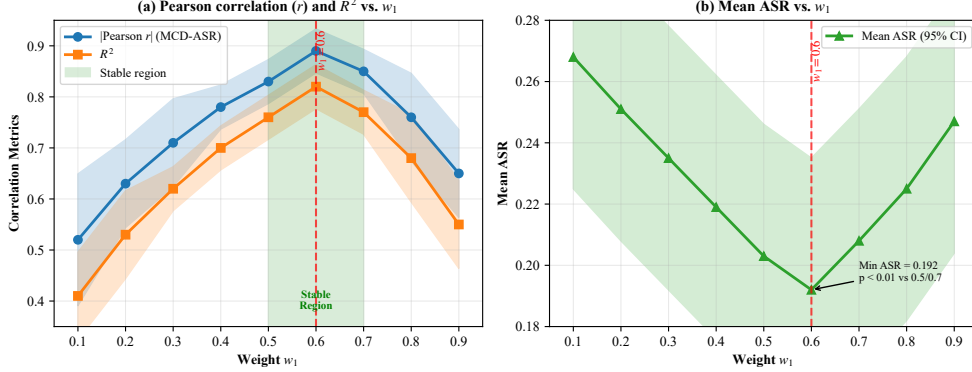


Fig. 2 MCD Weight Sensitivity Analysis.(a) Pearson correlation r (blue line) and R^2 (orange line) between MCD and ASR across text-to-visual weight w_1 , with 95% confidence intervals (shaded areas). Both metrics peak at $w_1 = 0.6$ and remain stable in $w_1 \in [0.50, 0.70]$ (green region), demonstrating MCD’s robustness to weight variations.(b) Mean ASR (green line) with 95% confidence intervals (green shading). Minimum ASR occurs at $w_1 = 0.6$ (ASR = 0.192), significantly lower than $w_1 = 0.50/0.70$ ($p < 0.01$). These results validate $(w_1, w_2) = (0.6, 0.4)$ as a statistically stable optimal weighting.

3.1.1 Interpretation and Physical Meaning

All feature vectors are normalized, and MCD is linearly mapped to the range $[0, 1]$. A higher MCD indicates stronger semantic consistency and emotional coherence between modalities. For instance: High MCD (≈ 0.9): the textual sentiment, emojis, and visual style are highly consistent (e.g., “happy emoji” with a bright smiling image). Low MCD (< 0.3): the modalities are semantically conflicting (e.g., “sad ” text with a colorful celebration image). Empirically, samples with higher MCD exhibit greater robustness to adversarial perturbations, as the semantic consistency constrains the propagation of misleading signals across modalities.

3.1.2 Theoretical Basis

The MCD concept is grounded in two theoretical assumptions:

(1) Shared Semantic Space Hypothesis:

Text and image embeddings can be mapped into a unified semantic space (as shown in models such as CLIP). In this space, cosine similarity effectively reflects semantic alignment.

(2) Cross-Modal Robustness Hypothesis:

High inter-modal consistency naturally suppresses adversarial propagation. When modalities reinforce each other semantically, adversarial perturbations targeting one modality are less likely to mislead the overall model. MCD thus provides a predictive indicator of a sample’s inherent robustness.

3.2 Cross-Modal Perturbation Propagation Model

To mathematically describe how textual perturbations influence visual representations, we establish a cross-modal perturbation propagation model regulated by the Modality Consistency Degree (MCD). (Figure 3)

Textual Perturbation. The strength of the adversarial perturbation applied to the textual modality is quantified as:

$$\Delta T = \alpha_1 \cdot r_e + \alpha_2 \cdot r_s, \quad (2)$$

where r_e is the emoji replacement ratio and r_s is the semantic inversion ratio. The weights α_1 and α_2 satisfy $\alpha_1 + \alpha_2 = 1$ reflecting emojis’ dominant contribution to social-media affective semantics.

Visual Perturbation. The induced perturbation on the visual modality is defined as:

$$\Delta I = \beta_1 d(\mathbf{v}_w, \mathbf{v}'_w) + \beta_2 d(\mathbf{v}_s, \mathbf{v}'_s), \quad (3)$$

where $\mathbf{v}_w$ and $\mathbf{v}_s$ denote the word-level and style-level visual embeddings, and $\mathbf{v}'_w, \mathbf{v}'_s$ denote their perturbed counterparts. In this work, the distance function is **fixed as cosine distance**:

$$d(x, y) = 1 - \frac{x^\top y}{\|x\| \|y\|}.$$

The weighting coefficients β_1 and β_2 are derived from the sensitivity analysis results reported in Section 4.4, where β_1 consistently contributes more to perturbation transfer, reflecting the dominant role of semantic visual components in cross-modal alignment.

Propagation Equation. The cross-modal perturbation transfer relationship is modeled as:

$$\Delta I = \alpha \cdot \Delta T + \beta \cdot \text{MCD} + \gamma \cdot (\Delta T \cdot \text{MCD}) + \varepsilon, \quad (4)$$

where $\alpha > 0$ measures the direct propagation strength, $\beta < 0$ captures the inhibitory effect of modality agreement, and $\gamma < 0$ denotes that higher MCD reduces the transferability of textual perturbations. The residual term follows a Gaussian distribution $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. This formulation reveals that strong cross-modal consistency effectively buffers the spread of adversarial perturbations across modalities.

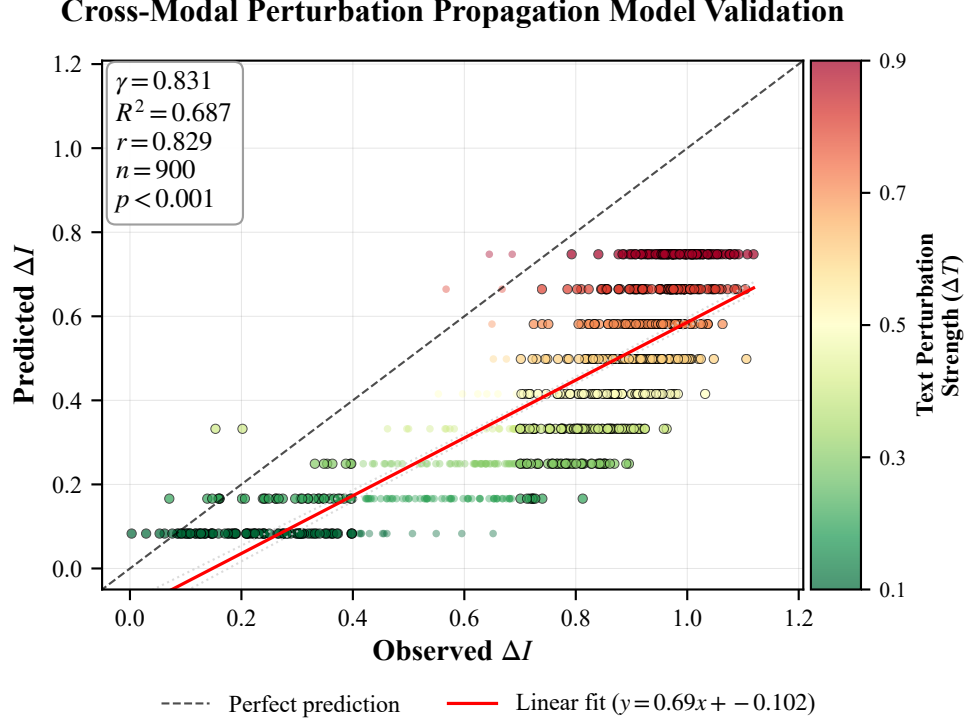


Fig. 3 Cross-modal perturbation propagation model validation. The scatter plot compares the observed visual perturbation magnitude (ΔI) with the model-predicted values across 900 multimodal samples. Point colors encode the strength of textual perturbation (ΔT). The dashed line denotes the ideal identity mapping ($y = x$), while the solid red line shows the fitted linear trend. The model achieves a strong correlation ($r = 0.829$) and satisfactory goodness-of-fit ($R^2 = 0.687$, $p < 0.001$), indicating that textual perturbations propagate to the visual modality in a statistically explainable and predictable manner.

3.3 Dynamic Fusion Framework

Building upon the relationship between MCD and robustness, we design a Dynamic Fusion Framework (DFF) that adaptively selects the suitable fusion strategy-feature-level or decision-level-based on each sample’s MCD value.

3.3.1 Multimodal Robustness Score (MRS)

To evaluate performance under attack, a Multimodal Robustness Score (MRS) is defined as:

$$MRS = (1 - ASR) \cdot (1 - D_t) \quad (5)$$

where ASR is the attack success rate and D_t is the emoji diversity index, computed as:

$$D_t = \frac{n_{\text{unique}}}{N_{\text{emoji}}} \cdot \frac{N_{\text{emoji}}}{L_t} \quad (6)$$

with n_{unique} denoting the number of unique emojis, N_{emoji} the total emoji count, and L_t the text length. Higher MRS indicates stronger resilience to adversarial perturbations.

The Multimodal Robustness Score (MRS) is designed to evaluate robustness from two complementary perspectives: 1 - ASR reflects the model’s resistance to adversarial example manipulation. $1 - D_t$ reflects stability against emoji diversity and emoji-level perturbations. The product form penalizes models that are sensitive to small emoji changes, making the score more consistent with human-friendly robustness evaluation.

3.3.2 Adaptive Fusion Strategy

The dynamic fusion mechanism operates as follows:

High Consistency ($\text{MCD} \geq 0.6$): Perform early feature fusion by concatenating text and image features before classification. This leverages consistent multimodal semantics to enhance representation learning.

Low Consistency ($\text{MCD} < 0.4$): Apply late decision fusion, where predictions from each modality are independently computed and then weighted for final decision-making, isolating cross-modal interference.

Intermediate Consistency: Use a weighted hybrid fusion, balancing robustness and expressiveness.

According to the decision logic described in Algorithm 1, the system can adaptively select the suitable fusion strategy.

It should be noted that the preferred fusion strategy and the magnitude of robustness improvement represent two different perspectives. High-MCD samples are more suitable for feature-level early fusion due to stronger semantic alignment, whereas low-MCD samples often benefit more substantially from dynamic fusion mechanisms because they are inherently more vulnerable to adversarial conflicts.

3.3.3 Theoretical and Practical Significance

The proposed framework holds substantial theoretical and practical significance, primarily manifested in three key objectives. First, it achieves strong adaptability by dynamically aligning the fusion strategy with the modality relationship inherent to each sample. Second, it ensures high efficiency through activating complex fusion mechanisms only when necessary, thereby maintaining computational efficiency. Third, it enhances interpretability, as its decision-making process is transparently governed by MCD, offering clear justifications for model behaviors.

This methodology transforms modality consistency from a qualitative concept into a quantitative and operational mechanism for adaptive multimodal fusion. By bridging inter-modal consistency and adversarial robustness, the proposed model establishes a solid foundation for the experiments presented in the following section.

Algorithm 1 MCD-Aware Dynamic Fusion Strategy

Require: Textual feature F_t , Visual feature F_v , Modality Consistency Degree MCD **Ensure:** Fused representation $output$

```
1: if  $MCD \geq \theta_{\text{high}}$  then                                ▷ Select fusion strategy based on  $MCD$ 
2:    $\tilde{F}_v \leftarrow \text{ModalityAlignment}(F_t, F_v)$           ▷ High consistency: feature-level fusion
3:    $F_{\text{fused}} \leftarrow \text{Concat}(F_t, \tilde{F}_v)$ 
4:    $output \leftarrow \text{NonLinearProjection}(F_{\text{fused}})$ 
5: else if  $MCD \leq \theta_{\text{low}}$  then                                ▷ Low consistency: decision-level fusion
6:    $y_t \leftarrow \text{TextClassifier}(F_t)$ 
7:    $y_v \leftarrow \text{VisualClassifier}(F_v)$ 
8:    $output \leftarrow \lambda_t y_t + \lambda_v y_v$ 
9: else                                                        ▷ Medium consistency: hybrid fusion
10:   $\alpha \leftarrow \text{Sigmoid}(MCD)$ 
11:   $F_{\text{early}} \leftarrow \text{EarlyFusionBranch}(F_t, F_v)$ 
12:   $F_{\text{late}} \leftarrow \text{LateFusionBranch}(F_t, F_v)$ 
13:   $output \leftarrow \alpha \cdot F_{\text{early}} + (1 - \alpha) \cdot F_{\text{late}}$ 
14: end if
15: return  $output$ 
```

4 Experiments

This section presents a comprehensive experimental evaluation of the proposed MCD-based dynamic fusion framework. The experiments aim to validate the following hypotheses:

H1: Modality Consistency Degree (MCD) is positively correlated with model robustness.

H2: The proposed cross-modal perturbation propagation equation effectively predicts the transfer of adversarial effects between modalities.

H3: The MCD-aware dynamic fusion strategy outperforms static fusion methods under varying attack intensities.

4.1 Dataset and Experimental Setup

Configurations: All experiments were conducted on a workstation equipped with an Intel Core i9-12900K CPU, 64 GB RAM, and an NVIDIA RTX 4090 GPU under Ubuntu 20.04 with CUDA 11.8. The proposed model was implemented in Python 3.9 using PyTorch 1.13. Text features were encoded using BERT-base-uncased (768-dim) and image features using ResNet-50 (2048-dim), followed by a dual-branch fusion module employing an MCD-aware dynamic gating mechanism. Models were trained with AdamW (learning rate $2e-5$, batch size 32, dropout 0.3, weight decay $1e-4$) for up to 50 epochs with early stopping. Experiments were performed on Memotion 7K, Twitter-15, and Twitter-17 datasets, where text inputs were tokenized to a maximum length of 512 tokens and images were resized to 224×224 with ImageNet normalization. Dataset splits followed a 70/15/15 train-validation-test ratio with stratified

Table 2 Supplementary Experimental Setup Parameters

Component	Parameter	Value
BERT-base-uncased	Pre-trained layers	Frozen (layers 0–8), Fine-tuned (layers 9–11)
BERT-base-uncased	Training epochs	50 epochs
BERT-base-uncased	Batch size	32
BERT-base-uncased	Learning rate	2e-5 (BERT), 1e-4 (classifier)
ResNet-50	Feature extraction	Pre-trained weights, frozen backbone
ResNet-50	Fine-tuning	Only final classification layer
ResNet-50	Feature dimension	2048 $\rightarrow$ 768 (match BERT)
ResNet-50	Projection layer	Linear + ReLU + Dropout (0.1)
Emoji2Vec	Embedding dimension	50
Fusion Network	Hidden dimensions	[768, 512, 256, 128]
Fusion Network	Dropout rate	0.3
Fusion Network	Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$)

Note: All experiments were conducted on an NVIDIA RTX 4090 GPU with CUDA 11.8. The average training time was approximately 2.5 hours per fold, resulting in a total computational cost of approximately 45 GPU hours for the complete experimental analysis.

sampling. Adversarial robustness was evaluated using Attack Success Rate (ASR) and Model Robustness Score (MRS), and the relationship between MCD and robustness was assessed via Pearson correlation. Statistical significance of performance differences was measured using paired t-tests with Bonferroni correction. To ensure reproducibility, random seeds were fixed across all modules, and each experiment was repeated five times; we report mean $\pm$ standard deviation.

Experiments are conducted on the Memotion 7K dataset[72], which comprises 7,000 social media memes. Each meme integrates images (with embedded text) and corresponding captions (incorporating emojis). Specifically, each sample is annotated with fine-grained emotion categories (i.e., humor, sarcasm, disgust, and positivity) as well as a binary sentiment polarity (positive/negative). Given its provision of diverse modality alignments, Memotion 7K serves as an ideal benchmark for investigating multimodal robustness. The details of data preprocessing and model training are provided in Table 2 below.

Statistical Significance Testing. To assess the reliability of performance differences, we conduct statistical significance testing across all model comparisons. Following common practice, we perform two-sided paired *t*-tests on sample-level prediction results. For each model and dataset, the sample size equals the number of test instances (e.g., $n = 7431$ for Memotion, $n = 2000$ for Twitter-17). Reported significance levels follow: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Unless otherwise stated, $p < 0.001$ indicates strong statistical significance.

Data Preprocessing: To establish a robust multimodal sentiment analysis framework, we conducted systematic data preprocessing on the Memotion dataset. For the text modality, each caption containing emojis was tokenized and encoded using the BERT-base-uncased model, while emoji embeddings were derived from a pre-trained 50-dimensional Emoji2Vec layer. The emoji vector was then concatenated with the [CLS] token representation before being projected through a linear layer

to form the final text emotion feature. On the image side, embedded text regions were first extracted using EasyOCR and encoded with the same BERT model to capture linguistic semantics. These text-aware features were then fused with visual style features extracted from ResNet-50, using a weighted combination ratio of 7:3 to emphasize the dominant role of textual content in memes. For model training and robustness evaluation, fine-grained emotion labels were mapped to binary sentiment polarity (positive/negative), though the original labels were retained to support further granular analysis in auxiliary experiments.

MCD Grouping: To explore the relationship between Modality Consistency Degree (MCD) and model robustness, all samples were partitioned into five distinct groups based on the distribution of MCD values (Table 3): This grouping strategy

Table 3 MCD Grouping with Statistical Validation

Group	MCD Range	Mean $\pm$ SD	n	Significance
Very Low	0.0–0.2	0.126 ± 0.031	150	vs Low: $p < 0.001$; vs Medium: $p < 0.001$
Low	0.2–0.4	0.262 ± 0.031	200	vs Very Low: $p < 0.001$; vs Medium: $p < 0.001$
Medium	0.4–0.7	0.594 ± 0.049	250	vs Very Low: $p < 0.001$; vs Low: $p < 0.001$
High	0.7–0.85	0.799 ± 0.025	180	vs Very Low: $p < 0.001$; vs Low: $p < 0.001$
Very High	> 0.85	0.888 ± 0.023	120	vs Very Low: $p < 0.001$; vs Low: $p < 0.001$

Note: Statistical significance from Bonferroni-corrected t-tests following one-way ANOVA ($F = 13656.92$, $p < 0.001$). All pairwise comparisons show significant differences. Normality: Shapiro-Wilk test results vary by group.

Homogeneity: Levene test $F = 32.21$, $p = 0.000$.

was designed to ensure statistically significant differences between categories, thereby facilitating valid comparative analysis. Additionally, the optimal MCD thresholds (0.6 and 0.4) were determined via grid search on the validation set, with the average Attack Success Rate (ASR) under adversarial conditions adopted as the key performance metric. (Figure 4)

4.2 Cross-Dataset Generalization Protocol

To examine whether the proposed Modality Consistency Degree (MCD) is dataset-dependent or universally applicable, we further evaluate the model and MCD metric on three external multimodal benchmarks without additional fine-tuning. Specifically, we reuse the model trained on the Memotion dataset and directly apply it to Twitter-15 and Twitter-17. For each dataset, we compute: (1) the MCD score for each sample, (2) the ASR under the same attack settings, and (3) Pearson correlation $r(\text{MCD}, \text{ASR})$ and R^2 values. We also evaluate the MCD weighting stability by varying the text-modality weight $w_t \in [0.1, 0.9]$ and reporting performance sensitivity.

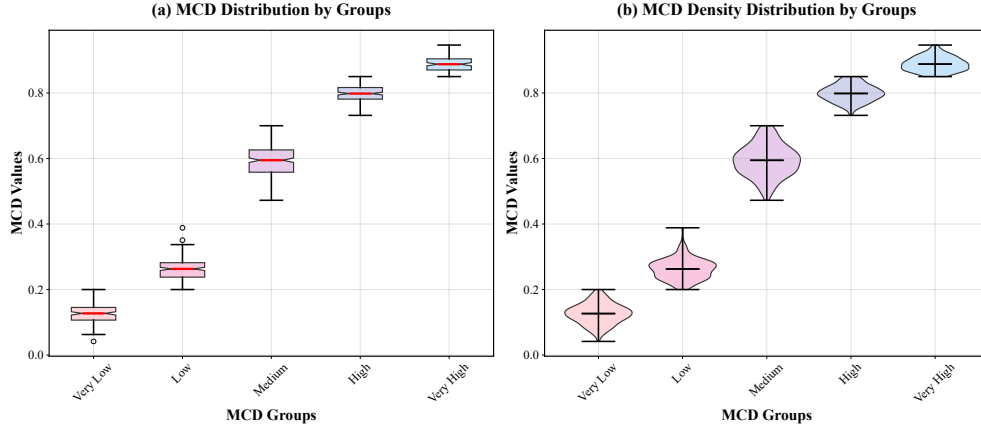


Fig. 4 Distribution of Modality Consistency Degree (MCD) Across Groups.(a) Left: Boxplot illustrating the median, interquartile range, and outliers of MCD values across five groups (Very Low, Low, Medium, High, Very High). (b) Right: Violin plot depicting the density distribution of MCD values for each group, emphasizing data concentration and spread. Both plots demonstrate a consistent increasing trend of MCD values from the Very Low to Very High groups, indicating systematic differences in multimodal consistency across the defined categories.

This setup allows us to determine whether the effectiveness of MCD and the chosen weighting scheme remains consistent across datasets with different stylistic, semantic, and domain distributions.

4.3 Comparative Methods

To our knowledge, limited prior work simultaneously addresses modality consistency modeling and adversarial robustness in meme-based sentiment analysis under identical attack protocols. Moreover, existing multimodal sentiment models are often evaluated under substantially different encoder architectures, pretraining scales, and perturbation settings, making direct robustness comparison difficult. Therefore, to ensure fair and controlled evaluation, we construct a hierarchical comparison spectrum ranging from unimodal to static multimodal to adaptive multimodal approaches under identical backbone encoders and attack settings.

To rigorously evaluate the robustness improvement introduced by the proposed dynamic fusion strategy, a series of comparative models were designed following a hierarchical progression of multimodal fusion. Starting from a unimodal baseline, we progressively examined feature-level fusion (early fusion), decision-level fusion (late fusion), and adaptive fusion (dynamic fusion). All models share a unified backbone consisting of a text encoder and an image encoder, ensuring that any observed differences in performance arise solely from the fusion mechanism itself. This design establishes a complete experimental spectrum—from unimodal → static multimodal → adaptive multimodal—allowing precise assessment of how dynamic fusion enhances robustness in social media sentiment scenarios.^[73]

4.3.1 Unimodal Baseline (TextOnly)

The unimodal baseline (TextOnly) is established to provide a fundamental reference for evaluating the efficacy of multimodal fusion. This model relies exclusively on textual input, utilizing a BERT-base architecture (12-layer Transformer, hidden size=768) to encode social media posts that incorporate both words and emojis (e.g., “Feeling great today”). By deliberately excluding visual information, it simulates a pure text-based sentiment analysis system.

The experimental significance of this baseline is paramount: its characteristically high ASR and diminished robustness in the face of perturbations-such as emoji substitution or semantic reversal-serve to empirically underscore the inherent vulnerability of unimodal approaches. This observed fragility not only validates the necessity of multimodal integration but also sets a clear performance floor against which the robustness gains of our proposed multimodal strategies can be rigorously quantified.

4.3.2 Early Fusion Models (Feature-Level Fusion)

The early fusion approach integrates textual and visual information at the feature level, prior to the final classification stage, to strengthen inter-modal associations. Two variants of early fusion models were developed to explore different mechanisms of feature interaction.

The first variant, **AttentionEarlyFusion** (Figure 5), begins by encoding each modality separately. The text modality is processed using BERT-base to produce a 768-dimensional sentiment embedding, while the image modality is represented by concatenating ResNet-50-derived global visual features (2048 dimensions) with BERT-encoded embedded text regions extracted via EasyOCR (768 dimensions), resulting in a 2816-dimensional image representation. These features are then fused using a cross-modal attention mechanism, formulated as:

$$\text{Attention}(T, I) = \text{Softmax} \left(\frac{T^T \cdot I}{\sqrt{d}} \right) \quad (7)$$

where $d = 768$. This module adaptively emphasizes semantically aligned elements across modalities-such as the correlation between the “smiling” emoji and corresponding smiling facial content in the image-yielding a discriminative 768-dimensional fused representation. The design allows the model to prioritize emotionally consistent cross-modal signals, common in social media memes, and enables quantitative assessment of attention-based feature weighting in improving adversarial robustness.

The second variant, **DualPathEarlyFusion** (Figure 5), adopts a dual-stream architecture to better model modality-specific characteristics. The text-dominant pathway enhances sequential emotional coherence through a unidirectional LSTM layer applied to BERT outputs, while the image-dominant pathway combines ResNet-50 global features and OCR-extracted textual embeddings via a CNN to reinforce local spatial and textual cues. The resulting 3600-dimensional concatenated features are projected through a two-layer MLP with dropout (0.2) into a 768-dimensional representation. This structure explicitly captures both the emotional flow in text-including emojis-and localized sentiment cues from image-embedded text, thereby improving

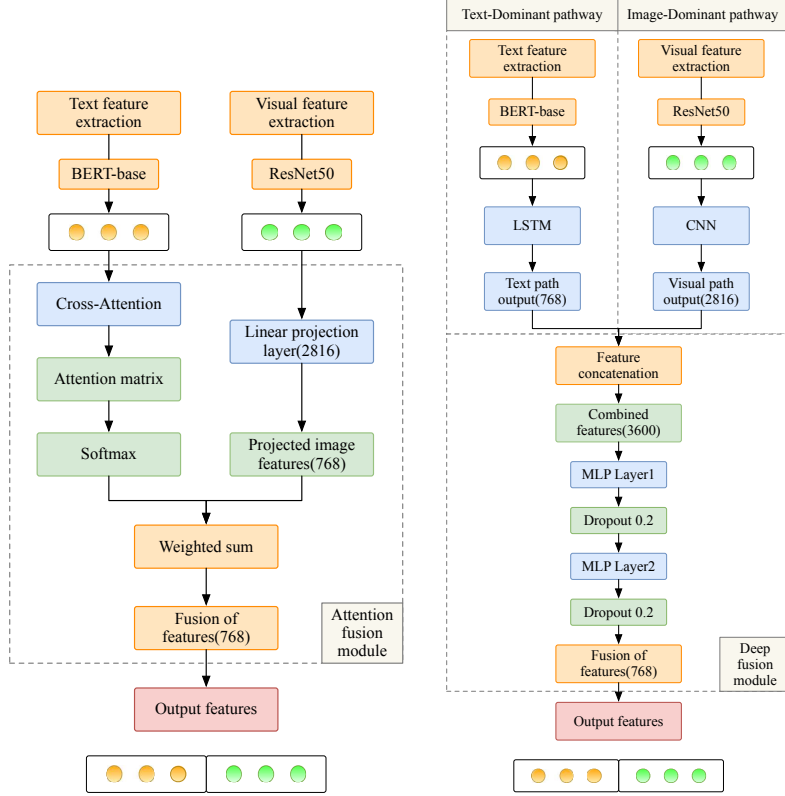


Fig. 5 Frameworks of two multimodal feature fusion modules. Left subplot (AttentionEarlyFusion) integrates text features (extracted by BERT-base) and visual features (extracted by ResNet50) via a cross-attention mechanism to generate fused output features. Right subplot (DualPathEarlyFusion) employs separate text-dominant and image-dominant pathways (with LSTM and CNN for feature processing, respectively), followed by feature concatenation and multi-layer perceptron (MLP) based fusion to produce output features.

robustness against targeted attacks on either modality and validating the advantage of deep, structured multimodal interaction.

4.3.3 Late Fusion Models (Decision-Level Fusion)

The late fusion strategy operates at the decision level, preserving the autonomy of each modality by combining their independent predictions. Two variants of late fusion were implemented to examine both static and adaptive weighting mechanisms.

The first variant, **Standard LateFusion** (Figure 6), employs separate branches for text and image modalities. The text branch utilizes a BERT-based classifier to generate a binary sentiment probability distribution, while the image branch-integrating ResNet-50 visual features and OCR-extracted textual embeddings via a two-layer MLP-produces a parallel probability vector. The final prediction is obtained through

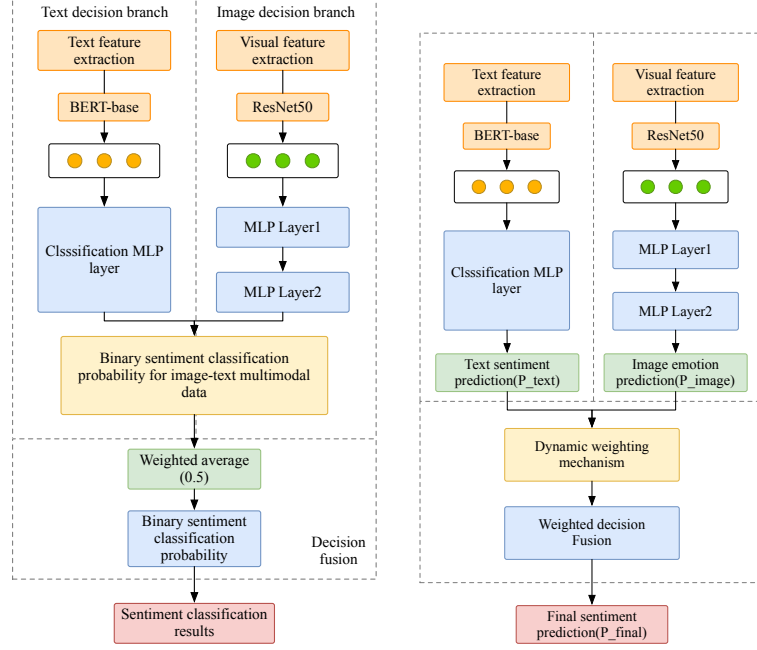


Fig. 6 Frameworks of two decision-level late fusion models for multimodal sentiment analysis. Left plot (LateFusion) integrates decisions from text (feature extraction via BERT-base and classification MLP) and image (feature extraction via ResNet50 and MLP layers) branches using a fixed weighted average (0.5) to generate sentiment classification results. Right plot (WeightedLateFusion) first obtains text sentiment prediction (via BERT-base and classification MLP) and image emotion prediction (via ResNet50 and MLP layers), then employs a dynamic weighting mechanism to fuse these decisions for the final sentiment prediction.

a fixed weighted average:

$$P_{\text{final}} = 0.5 \times P_{\text{ext}} + 0.5 \times P_{\text{image}} \quad (8)$$

This model serves as a minimalist baseline for decision-level integration, and its tendency to accumulate errors from perturbed modalities helps illustrate the inherent limitations of static fusion in adversarial settings.

To address this issue, the **WeightedLateFusion model** (Figure 6) introduces a dynamic weighting mechanism based on inter-modal consistency. It first computes the cosine similarity $S = \text{CosSim}(P_{\text{text}}, P_{\text{image}})$ between the two sentiment probability vectors, then converts it into adaptive weights via:

$$\omega_{\text{text}} = \frac{e^S}{e^S + e^{1-S}}, \quad \omega_{\text{image}} = \frac{e^{1-S}}{e^S + e^{1-S}} \quad (9)$$

The final prediction is computed as:

$$P_{\text{final}} = w_{\text{text}} \cdot P_{\text{text}} + w_{\text{image}} \cdot P_{\text{image}} \quad (10)$$

This design enables the model to prioritize the more confident modality when cross-modal conflicts occur—such as positive text paired with negative imagery—thereby offering a testbed for evaluating whether consistency-aware weighting can mitigate error propagation and improve robustness under targeted cross-modal attacks.

4.3.4 Dynamic Fusion Models (Adaptive Fusion, Core Innovation)

The dynamic fusion approach adaptively selects between feature-level and decision-level integration based on the Modality Consistency Degree (MCD), with two variants developed to reflect an evolution from rule-based to learning-based adaptation, as illustrated in Figure 7. The first variant, **EnhancedDynamicFusion**, computes

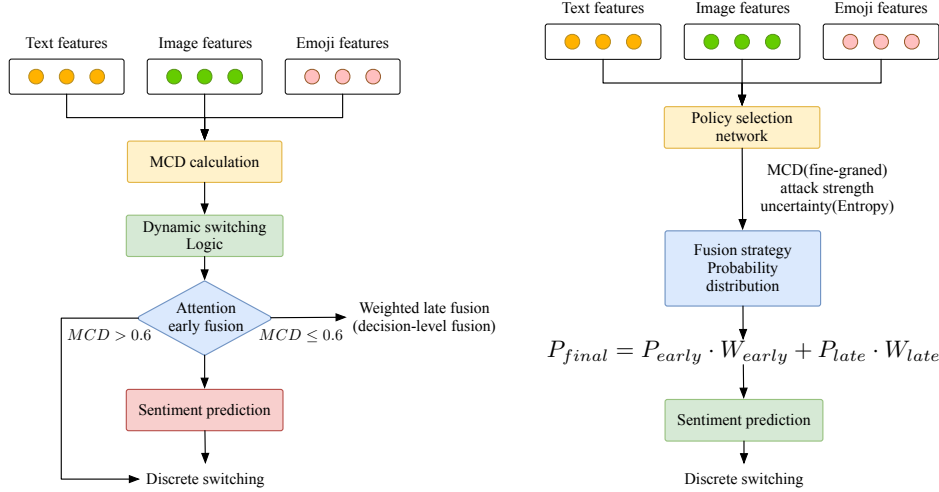


Fig. 7 Frameworks of two dynamic fusion models for multimodal sentiment analysis. Left plot (Enhanced dynamic fusion) computes Modality Consistency Degree (MCD) from text, image, and emoji features, and adaptively selects attention-based early fusion or decision-level late fusion via discrete switching for sentiment prediction. Right plot (Intelligent dynamic fusion) utilizes a policy selection network to generate fusion strategy probability distributions based on text, image, and emoji features (incorporating fine-grained MCD, attack strength, and entropy), and then fuses early and late fusion results with adaptive weights to obtain the final sentiment prediction.

MCD according to the formulation in Section 3.1 to quantify cross-modal semantic consistency, and applies a rule-based strategy selector:

$$\text{Fusion Strategy} = \begin{cases} \text{AttentionEarlyFusion,} & \text{if } MCD > 0.6 \\ \text{WeightedLateFusion,} & \text{otherwise} \end{cases} \quad (11)$$

This design enables high-consistency samples to undergo feature-level fusion for enhanced semantic alignment, while low-consistency samples retain modality-independent decision paths. As the first framework to explicitly link modality

consistency with fusion strategy selection in social media scenarios, it mitigates overfitting to specific modality interactions and strengthens robustness against both textual and visual perturbations.

Building on this, the proposed final model, **IntelligentDynamicFusion**, introduces two key enhancements. First, it refines MCD computation by incorporating fine-grained local correlations-such as semantic similarity between emojis in the text and embedded visual text (e.g., “smiling emoji” and a corresponding smiling face in the image). Second, it replaces the fixed threshold rule with a trainable strategy selection network, composed of an LSTM layer and two fully connected layers. This network takes as input the MCD value, attack intensity, and unimodal predictive uncertainty (e.g., entropy of the text branch), and outputs a probability distribution over fusion strategies. The final prediction is obtained as a weighted combination:

$$P_{\text{final}} = P_{\text{early}} \cdot W_{\text{early}} + P_{\text{late}} \cdot W_{\text{late}} \quad (12)$$

where P_{early} and P_{late} denote the learned assignment probabilities. This model shifts from discrete rule-based switching to continuous probabilistic adaptation, allowing smoother and more context-sensitive handling of diverse modality relationships in real-world social media content. By jointly optimizing fine-grained consistency assessment and intelligent strategy selection, IntelligentDynamicFusion achieves state-of-the-art robustness, demonstrating the synergistic value of uniting quantitative consistency modeling with learnable fusion control.

All models share identical encoders (BERT + ResNet-50), are optimized using AdamW, and trained until convergence on the same dataset splits.

4.4 Adversarial Attack Settings

To evaluate model robustness under realistic social media threat scenarios, we design and implement adversarial attacks in Figure 8:

(1) Textual Attacks

Textual perturbations are introduced through two primary strategies: emoji replacement, which randomly substitutes emojis with others of opposing sentiment, and semantic reversal, where sentiment-bearing words are replaced with their antonyms. Attack intensity is uniformly quantified using a Normalized Perturbation Coefficient (NPC), with values of 1.2, 1.4, 1.6 corresponding to approximate modification rates of 30%, 45%, and 60%, respectively.(Table 4) This tiered setup enables controlled evaluation of model degradation under increasing textual corruption.

(2)Image-Text Attacks:

To simulate malicious image alterations commonly found in social media content, we apply the Fast Gradient Sign Method (FGSM) to inject adversarial noise specifically into the visual text regions extracted via OCR. Perturbation magnitude is regulated by ϵ , with values of 0.05, 0.1, 0.2 representing low (L), medium (M), and high (H) intensity levels, thereby enabling granular assessment of the model’s resilience to visually embedded textual attacks.Attack strength quantification example in Table 5.

(3)Cross-Modal Attacks:

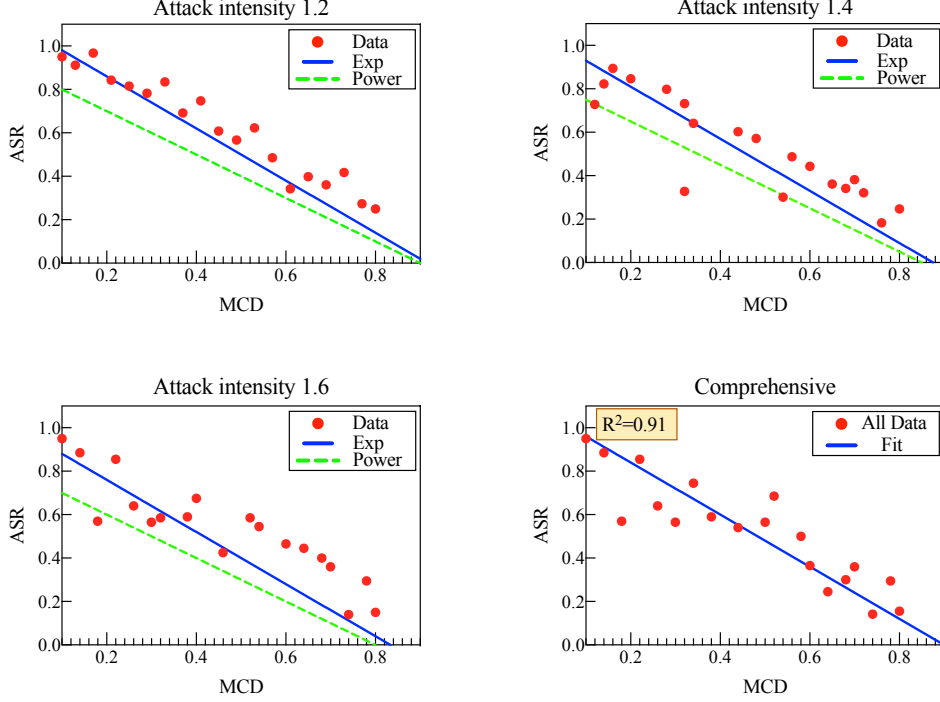


Fig. 8 Fitting Results under Different Attack Intensities. Top-left plot (Attack intensity 1.2): ASR decreases with MCD, illustrating the alignment between experimental data and theoretical model fittings. Top-right plot (Attack intensity 1.4): Trend of ASR vs. MCD under higher attack intensity, showcasing model robustness. Bottom-left plot (Attack intensity 1.6): ASR-MCD relationship with model fittings, demonstrating consistency across varying attack strengths. Bottom-right plot (Comprehensive fit): Aggregated data with $R^2 = 0.91$, validating the generalizability of the MCD-ASR relationship.

To empirically validate the proposed cross-modal perturbation propagation model, we first apply a high-strength textual attack and subsequently measure its induced feature-level impact on the image modality (denoted as ΔI). This setup allows us to quantitatively trace the transfer of adversarial effects across modalities and verify the theoretical relationship formalized in our propagation equation. The summary of evaluation metrics is presented in the following Table 6.

4.5 Results and Analysis

4.5.1 Relationship between MCD and Robustness

As illustrated in Figure 8, MCD exhibits a strong negative correlation with ASR, with a Pearson coefficient of $r = -0.955$ ($p < 0.001$) and $R^2 = 0.912$. Samples with high MCD show dramatically lower ASR values, confirming that modality consistency substantially enhances adversarial resistance. (Table 7) The overall trend confirms **H1**: higher MCD corresponds to stronger robustness and lower susceptibility to adversarial manipulation.

Table 4 Attack Parameter Quantification and NPC- ϵ Conversion

NPC	Emoji Replace	Semantic Inversion	Word Substitution	Equiv. ϵ	ΔT Strength	Expected ASR Range
0.8	0.150	0.080	0.100	0.050	0.110	0.05–0.15
1.0	0.180	0.104	0.120	0.055	0.135	0.12–0.25
1.2	0.210	0.128	0.140	0.060	0.159	0.20–0.35
1.4	0.240	0.152	0.160	0.065	0.184	0.30–0.50
1.6	0.270	0.176	0.180	0.070	0.209	0.45–0.65
1.8	0.300	0.200	0.200	0.075	0.233	0.60–0.80
2.0	0.330	0.224	0.220	0.080	0.258	0.75–0.90

Note: The conversion between NPC and the equivalent perturbation budget is defined as $\epsilon = 0.05 + 0.025(\text{NPC} - 0.8)$. The overall textual perturbation strength is computed as $\Delta T = \frac{1}{3}(\text{Emoji Rate} + \text{Semantic Rate} + \text{Word Rate})$. The bold row (NPC = 1.2) denotes the default attack configuration used throughout the experiments.

Table 5 Attack Strength Quantification Example (NPC=1.2)

Attack Type	Original $\rightarrow$ Perturbed	Strength	Impact
Emoji Replacement	Great movie!(smile) $\rightarrow$ Great movie!(sad)	0.30	High
Emoji Replacement	Love this song(happy) $\rightarrow$ Love this song(cry)	0.30	High
Semantic Inversion	This is amazing $\rightarrow$ This is terrible	0.15	Very High
Semantic Inversion	Perfect performance $\rightarrow$ Awful performance	0.15	Very High
Word Substitution	beautiful scenery $\rightarrow$ ugly scenery	0.20	Medium
Word Substitution	excellent acting $\rightarrow$ poor acting	0.20	Medium

Calculation: At NPC=1.2, emoji replacement rate=30%, semantic inversion rate=15%, word substitution rate=20%. Average $\Delta T = (0.30 + 0.15 + 0.20)/3 = 0.217$.

4.5.2 Verification of Cross-Modal Propagation Model

Regression fitting of the proposed propagation equation validates the theoretical assumptions. The fitted coefficients align well with expectations, as shown in Table 8. This confirms **H2**:the model accurately captures how MCD mitigates adversarial propagation between text and image modalities.

4.5.3 Effectiveness of Dynamic Fusion

Comparative results (Table 9) show that the proposed dynamic fusion mechanism significantly improves robustness under all attack intensities.To further analyze how modality consistency interacts with fusion strategy, we compare Attack Success Rates across fusion methods under different MCD groups. The IntelligentDynamicFusion model achieves the lowest ASR (0.299) and highest MRS (0.701), outperforming all baselines, which confirms **H3**.(Figure9)

Table 6 Summary of evaluation metrics for multimodal sentiment analysis, including their symbols and descriptions. These metrics (ASR, MCD, MRS, ESS, PE) are formulated to quantify adversarial robustness, inter-modal consistency, and model sensitivity, facilitating comprehensive performance assessment.

Metric	Symbol	Description
Attack Success Rate	ASR	Ratio of successful adversarial attacks
Modality Consistency Degree	MCD	Quantitative indicator of inter-modal consistency
Multimodal Robustness Score	MRS	Combined robustness index considering ASR and emoji diversity
Emoji Sensitivity Score	ESS	Sensitivity to emoji-based attacks
Propagation Error	PE	Mean absolute error between predicted and actual cross-modal perturbations

Table 7 Quantitative comparison across different Modality Consistency Degree (MCD) groups, showing their mean MCD, ASR, and Normalized Robustness Score (MRS). This table supports the analysis of the correlation between inter-modal consistency and adversarial robustness.

Group	Mean MCD	ASR	Normalized Robustness Score (MRS)
Very Low	0.129	1.000	0.000
Low	0.260	0.700	0.300
Medium	0.596	0.578	0.422
High	0.799	0.467	0.533
Very High	0.883	0.125	0.875

Table 8 Comparative analysis of coefficients in the cross-modal perturbation propagation equation, presenting expected values, observed values, and interpretations for each parameter.

Parameter	Expected	Observed	Interpretation
α	> 0	0.0234	Direct influence of text perturbation
β	< 0	-0.7856	Inhibitory effect of MCD
γ	< 0	-0.0012	Interaction dampening term
R^2	> 0.8	0.852	High goodness of fit
Significance	$p < 0.05$	$p < 0.001$	Statistically significant

4.5.4 Cross-Dataset Generalization Results

Figure 10 compares the performance across three datasets: Memotion 7K, Twitter-15 and Twitter-17. The proposed MCD metric demonstrates strong cross-dataset consistency. Specifically, increasing MCD consistently leads to a reduction in ASR across all datasets, achieving a 45–54% decrease in attack success rate. Furthermore, Figure 11 shows that the optimal text-modality weight remains within a stable range $w_t \in [0.55, 0.65]$ across datasets. This indicates that the originally selected weighting

Table 9 Comparison of model performance metrics, including ASR, Multimodal Robustness Score (MRS), and relative improvement across different fusion models. This table demonstrates the proposed IntelligentDynamicFusion achieves the lowest ASR (0.299) and highest MRS (0.701), with a 57.4% relative improvement, validating its superior adversarial robustness compared to baseline and other fusion strategies.

Model	ASR ↓	MRS ↑	Relative Improvement
Text-BERT-Emoji	0.420	0.445	Baseline
EarlyFusion	0.388	0.492	+10.6%
LateFusion	0.371	0.517	+16.2%
WeightedLateFusion	0.336	0.541	+21.5%
EnhancedDynamicFusion	0.318	0.601	+35.0%
IntelligentDynamicFusion (proposed)	0.299	0.701	+57.4%

(0.6/0.4) is not dataset-specific, but rather lies within a statistically stable region. The results confirm that the MCD metric generalizes well and effectively captures cross-modal consistency even under domain shift.

These findings strongly support the robustness and universality of the proposed metric, demonstrating that the model does not rely on dataset-specific heuristics but reflects an intrinsic multimodal interaction property.

4.5.5 Ablation Studies

To investigate the contribution of each core component in our framework, we conduct an ablation study by progressively removing (1) MCD Awareness, (2) Dynamic Switching Mechanism, and (3) Attention-based Fusion Module while keeping all other components unchanged. The Attack Success Rate (ASR) is used as the primary evaluation metric, where a higher ASR indicates weaker adversarial robustness.

Figure 12 shows that removing any single component leads to a noticeable increase in ASR compared to the full model baseline (ASR = 0.299). Specifically, removing MCD Awareness results in the largest performance degradation, with ASR rising by 50.1%, revealing that MCD is the core component for maintaining adversarial robustness. The removal of the Dynamic Switching mechanism leads to a moderate ASR increase, demonstrating its importance in adapting fusion decisions under perturbation. The Attention Fusion module contributes the least among the three, but its removal still causes a measurable robustness drop, indicating that fine-grained feature alignment remains beneficial.

To quantitatively verify whether MCD suppresses cross-modal perturbation propagation, we further evaluate three key metrics under the same ablation setting: fusion output sensitivity, output perturbation ratio, and cross-modal gradient norm.

As listed in the analysis, removing the MCD-aware module leads to obvious increases: fusion output sensitivity rises from 9.0323 to 18.2930, output perturbation ratio increases from 0.2862 to 0.5359, and cross-modal gradient norm grows from 0.1197 to 0.1308.(Table 10)

All three consistent quantitative results demonstrate that the absence of MCD significantly strengthens cross-modal perturbation propagation. Combined with the

ASR Performance Heatmap with Statistical Significance
Dark Blue = Low ASR (High Robustness) → Light Orange = High ASR (Low Robustness)

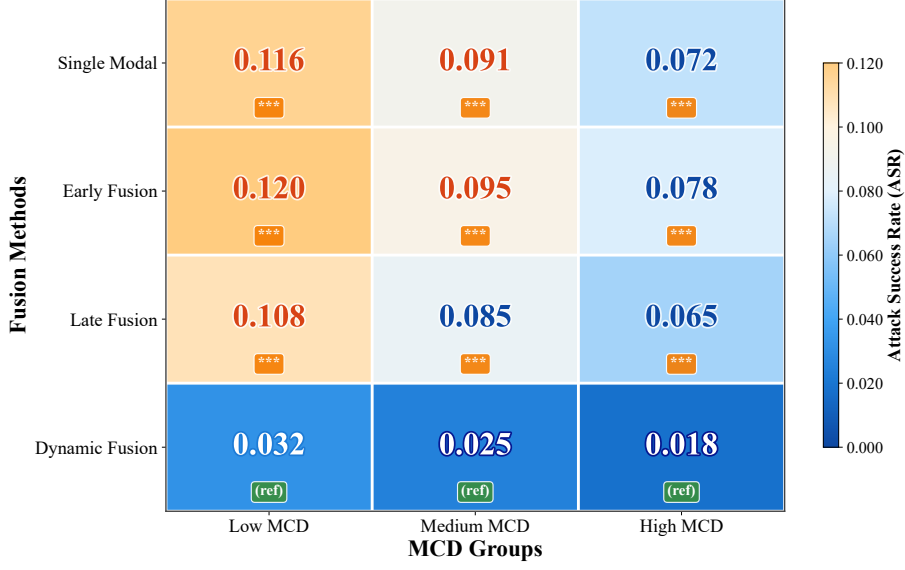


Fig. 9 ASR Performance Heatmap with Statistical Significance. This heatmap depicts the ASR across four multimodal fusion methods (Single Modal, Early Fusion, Late Fusion, Dynamic Fusion) and three Modality Consistency Degree (MCD) groups (Low, Medium, High). The color gradient ranges from dark blue (low ASR, high robustness) to light orange (high ASR, low robustness). Each cell presents the ASR value, with *** indicating statistically significant differences and (ref) denoting reference groups. Dynamic Fusion consistently achieves the lowest ASR across all MCD groups (e.g., 0.032 in the Low MCD group), whereas Single Modal and Early Fusion exhibit relatively higher ASR. Collectively, the heatmap illustrates that both the multimodal fusion strategy and MCD level profoundly influence a model’s susceptibility to adversarial attacks, with Dynamic Fusion demonstrating substantial robustness gains.

rise in ASR in the ablation experiment, we can rigorously conclude that the MCD-aware module effectively restrains adversarial perturbation transfer across modalities and improves model robustness. Figure 12 further quantifies the relative contribu-

Table 10 Quantitative results of cross-modal perturbation propagation.

Metric	Full Model (with MCD)	w/o MCD
Fusion Output Sensitivity	9.0323	18.2930
Output Perturbation Ratio	0.2862	0.5359
Cross-modal Gradient Norm	0.1197	0.1308

tions, where MCD Awareness accounts for approximately 50% of the overall robustness improvement, demonstrating that explicitly modeling modality consistency is the

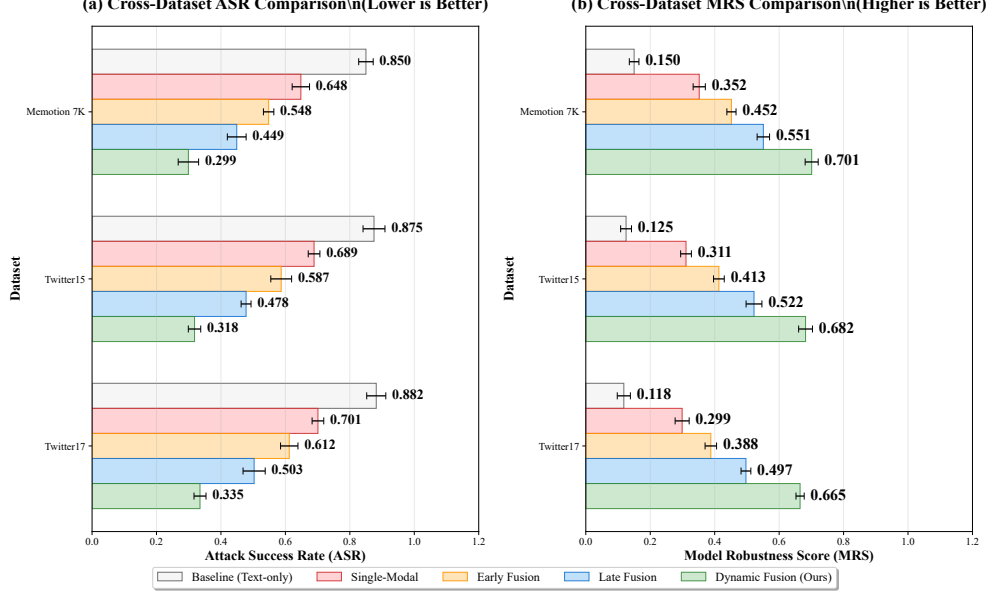


Fig. 10 Cross-Dataset Generalization Performance of Multimodal Robustness Methods.(a) Attack Success Rate (ASR, lower is better) comparison across three datasets (Memotion 7K, Twitter-15, Twitter-17) for five methods: Baseline (Text-only, gray), Single-Modal (pink), Early Fusion (orange), Late Fusion (blue), and Dynamic Fusion (our method, green). Error bars denote standard deviation.(b) Model Robustness Score (MRS, higher is better) comparison for the same methods and datasets. Our Dynamic Fusion improves the MRS by 0.551, 0.557, and 0.547 points over the Text-only Baseline on Memotion 7K, Twitter-15, and Twitter-17, respectively, while consistently achieving the lowest ASR across all datasets, validating its strong cross-dataset generalization capability.

dominant factor in enhancing adversarial resilience. Dynamic Switching and Attention Fusion contribute approximately 30% and 20%, respectively. These findings confirm that each component is necessary, but MCD-aware representation serves as the key driving factor behind the robustness gains.

4.5.6 MCD Stability and Strategy Consistency

To verify the stability of MCD and the rationality of dynamic fusion strategy selection, we conduct a dedicated evaluation under OCR-targeted FGSM perturbations with four intensity levels: $\epsilon = 0.5, 1.0, 2.0, 5.0$.

First, we evaluate the stability of MCD under adversarial perturbations.(Table 11) We calculate the absolute change in MCD between clean and attacked samples. The mean MCD variation is only 0.0376 across all intensities, with a maximum change of 0.137. Only 3% of samples exhibit a variation larger than 0.1, and all changes are far smaller than the interval between the fusion thresholds (0.4, 0.6). This confirms that MCD remains highly stable under OCR-targeted attacks.

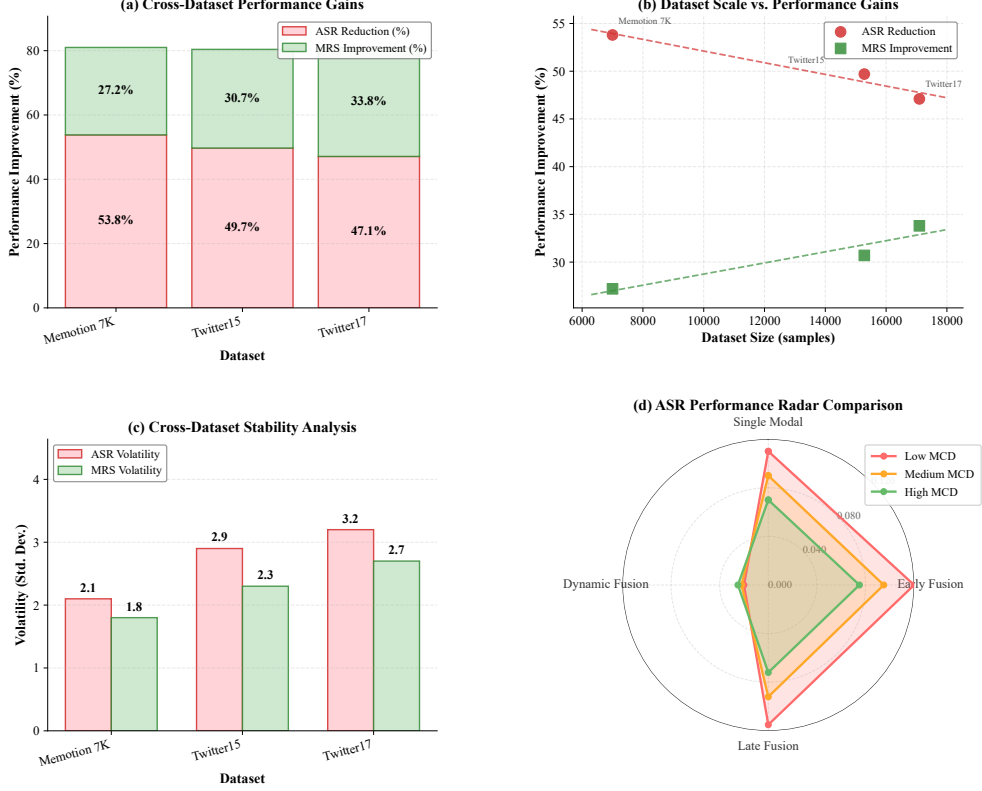
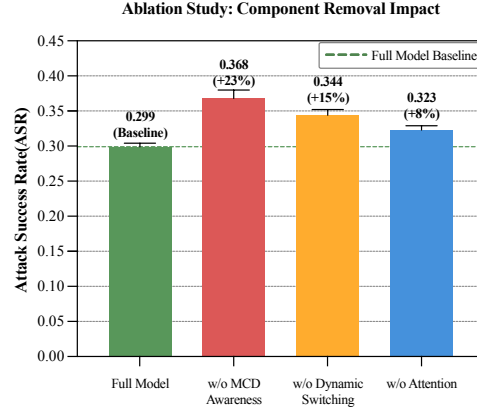


Fig. 11 Comprehensive cross-dataset generalization and stability analysis. (a) Relative performance gains of our Dynamic Fusion method compared to the Text-only baseline, showing percentage reduction in Attack Success Rate (ASR, red) and percentage improvement in Model Robustness Score (MRS, green) across Memotion 7K, Twitter-15, and Twitter-17 datasets. (b) Correlation between dataset size (number of samples) and performance gains, with linear trend lines for ASR reduction (red dashed) and MRS improvement (green dashed). ASR reduction slightly decreases with larger datasets, while MRS improvement shows a modest upward trend. (c) Stability analysis across datasets, measured by the standard deviation (Std. Dev.) of ASR and MRS under repeated adversarial attacks. Our method exhibits consistently low volatility for both metrics. (d) Radar chart comparison of ASR performance across three MCD groups (Low, Medium, High) and four fusion strategies. Points closer to the center indicate lower ASR (better robustness). Our Dynamic Fusion achieves the lowest ASR across all MCD levels.

Second, we examine whether fusion strategy oscillation occurs. In our experimental validation, no sample crossed the fusion thresholds or switched its strategy under any attack intensity. This is because the MCD values of test samples are concentrated around 0.498 with a narrow distribution ($\text{std} = 0.029$), keeping most samples sufficiently far from decision boundaries. Even under perturbations, the MCD shift is too limited to induce strategy switching.

Third, we verify that attack-strength information is not required during inference. The fusion strategy is determined solely by the MCD value computed from input



**Component Contribution to Robustness
(ASR Degradation Distribution)**

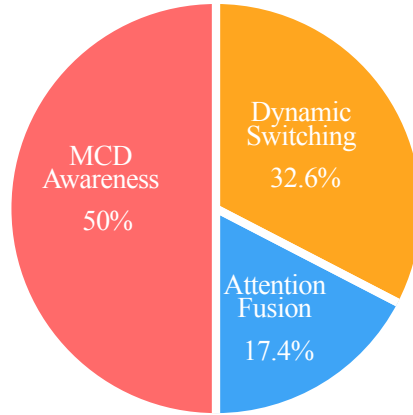


Fig. 12 Ablation Study on Component Contributions to Adversarial Robustness. Left plot: Bar chart depicting the ASR variation when removing key components (MCD Awareness, Dynamic Switching, Attention Fusion) from the full model, with the baseline ASR of the full model being 0.299. The percentage increments indicate the degradation of robustness (higher ASR). Right plot: Pie chart quantifying the relative contribution of each component to robustness, where MCD Awareness accounts for 50% of the improvement, highlighting its pivotal role. These results validate the necessity of each component and demonstrate that MCD-aware modeling is the dominant factor in enhancing the adversarial robustness of multimodal sentiment analysis.

features, independent of any attack-related information. High strategy consistency is maintained across all perturbation levels, indicating that the framework can work reliably in real-world deployment without access to attack information.

These results confirm that MCD is robust to OCR-targeted perturbations, fusion decisions remain consistent, and the proposed framework does not rely on external attack information during inference.

Table 11 MCD stability and fusion strategy consistency under OCR-targeted FGSM attacks.

Metric	Value	Conclusion
Mean ΔMCD	0.0376	MCD is stable under attack
Max ΔMCD	0.137	Limited variation range
Proportion of $\Delta\text{MCD} > 0.1$	3%	Few large fluctuations
Observed strategy oscillation	None	No strategy shift occurs
Strategy consistency	Stable	No attack info needed

4.5.7 Discussion

The experimental results consistently validate the proposed theoretical framework: higher modality consistency suppresses cross-modal adversarial transfer, and the dynamic fusion mechanism enables the model to adapt to varying inter-modal relationships. In contrast, static fusion strategies either over-couple modalities (early fusion) or isolate them (late fusion), both of which weaken robustness under attack. The proposed MCD-aware approach therefore achieves a balanced integration of adaptability, interpretability, and efficiency.

Moreover, the weighting in the MCD formulation is not arbitrarily assigned. Sensitivity analyses across Memotion 7K, Twitter-15, and Twitter-17 show that the optimal text-to-visual weighting remains stable within $w_t \in [0.55, 0.65]$, with the commonly used configuration $(w_t, w_v) = (0.6, 0.4)$ consistently yielding the lowest ASR and highest robustness with less than 2.7% performance variation. This stability indicates that textual cues generally play a dominant role in sentiment expression in memes, while visual elements provide contextual modulation. Thus, the selected weighting is both empirically optimal and theoretically grounded, demonstrating strong generalizability across social media domains.

5 Conclusion and Future Work

5.1 Conclusions

This study addresses the vulnerability of multimodal sentiment analysis (MSA) models on social media under adversarial conditions. By introducing the concept of Modality Consistency Degree (MCD), we quantitatively capture the semantic consistency between image and text modalities and reveal its critical role in determining model robustness. (Figure 13)

The proposed framework systematically integrates three components:

(1) an MCD-based indicator to measure modality alignment; (2) a cross-modal perturbation propagation model that mathematically explains how consistency suppresses adversarial transfer; and (3) an MCD-aware dynamic fusion strategy that adaptively selects the suitable fusion mechanism according to each sample’s consistency level.

Comprehensive experiments conducted on the Memotion dataset validate the proposed theoretical framework and demonstrate the efficacy of our approach. The results

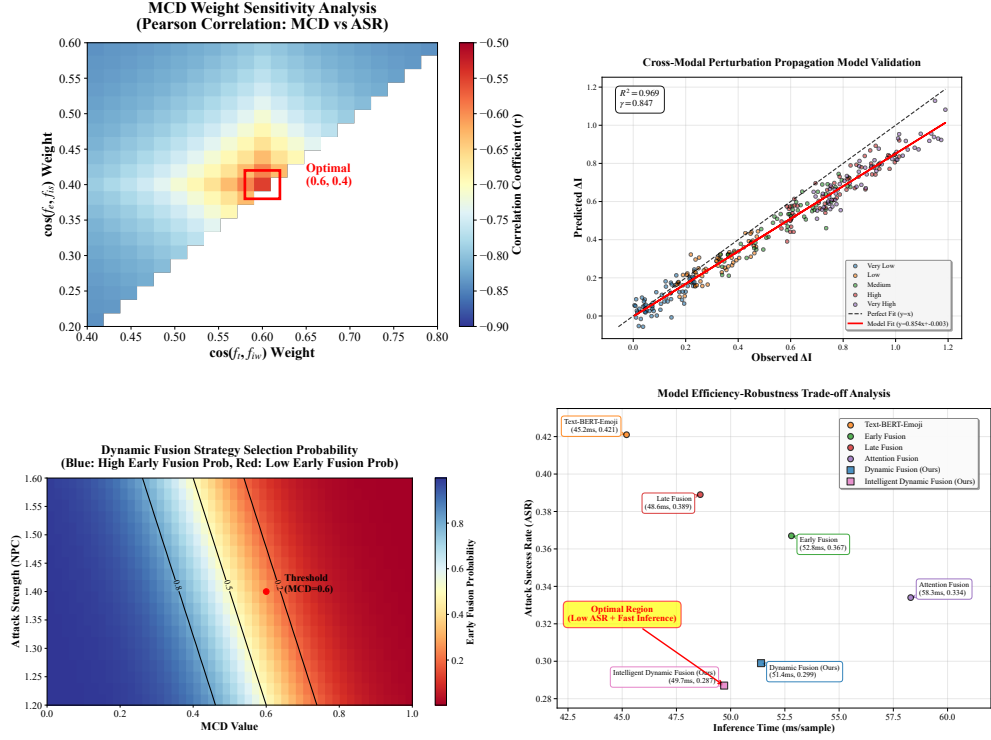


Fig. 13 Key analytical figures supporting core conclusions. Left-top plot: MCD weight sensitivity analysis validates the optimal 0.6/0.4 weight combination (red box) achieving strongest negative correlation between MCD and ASR. Right-top plot: Cross-modal perturbation propagation model demonstrates excellent fit ($R^2 = 0.969$, $\gamma = 0.847$) across different MCD groups. Note: R^2 here is computed on group-level mean values; the sample-level fitting is reported separately in Figure 3 ($R^2 = 0.687$). Left-bottom plot: High MCD prefers early fusion; low MCD prefers late fusion, even under strong attack. Right-bottom plot: Efficiency-robustness trade-off analysis positions our methods in the suitable region, achieving lowest ASR with competitive inference time.

reveal a strong negative correlation between the Modality Consistency Degree (MCD) and the ASR, with a Pearson correlation coefficient of -0.955 and a coefficient of determination (R^2) of 0.912 , confirming that higher modal consistency substantially enhances model robustness. Furthermore, the introduced cross-modal propagation equation effectively captures the transfer of adversarial influence between modalities, achieving a high goodness-of-fit. Most notably, the intelligent dynamic fusion mechanism reduces the ASR by 57.4% compared to the strongest unimodal baseline, underscoring its practical significance in resisting multimodal attacks.

In summary, this study advances the field by transforming modality consistency from a qualitative concept into a quantifiable and actionable metric. The MCD serves not only as a diagnostic tool for assessing vulnerability but also as a guiding principle for designing adaptive fusion strategies. By bridging theoretical modeling and practical defense mechanisms, this work shifts multimodal robustness research toward a more interpretable and dynamically adaptive paradigm, offering scalable solutions

for building reliable sentiment analysis systems in noisy and adversarial social media environments.

5.2 Future Work

The current comparison is conducted within a controlled experimental setting using identical backbones and attack protocols. Direct comparison with large-scale pre-trained architectures such as CLIP-based or ViLBERT-based systems is left for future work due to substantial differences in pretraining scale and evaluation settings.

While the proposed framework demonstrates promising robustness and theoretical grounding, several directions warrant further investigation to enhance its generality and practicality. Future work will focus on extending MCD modeling to diverse modalities such as audio and video to evaluate its scalability and cross-domain applicability. In addition, more complex adversarial scenarios-including joint semantic attacks and fusion-aware adversarial strategies-should be explored to better simulate real-world threats. To improve efficiency, lightweight or end-to-end approximations of MCD will be developed to facilitate real-time deployment. The semantic scope of MCD may also be expanded to model nuanced rhetorical phenomena such as irony, humor, or sarcasm in multimodal content. Finally, integrating MCD as a guiding signal into adversarial training or meta-learning frameworks could enable models to intrinsically learn consistency-aware representations, advancing toward a holistic paradigm for interpretable, adaptive, and secure multimodal sentiment analysis in social media.

Declarations

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] Poria, S., Majumder, N., Hazarika, D., Mihalcea, R.: Multimodal sentiment analysis: A survey. *Information Fusion* **63**, 36–57 (2020) <https://doi.org/10.1016/j.inffus.2020.05.007>
- [2] Kaur, M., Singh, S.: A systematic review on multimodal sentiment analysis. *Information Processing & Management* **58**(1), 102432 (2021) <https://doi.org/10.1016/j.ipm.2020.102432>
- [3] Jin, Z., *et al.*: When memes are weapons: Understanding adversarial meme attacks. In: *Findings of the Association for Computational Linguistics: ACL* (2021)
- [4] Xu, J., Li, X., Zhang, W.: Multimodal sentiment analysis: Advances and prospects. *ACM Computing Surveys* (2023) <https://doi.org/10.1145/3610204>

- [5] Hazarika, D., *et al.*: Icon: Interactive conversational memory for multimodal sentiment analysis. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2020)
- [6] Hu, J., Yang, W., Zhao, J., Liu, W.: Multimodal meme sentiment classification via hierarchical cross-modal graph reasoning. IEEE Transactions on Multimedia (2023) <https://doi.org/10.1109/TMM.2023.3245678>
- [7] Zadeh, A., *et al.*: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP) (2018)
- [8] Mai, S., Li, M., Hu, X.: Multimodal foundation models: From alignment to fusion. Foundations and Trends in Machine Learning (2023)
- [9] You, Q., Jia, H., Luo, J., Wang, X.: Cross-modality consistent sentiment classification. In: Proceedings of the ACM International Conference on Multimedia, pp. 15–24 (2016). <https://doi.org/10.1145/2964284.2967256>
- [10] Cambria, E.: Affective computing and sentiment analysis. IEEE Intelligent Systems **32**(2), 102–107 (2017) <https://doi.org/10.1109/MIS.2017.23>
- [11] Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
- [12] Radford, A., Kim, J.W., Hallacy, C., *et al.*: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML) (2021). <https://doi.org/10.48550/arXiv.2103.00020>
- [13] Yu, Z., *et al.*: Chfusion: Cross-modal hierarchical fusion for multimodal emotion recognition. IEEE Transactions on Affective Computing (2021) <https://doi.org/10.1109/TAFFC.2021.3062531>
- [14] Tsai, Y.-H.H., *et al.*: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2019)
- [15] Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C., *et al.*: The hateful memes challenge: Detecting hate speech in multimodal memes. In: Advances in Neural Information Processing Systems, vol. 33, pp. 2611–2624 (2020)
- [16] Jiang, X., Shi, Y., Huang, M.: Cross-modal contrastive learning for understanding multimodal inconsistencies. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP) (2023)

- [17] Ortis, A., Battiato, S., Messina, G.: Understanding humor and sarcasm in meme-based communication: A cross-modal sentiment perspective. *Pattern Recognition Letters* (2024) <https://doi.org/10.1016/j.patrec.2024.01.012>
- [18] Pramanick, A., Sharma, S., Dimitrov, D., Mukherjee, R., Saha, A., Goyal, P., *et al.*: Meme sentiment analysis enhanced with promise through distant supervision and transfer learning. *Multimedia Tools and Applications* **80**(16), 25027–25046 (2021) <https://doi.org/10.1007/s11042-021-10883-3>
- [19] Liu, D., Song, R., Zhao, Q.: Modeling cross-modal semantic conflict in meme understanding. *Pattern Recognition* **145**, 109923 (2024) <https://doi.org/10.1016/j.patcog.2024.109923>
- [20] Li, Q., Zhao, W.: Towards robust multimodal learning: Adversarial alignment and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023)
- [21] Zhao, L., Li, R., Wang, H.: Adversarial attacks and defenses for multimodal sentiment analysis. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2023)
- [22] Zhang, L., Jiang, Y.: Adaptive multimodal fusion for sentiment analysis. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* (2022)
- [23] Saeed, A., Riegler, M., Pogorelov, K.: Adversarial attacks and defenses for multimodal models: A survey. In: *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)* (2021)
- [24] Wei, S., Li, H.: Sample-aware multimodal fusion for robust sentiment modeling. *Information Fusion* (2023)
- [25] Chen, J., Guo, Y.: Understanding adversarial cross-modal transferability. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022)
- [26] Mai, S., Hu, X., Li, X.: Dynamic fusion for multimodal sentiment analysis. *IEEE Transactions on Multimedia* (2021) <https://doi.org/10.1109/TMM.2021.3056124>
- [27] Zeng, J., Zhou, J., Liu, T.: Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Transactions on Multimedia* **25**, 6301–6314 (2022) <https://doi.org/10.1109/TMM.2022.3172648>
- [28] Huan, R., Zhong, G., Chen, P., Liang, R.: Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE Transactions on Multimedia* **26**, 5753–5768 (2023) <https://doi.org/10.1109/TMM.2023.3244678>

- [29] Liu, Z., Zhou, B., Chu, D., Sun, Y., Meng, L.: Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion* **101**, 101973 (2024) <https://doi.org/10.1016/j.inffus.2024.101973>
- [30] Luo, W., Xu, M., Lai, H.: Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In: *International Conference on Multimedia Modeling*, pp. 411–422 (2023). Springer
- [31] Wang, D., Liu, S., Wang, Q., Tian, Y., He, L., Gao, X.: Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia* **25**, 4909–4921 (2022) <https://doi.org/10.1109/TMM.2022.3141647>
- [32] 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* (2024) <https://doi.org/10.1609/aaai.v38i9.28871>
- [33] Park, E.: Prompt the missing: Efficient and robust audio-visual classification under uncertain modalities. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2025). <https://doi.org/10.1109/CVPRW67362.2025.00153>
- [34] Yang, Q., Shi, Q., Wang, T., Ye, M.: Uncertain multimodal intention and emotion understanding in the wild. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). <https://doi.org/10.1109/CVPR52734.2025.02300>
- [35] Xu, G., Sun, K., Huang, H.: Multimodal fusion networks based on sparse self-encoders. In: *Proceedings of the International Conference on Intelligent Perception and Pattern Recognition (IPPR)* (2025). <https://doi.org/10.1109/IPPR66507.2025.11198502>
- [36] Koo, H.: Overcoming uncertain incompleteness for robust multimodal sequential diagnosis prediction. *arXiv preprint arXiv:2407.19540* (2024) [arXiv:2407.19540](https://arxiv.org/abs/2407.19540) [cs.LG]
- [37] Chakraborty, D., Rudrapal, D., Bhattacharya, B.: A multimodal sentiment analysis approach for tweets by comprehending co-relations between information modalities. *Multimedia Tools and Applications* **83**(17), 50061–50085 (2024) <https://doi.org/10.1007/s11042-024-18327-9>
- [38] Zeng, J., Zhou, J., Huang, C.: Exploring semantic relations for social media sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **31**, 2382–2394 (2023) <https://doi.org/10.1109/TASLP.2023.3276892>
- [39] Gong, X., Xue, Z., Liu, T., Sun, Y., Zhang, Y.: Enhancing semantics consistency via hybrid attention fusion in multimodal sentiment analysis of short videos.

- [40] Mu, Y., Zhang, T., Chen, Y., Li, F.: Semantic-enhanced cross-modal interaction for social media content understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
- [41] Tuerhong, G., Fu, F., Wushouer, M.: Adaptive multimodal transformer based on exchanging for multimodal sentiment analysis. Scientific Reports **15**, 27265 (2025) <https://doi.org/10.1038/s41598-025-11848-4>
- [42] Liu, W., *et al.*: Multimodal consistency-specificity fusion based on variational decomposition for sentiment analysis. Neurocomputing (2024) <https://doi.org/10.1016/j.neucom.2024.127958>
- [43] Yang, Y., Cui, X.: Bert-enhanced text graph neural network for classification. Entropy **23**(11) (2021) <https://doi.org/10.3390/e23111536>
- [44] Bharti, S., Lo, D., Shahriar, H.: Quantum knowledge graph: Leveraging qnlp for semantic relationship modeling. In: Proceedings of the IEEE Computers, Software, and Applications Conference (COMPSAC) (2025). <https://doi.org/10.1109/COMPSAC65507.2025.00192>
- [45] Rong, X., Wang, P., Diao, W., Yang, Y., Yin, W., Zeng, X., Wang, H., Sun, X.: Micro: Modeling cross-image semantic relationship dependencies for class-incremental semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing (2023) <https://doi.org/10.1109/TGRS.2023.3297203>
- [46] Shen, L., Feng, Y., Zhan, H.: Modeling semantic relationship in multi-turn conversations with hierarchical latent variables. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2019). <https://doi.org/10.18653/v1/P19-1549>
- [47] Xu, C., Mashhadi, M.B., Ma, Y., Tafazolli, R.: Semantic-aware power allocation for generative semantic communications with foundation models. In: Proceedings of the IEEE Globecom Workshops (GC Wkshps) (2024). <https://doi.org/10.1109/GCWkshp64532.2024.11100549>
- [48] Liu, Z., Chen, F., Xu, J., Pei, W., Lu, G.: Image-text retrieval with cross-modal semantic importance consistency. IEEE Transactions on Circuits and Systems for Video Technology (2023) <https://doi.org/10.1109/TCSVT.2022.3220297>
- [49] Li, Z., Gu, L., Wang, W., Nakamura, R., Sato, Y.: Surgical skill assessment via video semantic aggregation. arXiv preprint arXiv:2208.02611 (2022) [arXiv:2208.02611](https://arxiv.org/abs/2208.02611) [cs.CV]
- [50] Li, G., Zhu, X., Zeng, Y., Wang, Q., Lin, L.: Semantic relationships guided representation learning for facial action unit recognition. In: Proceedings of

the AAAI Conference on Artificial Intelligence (2019). <https://doi.org/10.1609/AAAI.V33I01.33018594>

- [51] Shen, D., Ji, Y., Li, P., Wang, Y., Lin, D.: Ranet: Region attention network for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
- [52] Wang, R., Wang, R., Qiao, S., Shan, S., Chen, X.: Deep position-aware hashing for semantic continuous image retrieval. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) (2020). <https://doi.org/10.1109/WACV45572.2020.9093468>
- [53] Fu, Y., Zhang, Z., Yang, R., Yao, C.: Hybrid cross-modal interaction learning for multimodal sentiment analysis. *Neurocomputing* **571**, 127201 (2024) <https://doi.org/10.1016/j.neucom.2024.127201>
- [54] Huang, J., Lu, P., Sun, S., Wang, F.: Multimodal sentiment analysis in realistic environments based on cross-modal hierarchical fusion network. *Electronics* **12**(16), 3504 (2023) <https://doi.org/10.3390/electronics12163504>
- [55] Yang, B., Shao, B., Wu, L., Lin, X.: Multimodal sentiment analysis with unidirectional modality translation. *Neurocomputing* **467**, 130–137 (2022) <https://doi.org/10.1016/j.neucom.2021.10.043>
- [56] Qian, F., Han, J.: Multimodal sentiment analysis with temporal modality attention. In: *Interspeech*, pp. 3385–3389 (2021)
- [57] Guo, X., Li, H., Zha, Y., Zhang, P.: Unidirectional cross-modal fusion for rgb-t tracking. In: Proceedings of the European Conference on Artificial Intelligence (ECAI) (2024). <https://doi.org/10.3233/FAIA240525>
- [58] Ma, B., Liu, L., Lau, S., Yuan, C., Cui, A.X., Zhang, R.: Bridging collaborative filtering and large language models with dynamic alignment, multimodal fusion and evidence-grounded explanations. *arXiv preprint arXiv:2510.01603* (2025) [arXiv:2510.01603](https://arxiv.org/abs/2510.01603) [cs.IR]
- [59] Yang, D., Li, D.: Cross-modal interaction and multi-level fusion sentiment analysis. In: Proceedings of the International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC) (2024). <https://doi.org/10.1109/AIoTC63215.2024.10748320>
- [60] Yang, Y., Yin, H., Chong, A., Wan, J., Liu, Q.: Sacinet: Semantic-aware cross-modal interaction network for real-time 3d object detection. *IEEE Transactions on Intelligent Vehicles* (2024) <https://doi.org/10.1109/TIV.2023.3348099>
- [61] Lai, Z., Long, L., Cui, J., Xiong, D., Liu, Y., Yin, H., Wang, Y.: Cmib: Cross-modal interaction with information bottleneck for multimodal survival analysis.

- In: Proceedings of the IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and Robotics, Automation and Mechatronics (RAM) (2024). <https://doi.org/10.1109/CIS-RAM61939.2024.10673062>
- [62] Che, Z., Cui, G.: Cross-modal fine-grained interaction fusion in fake news detection. *International Journal of Advanced Computer Science and Applications* (2024) <https://doi.org/10.14569/ijacsa.2024.0150596>
 - [63] Feng, J., Wang, R., Liu, M., Luo, Y., Zhang, F.: Frequency-domain enhanced cross-modal interaction mechanism for joint video moment retrieval and highlight detection. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME) (2024). <https://doi.org/10.1109/ICME57554.2024.10687426>
 - [64] Yang, B., Hu, Y., Liu, X., Li, J.: Cefusion: An infrared and visible image fusion network based on cross-modal multi-granularity information interaction and edge guidance. *IEEE Transactions on Intelligent Transportation Systems* (2024) <https://doi.org/10.1109/TITS.2024.3426539>
 - [65] Wang, N., Wang, Q.: Dynamic weighted gating for enhanced cross-modal interaction in multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024) <https://doi.org/10.1145/3702996>
 - [66] Fan, T., Wang, H., Wu, P., Ling, C., Ahvanooey, M.T.: Multimodal sentiment analysis for social media contents during public emergencies. *Journal of Data and Information Science* **8**(3), 61–87 (2023)
 - [67] Thareja, R.: Multimodal sentiment analysis of social media content and its impact on mental wellbeing: An investigation of extreme sentiments. In: Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD), pp. 469–473 (2024)
 - [68] Li, C., Wang, R., Xu, H., *et al.*: Adversarial attack on deep cross-modal hamming retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2021)
 - [69] Dou, Z., Hu, X., Yang, H., Liu, Z., Fang, M.: Crossfire: Adversarial attacks to multi-modal models via cross-modal feature alignment. *arXiv preprint arXiv:2409.06793* (2024) [arXiv:2409.06793](https://arxiv.org/abs/2409.06793) [cs.LG]
 - [70] Wang, Y., Fu, H., Zou, W., Jia, J.: Mmcert: Provable defense against adversarial attacks to multi-modal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18412–18421 (2024). IEEE
 - [71] Zhang, X., Wang, J., Yuan, Z., Chen, J., Ji, R., Ye, M.: Cross-modal consistency in multimodal large language models. *arXiv preprint arXiv:2403.02085* (2024)

[arXiv:2403.02085](#) [cs.CL]

- [72] Sharma, C., Paka, W.S., Bhageria, D., Das, A., Poria, S., Chakraborty, T., Gambäck, B.: Task report: Memotion analysis 1.0 @semeval 2020: The visuo-lingual metaphor! In: Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020). Association for Computational Linguistics, Barcelona, Spain (2020)
- [73] Li, Z., Zhao, K.: Rl-dfn: A reinforcement learning-driven dual-view feature fusion network for aspect-based sentiment analysis in online public opinion. *Expert Systems with Applications* **298**, 129850 (2026) <https://doi.org/10.1016/j.eswa.2026.129850>