

Article

Reward-Guided Dynamic Fusion and Modality Decoupling for Enhanced Multimodal Sentiment Analysis

He Zhang ¹, Zichen Gao ², Qi Yan ², Yu Gu ^{3,*}, Shuang Wang ³, Linsong Liu ³ and Dequan An ³¹ School of Journalism and Communication, Northwest University, Xi'an 710127, China; zhanghe@nwu.edu.cn² Yantai Research Institute, Harbin Engineering University, Yantai 264000, China; zuzhen-chen@hrbeu.edu.cn (Z.G.); yanqi11@hrbeu.edu.cn (Q.Y.)³ Department of Artificial Intelligence, Xidian University, Xi'an 710126, China; shwang@mail.xidian.edu.cn (S.W.); 23171214603@stu.xidian.edu.cn (L.L.); dqan@stu.xidian.edu.cn (D.A.)

* Correspondence: guyu@xidian.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) integrates multiple modalities to better understand human emotions. However, existing methods often neglect heterogeneity among modal features, causing redundancy and inconsistencies. Additionally, the dynamic interplay between modalities is frequently ignored during fusion, limiting performance. To address these issues, we propose Reward-Guided Dynamic Fusion and Modality Decoupling (RDFD). RDFD includes two key components: (1) a feature decoupling module that separates modality-specific and modality-shared features, reducing redundancy and conflicts; (2) a Reward-Guided Dynamic Fusion module that adaptively selects guiding modalities to enhance modality-specific representations and enable flexible fusion. Experiments on the CMU-MOSI and CMU-MOSEI datasets show that RDFD achieves state-of-the-art performance, demonstrating its effectiveness in advancing Multimodal Sentiment Analysis.

Keywords: multimodal sentiment analysis; modality decoupling; dynamic fusion; reward-guided learning

1. Introduction

With the rapid growth of social media, people are increasingly shifting from single-modal interactions, such as speech, to multimodal interactions. As a result, Multimodal Sentiment Analysis (MSA) has emerged as a crucial subfield in Artificial Intelligence [1–6]. MSA consists of two key components: multimodal feature extraction and multimodal feature fusion. Early research in multimodal feature extraction primarily focused on mapping different modal data into a shared feature space for representation learning [7–10]. For instance, tensor fusion network (TFN) utilized a tensor outer product operation to map multimodal features into a 3D feature space [11], while Low-rank Multimodal Fusion (LMF) applied low-rank decomposition to combine highly correlated feature vectors [12]. However, these approaches forced heterogeneous modal features (e.g., video, audio, and text) into a common representation space, failing to eliminate modality-specific heterogeneity [13–17]. To address these limitations, modality decoupling was introduced, aiming to separate different variation factors across multiple modalities into independent subspaces [18]. This technique helps distinguish modality-specific and modality-shared features, thereby reducing redundancy and conflicts, ultimately enhancing the accuracy of sentiment analysis [19]. Several methods have been proposed to achieve modality decoupling. For example, Modality-Invariant and Specific Representations for Multimodal



Academic Editor: Stefanos Kollias

Received: 21 March 2026

Revised: 14 April 2026

Accepted: 16 April 2026

Published: 24 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Sentiment Analysis (MISA) projected each modality into shared and specific feature spaces, reducing inter-modality differences while improving representation diversity [20]. Similarly, Feature-Disentangled Multimodal Emotion Recognition (FDMER) leveraged metric learning and adversarial learning to construct modality-shared and modality-specific subspaces, significantly enhancing multimodal fusion [21].

In multimodal fusion, two primary approaches dominate: ternary symmetry methods and text-dominated methods. Ternary symmetry methods, such as CubeMLP, balance the contributions of all modalities by equally processing text, vision, and audio across sequence, modality, and channel axes [22].

In contrast, text-dominated methods prioritize text as the core modality, using it to guide the integration of visual and auditory features. Examples include ALMT, which employs text queries in cross-modal transformers [23]; TETFN, which applies a text-oriented attention mechanism [24]; and MMCL, which optimizes audio and visual features alongside text [25]. Additionally, DLF enhances language representations through a text-guided cross-attention mechanism [18]. While effective, both approaches have limitations: ternary symmetry methods fail to account for modality-specific features, potentially suppressing unique modality-specific differences during fusion, whereas text-dominated methods statically set text as the dominant modality, which may not generalize well across different scenarios. In some cases, over-reliance on text can misdirect the model's attention, especially when visual or audio features are more informative.

To illustrate this point, we will use a set of data from the MOSI dataset shown in Figure 1 as an example: a speaker says “Now the title of the movie basically says it all,” which is semantically neutral (T: Neutral), yet their facial expression such as frowns conveys clear negativity (V: Negative), and the audio waveform lacks strong emotional cues (A: Neutral). The ground truth label is Negative, clearly indicating that visual modality dominates in this case. If a text-dominated method were applied, it would likely predict Neutral, while a ternary symmetric fusion might dilute the visual signal by averaging with neutral text and audio. This highlights the critical need for dynamic, sample-adaptive fusion that can identify and amplify the most informative modality per instance.



Figure 1. An illustrative example where visual modality dominates sentiment prediction. Although the text is neutral (“Now the title of the movie basically says it all”), the speaker’s facial expression is negative, and the ground truth label is Negative. Audio provides no strong cue. This demonstrates the limitation of static fusion strategies and motivates our dynamic, reward-guided approach.

A survey of commonly used MSA datasets indicates that the dominant modality varies across different contexts [18]. In real-world applications, text, vision, and audio may each serve as the primary modality depending on the situation. To address these challenges,

we propose Reward-Guided Dynamic Fusion, which dynamically determines the most informative modality based on the characteristics of each sample, ensuring adaptive and robust Multimodal Sentiment Analysis. To address these challenges, we propose Reward-Guided Dynamic Fusion, which dynamically determines the most informative modality based on the characteristics of each sample, ensuring adaptive and robust Multimodal Sentiment Analysis.

Unlike existing works, we propose a Reward-Guided Dynamic Fusion and Modality Decoupling (RDFD) method, which first decouples shared and specific features through modality separation. It then employs a reward mechanism to dynamically adjust modality weights, establishing a dominant modality to guide the fusion of modality-specific features. Our main contributions can be summarized as follows:

- **Proposed Method:** We propose the RDFD multimodal representation learning method to enhance MSA tasks. The RDFD framework provides a structured approach comprising feature extraction, modality decoupling, feature enhancement, and prediction.
- **Reward-Guided Dynamic Fusion:** The Reward-Guided Dynamic Fusion mechanism consists of three key components: modal refactor, reward computation, and dynamic fusion. Initially, the modal refactor component establishes initial modality weights. These weights are then dynamically updated using a diverse set of reward mechanisms, such as loss and accuracy. Finally, the dominant modality is determined using the updated weights, enabling adaptive dynamic fusion for improved performance.

2. Methods

2.1. Proposed Method

In RDFD, the subscript $m \in \{a, v, t\}$ denotes the following modalities: audio (a), visual (v), and text (t). Each modality is represented by a tensor $I_m \in \mathbb{R}^{N_m \times d_m}$, where N_m is the sequence length and d_m is the embedding dimension. Figure 2 shows the RDFD method, which decomposes multimodal features into shared and specific spaces to reduce redundancy and conflicts. To achieve decoupling, three loss functions with discriminators maximize separation between shared and specific features. Additionally, we design a Reward-Guided Dynamic Fusion module to enhance modality-specific features. Four reward mechanisms are introduced for optimal reward computation and adaptive fusion. The following sections detail each module of our method. The overall algorithm can be found in Appendix A.

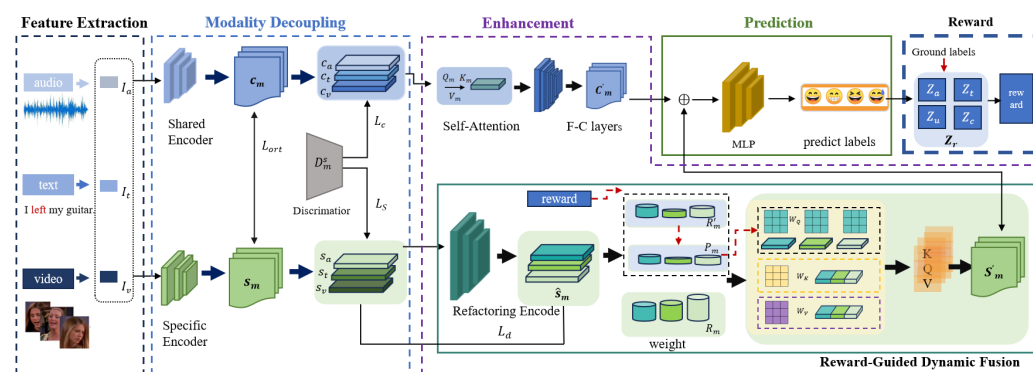


Figure 2. The overall RDFD method includes Feature Extraction, Modality Decoupling, Feature Enhancement, and Prediction.

2.2. Feature Decoupling Module

To analyze heterogeneous modal information and reduce redundancy, we design a feature decoupling module that separates modality-shared features c_m and modality-

specific features s_m . The shared encoder E_m^c extracts c_m , and the specific encoder E_m^s extracts s_m , ensuring a clear distinction for effective representation learning.

$$\begin{aligned} c_m &= E_m^c(I_m) \\ s_m &= E_m^s(I_m) \end{aligned} \quad (1)$$

This process requires separating both shared and specific features within and across modalities. Following [26], we enforce a soft orthogonality constraint between the representations. To achieve this, we introduce the L_{ort} loss function, ensuring effective feature separation and minimizing redundancy between modality-shared and modality-specific representations.

$$L_{ort} = \sum_{m \in \{t,a,v\}} \|c_m^\top s_m\|_F^2 + \sum_{(m_1, m_2) \in t,a,v} \|s_{m_1}^\top s_{m_2}\|_F^2 \quad (2)$$

Unlike the feature decoupling framework of MISA, on this basis, we enhance differentiation by introducing an emotion-centered discriminator. Separate loss functions (L_c/L_s) optimize shared and specific features to retain key classification information. Reverse optimization prevents information loss during the decoupling process while ensuring the completeness of task-related features, thereby implicitly enhancing the uniqueness and interpretability of the features.

The notation $\|\cdot\|$ denotes the square of the Frobenius norm, used to measure matrix similarity. While the orthogonal constraint aids feature separation, it may cause information loss during learning. To address this, we introduce a discriminator D and apply discriminant loss to both modality-shared and modality-specific features, further enhancing their separation in the decoupling process. Accordingly, we define

$$L_c = -\frac{1}{N_b} \sum_{m \in \{t,a,v\}} y \log(D(c_m; \theta_D)) \quad (3)$$

$$L_s = -\frac{1}{N_b} \sum_{m \in \{t,a,v\}} y \log(D(s_m; \theta_D)) \quad (4)$$

Here, θ_D represents the discriminator parameters, and N_b denotes the batch size, with y as the true label for c_m and s_m . Through the feature decoupling process and the three loss functions, the separation of modality-shared and modality-specific features is effectively achieved.

2.3. Feature Enhancement Module

The Feature Enhancement Module consists of two components: modality-shared feature enhancement and modality-specific feature enhancement. Among these, modality-specific feature enhancement is partially implemented through Reward-Guided Dynamic Fusion.

The modality-shared features c_m and modality-specific features s_m are obtained through feature decoupling. The specific implementation is as follows.

Modality-shared features represent the common feature information extracted from heterogeneous modalities. To refine these features, we first apply a self-attention mechanism, allowing the model to focus on the most informative aspects of the input features. This mechanism enables the model to effectively identify key features in affective expression while ignoring redundancy and conflicts. The process of applying the self-attention mechanism is as follows:

$$A = \text{softmax}\left(\frac{Q_m K_m}{\sqrt{d}}\right) V_m \quad (5)$$

where Q_m , K_m , and V_m represent the query, key, and value vectors for the same modality $m \in \{a, v, t\}$, and $\sqrt{d}$ denotes the dimension of the shared feature c_m . Next, a fully connected (FC) layer refines the extracted features, producing the modal-shared enhancement feature c'_m

$$c'_m = F(Ac_m; \theta_1) \quad (6)$$

In the above, c'_m represents the shared-enhancement feature, and θ_1 represents the parameters in the FC layers.

2.4. Reward-Guided Dynamic Fusion

The modality-specific feature enhancement network consists of three components: modal refactor, reward computation, and dynamic fusion. The D_m^s refactoring encoder processes modality-specific features to generate refactored features $\hat{s}_m$.

In the modal refactor stage, $\hat{s}_m$ is used to establish the initial weights of the modalities. By decoding the fusion features, the exponential error of the sum D_m is computed and transformed into emotional gravity, recorded as the initial weight R_m . The encoder's parameter reconstruction process is described as follows: In the modal refactor stage, the refactored modality-specific features $\hat{s}_m$ are utilized to establish the initial weights of the three modalities. First, calculate the exponential error D_m of each modal reconstruction. If the characteristics of a certain modal can be reconstructed well, it indicates that the information of this modal is well retained and utilized. Conversely, it also suggests that the information of this modal may be lost or unimportant. Therefore, the reconstruction exponential error D_m of $\hat{s}_m$ and s_m is calculated, and the relative importance of each mode is obtained through normalization, which is used as the initial weight. The process of reconstructing parameters in the encoder can be described as follows:

$$\hat{s}_m = D_m^s(s_m; \theta_2) \quad (7)$$

$$D_m = \exp(-k|\hat{s}_m - s_m|^2) \quad (8)$$

$$R_m = \frac{D_m}{D_a + D_t + D_v} \quad (9)$$

In this process, k represents the function's slope, which scales the sentiment ratio. Simultaneously, to minimize the difference between s_m and $\hat{s}_m$, we introduce a loss function L_d for evaluation:

$$L_d = \sum_{m \in \{t, a, v\}} \|s_m - \hat{s}_m\|_F^2 \quad (10)$$

The reward design comprises four reward mechanisms: accuracy, task-classification loss, uncertainty loss, and consistency loss, denoted as Z_a , Z_t , Z_u , and Z_c , respectively.

$$Z_a = \frac{n_c}{n} \quad (11)$$

$$Z_t = -\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (12)$$

Let n denote the total number of samples, and n_c represent the number of correctly predicted samples. Additionally, let y denote the true labels of the samples and $\hat{y}$ represent the predicted labels.

In dominant modality selection, certain modalities may contain task-irrelevant information or be affected by noise interference, necessitating quantitative measurement. The dominant modality is defined as the modality that contributes the most to the prediction re-

sult in the current task. Entropy can be used to measure the uncertainty of the classification result of a certain sample by the model.

Z_u is a concept derived from entropy, used to measure the uncertainty of a probability distribution. In this context, modal uncertainty can also be utilized to quantify a model's confidence in its predictions through entropy. Here, p_{ij} represents the probability that the n th sample belongs to the i th class, and ϵ is a small constant introduced to prevent taking the logarithm of zero.

$$Z_u = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C p_{ij} \log(p_{ij} + \epsilon) \quad (13)$$

Z_c is used to measure the similarity between the modal feature and its consistency feature using cosine similarity. A higher Z_c value indicates a higher likelihood of a modality being the dominant modality. Therefore, Z_c is incorporated into the reward mechanism to enhance modality selection.

$$Z_c = \frac{1}{M} \sum_{m \in M} \frac{1}{N} \sum_{i=1}^N \cos(s_m, c_m) \quad (14)$$

In the reward calculation process, the values of Z_t and Z_u play a crucial role in assessing model reliability. Higher Z_t and Z_u indicate greater model uncertainty, making them negatively correlated with the reward. Conversely, a larger Z_a reflects better model performance, while a higher Z_c signifies improved collaboration between modalities, both of which are positively correlated with the reward.

Since Z_a, Z_t, Z_u , and Z_c operate at different scales, ranges, and semantic levels, a naive linear summation is mathematically unsound. We therefore apply batch-wise z-score normalization and introduce learnable coefficients $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ to automatically balance their contributions:

$$\tilde{Z}_i = \frac{Z_i - \mu_{Z_i}}{\sigma_{Z_i} + \epsilon}, \quad i \in \{a, t, u, c\} \quad (15)$$

$$\tilde{Z}_r = \alpha_1 \tilde{Z}_a - \alpha_2 \tilde{Z}_t - \alpha_3 \tilde{Z}_u + \alpha_4 \tilde{Z}_c \quad (16)$$

where μ_{Z_i} and σ_{Z_i} are batch-wise mean and standard deviation, $\epsilon = 10^{-8}$ ensures numerical stability, and α_i are trainable parameters initialized to 1.0. The negative signs for Z_t and Z_u reflect that higher loss and uncertainty should reduce the reward.

Then, the reward is used to update the initial weight as follows: The normalized composite reward $\tilde{Z}_r$ modulates the initial weights in a single, fully differentiable step:

$$R'_m = R_m \cdot (1 + \lambda \cdot \tilde{Z}_r) \quad (17)$$

where λ is a learnable scaling factor initialized to 1.0. At this point, R'_m is using $\tilde{Z}_r$ to update the weights of each modality.

Although the composite reward signal $\tilde{Z}_r$ is not explicitly formulated as an independent loss term, it implicitly guides the optimization of the refactoring encoder parameters θ_2 through dynamic gradient modulation during backpropagation. Specifically, the updated modality weight R'_m incorporates $\tilde{Z}_r$ as a multiplicative scaling factor, which directly influences the magnitude of gradients flowing back to the reconstruction encoder. When a modality demonstrates high task relevance, the resulting gradient magnitude is proportionally amplified, compelling the encoder to prioritize and refine the reconstruction of discriminative features that contribute most to sentiment prediction. Conversely, for modalities exhibiting low task relevance or minimal initial importance, the gradient signal is naturally attenuated, which implicitly discourages the model from allocating representational capacity to uninformative or noisy components. Through this continu-

ous feedback mechanism across training iterations, the refactoring encoder progressively becomes task-aware: reconstruction fidelity R_m naturally converges with actual task importance, as measured by $\tilde{Z}_r$, ensuring that the model learns to preserve only the most semantically valuable modality-specific information without requiring additional explicit supervisory signals.

To ensure adaptive and robust feature integration, the dominant modality serves as a semantic anchor that actively queries complementary cues from auxiliary modalities. We first derive soft dominance probabilities p_m from the reward-updated weights R'_m ,

$$p_m = \frac{\exp(R'_m / \tau)}{\sum_{m' \in \{a, t, v\}} \exp(R'_{m'} / \tau)} \quad (18)$$

where $\tau = 0.2$ is a temperature parameter during training. This continuous formulation avoids gradient discontinuities while preserving the selection logic.

The query, key, and value matrices are constructed as follows:

$$Q = \sum_{m \in \{a, t, v\}} p_m \cdot W_Q s_m \quad (19)$$

$$K = W_K [s_i; s_j; s_k] \quad (20)$$

$$V = W_V [s_i; s_j; s_k] \quad (21)$$

where W_Q, W_K, W_V are shared linear projection matrices, and $[s_i; s_j; s_k]$ denotes the concatenation of all three modality-specific features. By weighting the query construction with p_m , the dominant modality effectively dictates the attention focus, actively retrieving complementary information from other modalities. Meanwhile, K and V preserve the full cross-modal context, preventing premature information loss.

To further emphasize task-relevant signals, the value matrix V is element-wise scaled by the updated weights R' before attention aggregation. This mechanism ensures that modalities with high reconstruction fidelity and reward alignment contribute proportionally stronger features to the fusion process. The enhanced modality-specific feature s'_m is computed as

$$s'_m = \text{LayerNorm} \left(\text{Softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) (V \odot R') + Q \right) \quad (22)$$

where d_k is the key dimension, $R' = [R'_a, R'_t, R'_v]$ is broadcast to match V 's dimensions, and $\odot$ denotes element-wise multiplication. This design strategically positions the dominant modality throughout the fusion pipeline: The dominant modality drives the query Q , ensuring the attention mechanism focuses on retrieving the most relevant cross-modal cues. Scaling V by R' dynamically amplifies reliable modalities and suppresses noisy ones, aligning the fusion process with both feature stability and task reward. The residual connection ($+Q$) guarantees that the original discriminative information from the dominant modality is retained, preventing over-smoothing during cross-modal interaction.

2.5. Prediction

As shown in the Figure 2, we concatenate the specific enhancement features with the shared enhancement features:

$$\hat{y} = M(s'_m \oplus c'_m) \quad (23)$$

The loss function for the classifier is

$$L_{\text{cls}} = \frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2 \quad (24)$$

The final total output loss of MSA is

$$L_{\text{RDFD}} = \lambda_1 L_{\text{ort}} + \lambda_2 L_c + \lambda_3 L_s + \lambda_4 L_d + \lambda_5 L_{\text{cls}} \quad (25)$$

L_{ort} is the orthogonality loss that enforces feature separation between modality-shared and modality-specific representations, minimizing redundancy and interference across modalities. L_c is the discriminant loss for modality-shared features, optimizing the shared encoder to retain task-relevant common information across different modalities. L_s is the discriminant loss for modality-specific features, ensuring that each modality preserves its unique characteristics while maintaining classification capability. L_d is the reconstruction loss that measures the fidelity of modality-specific feature reconstruction, serving as the basis for initial modality weight estimation. L_{cls} is the classification loss (mean squared error) that directly optimizes the sentiment prediction accuracy. $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are hyperparameters that balance the contributions of different loss components, with values detailed in the experimental section.

3. Experiment

3.1. Set up Experiment

We evaluated the RDFD method on CMU-MOSI [27] and CMU-MOSEI [28], with dataset statistics shown in Table 1.

MOSI: The MOSI dataset is a collection of 2199 opinion video clips, each of which is annotated with emotions within the range of $[-3, 3]$, representing the range of emotions from highly negative to highly positive. Its input data includes a combination of language, visual and acoustic modalities. The audio and visual features were extracted at frequencies of 12.5 Hz and 15 Hz, respectively. This dataset is divided into 1284 training samples, 229 validation samples and 686 test samples.

MOSEI: MOSEI is the next generation of MOSI. Its dataset is larger and consists of 22,856 video clips of movie reviews from YouTube. Each segment is also annotated with emotions within the range of $[-3, 3]$ for feature extraction in audio at 20 Hertz and visual patterns at 15 Hertz. The dataset is divided into 16,326 training samples, 1871 validation samples and 4659 test samples.

Evaluation Metrics: Following prior studies [29–31], we evaluated performance using multiple metrics: seven-level accuracy (Acc-7), five-level accuracy (Acc-5), binary accuracy (Acc-2), F1-score, correlation with human annotations (Corr), and Mean Absolute Error (MAE).

Table 1. Statistics of the CMU-MOSI and CMU-MOSEI datasets.

Dataset	Train	Val	Test	Total
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16,326	1871	4659	22,856

We implemented this method using PyTorch 2.2.1. The experiments were conducted on an NVIDIA GeForce RTX 3090. The detailed information about the hyperparameters is shown in Table 2.

Table 2. Hyperparameters for Different Datasets. ✓ indicates that the technique was used.

Hyperparameter	MOSI	MOSEI
Batch Size	64	64
Learning Rate	1×10^{-4}	1×10^{-4}
λ Weight	0.1, 0.2, 0.1, 0.2, 0.2, 0.2	0.1, 0.2, 0.1, 0.2, 0.2, 0.2
α Weight	0.3, 0.2, 0.2, 0.3	0.3, 0.2, 0.2, 0.3
Optimizer	AdamW	AdamW
Epochs	200	200
Early Stop	✓	✓

3.2. Comparative Models

Feature decoupling methods focus on separating shared and modality-specific features, ensuring a more distinct representation across modalities. Text-dominant methods prioritize language as the primary modality, guiding fusion through text-based attention mechanisms. Ternary symmetric methods, on the other hand, equally process all modalities, often utilizing operations such as tensor outer products or low-rank decomposition to achieve balanced fusion.

To comprehensively evaluate RDFD's effectiveness in multimodal fusion tasks, we compare it against various representative models, including

- (1) Feature decoupling methods: MISA [20], FEMER [21];
- (2) Text-dominant fusion methods: TETFN [24], MMCL [25], DLF [18];
- (3) Ternary symmetric methods: TFN [11], LMF [12], CubeMLP [22].

3.3. Experimental Results

Tables 3 and 4 show baseline results and performance comparisons for CMU-MOSI and CMU-MOSEI. The results demonstrate that RDFD is highly competitive, achieving significant improvements in key metrics such as accuracy, F1-score, and correlation.

The Reward-Guided Dynamic Fusion mechanism in RDFD allows the model to adaptively adjust modality weights based on validation loss fluctuations, leading to an overall performance enhancement.

Table 3. Performance comparison on CMU-MOSI. The numbers in [] represent the sources of the data, - represents that the result from the original paper is not provided, and * represents reproduced results from public code with hyper-parameters provided in the original paper. **Bold** values indicate the best results.

Method	Acc-7	Acc-5	Acc-2	F1	Corr	MAE
TFN [18]	34.90	39.90	80.80	80.87	0.698	0.901
LMF [18]	33.20	38.13	82.50	82.40	0.695	0.917
MISA [18]	41.73	47.08	83.54	83.58	0.778	0.777
FDMER [21]	42.10	-	84.20	83.90	0.845	0.732
CubeMLP [32]	43.44	-	82.32	81.77	0.772	0.755
TETFN [24]	-	-	84.05	83.83	0.800	0.717
MMCL [25]	46.50	-	84.00	83.80	0.797	0.705
DLF *	47.23	53.14	83.69	83.65	0.786	0.730
RDFD (ours)	46.66	53.21	84.76	84.70	0.802	0.708

On the CMU-MOSI dataset, RDFD outperforms existing methods in ACC-5, ACC-2, F1, and correlation. Compared to the latest DLF algorithm, RDFD achieves a 1.05% increase in F1 and a 1.07% improvement in ACC-2. However, its performance in ACC-7 and MAE is slightly lower than DLF and CubeMLP, respectively.

This pattern reveals a subtle trade-off in the RDFD design: Dynamic fusion excels in capturing coarse-grained emotional polarities (binary or five-category) but faces challenges in the fine-grained intensity distinctions. In MOSI, seven-category tasks require distinguishing between adjacent levels, such as “slightly negative” and “moderately negative”, where the differences often lie only in subtle variations in acoustic intonation or lexical choice. Our analysis indicates that the reward mechanism of RDFD tends to amplify the strongest modal signals but may inadvertently suppress weaker but discriminative cues from other modalities, which are crucial for boundary classification. For example, a neutral statement with a slightly tense tone can only be correctly classified as “weakly negative” when the audio pitch contour and text negation markers are considered simultaneously; however, if the reward assigns a higher weight to the audio, the text negation will be ignored, leading to a misjudgment as an excited “weakly positive” emotion, and the model will thus overlook this subtle difference. Therefore, ACC-7 is slower than DLF, which adopts a more conservative and balanced cross-modal interaction approach.

On the CMU-MOSEI dataset, due to its higher complexity, the model requires more training epochs to reach peak performance, resulting in a smaller performance gain compared to CMU-MOSI. Nevertheless, RDFD still improves in ACC-5, ACC-2, F1, and correlation. Compared to DLF, it achieves a 1.36% increase in ACC-2.

Table 4. Performance comparison on CMU-MOSEI. The numbers in [] represent the sources of the data, - represents that the result from the original paper is not provided, and * represents reproduced results from public code with hyper-parameters provided in the original paper. **Bold** values indicate the best results.

Method	Acc-7	Acc-5	Acc-2	F1	Corr	MAE
TFN [18]	50.20	53.10	82.50	82.10	0.700	0.593
LMF [18]	48.00	52.90	82.00	82.10	0.677	0.623
MISA [18]	52.05	53.63	84.67	84.66	0.752	0.558
FDMER [21]	53.80	-	83.90	83.80	0.736	0.568
CubeMLP [32]	53.35	-	83.76	83.82	0.760	0.531
TETFN [24]	-	-	84.25	84.18	0.748	0.551
MMCL [25]	53.60	-	84.80	83.80	0.765	0.537
DLF *	53.67	55.70	84.42	85.27	0.744	0.532
RDFD (ours)	51.75	53.70	85.78	85.45	0.778	0.538

On the CMU-MOSEI dataset, Figure 3 reveals a pronounced long-tail distribution in the sentiment labels, where over 40% of samples are concentrated in the positive and slightly positive categories, while extreme or negative categories remain relatively under-represented. One plausible explanation for the observed limitations in ACC-7 and MAE is that this label imbalance may potentially constrain the model’s capacity to learn robust decision boundaries for minority classes. During optimization, gradients from frequent categories could dominate the training process, which might reduce the model’s sensitivity to subtle intensity variations in the tail region. Alternatively, the dynamic modality selection mechanism, while highly effective for capturing coarse-grained emotional polarity, might inadvertently prioritize strongly dominant signals at the expense of weaker but discriminative cross-modal cues that are crucial for fine-grained boundary classification. Despite these potential bottlenecks, RDFD still achieves consistent and meaningful improvements over existing static and text-dominant fusion baselines across most evaluation metrics (e.g., Acc-2, F1, and Corr on both datasets). This demonstrates that the reward-guided adaptive fusion strategy provides a substantial advancement in multimodal representation learning, even if fine-grained intensity discrimination under highly imbalanced conditions remains a challenging direction for future research.

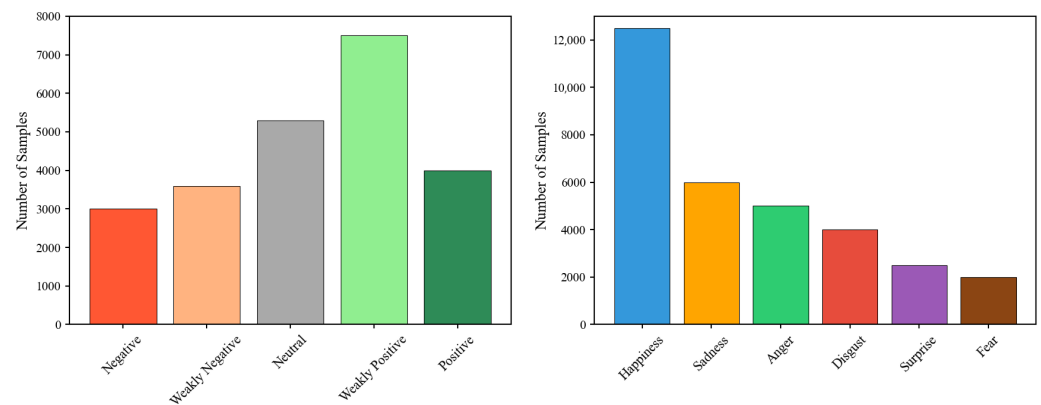


Figure 3. Label distribution of the MOSEI dataset.

Consequently, the improvement in RDFD on the ACC-7 metric, which requires precise discrimination among seven fine-grained emotion levels, appears limited. The relative scarcity of samples from boundary categories may impede robust classification and could increase the difficulty of distinguishing adjacent emotional intensities. Similarly, the Mean Absolute Error (MAE) might exhibit suboptimal performance, as outliers from under-represented classes can have a disproportionate impact on regression-based evaluations. Although our dynamic modality weighting effectively adapts to dominant signals, it may face inherent constraints when operating within highly imbalanced data distributions.

To further investigate the influence of emotion granularity on MSA, we visualized the confusion matrix and corresponding per-class accuracy for the MOSI dataset. As shown in Figure 4, we observe that while most emotional categories achieve accuracy rates above 35%, the extreme categories “HN” (Highly Negative) and “HP” (Highly Positive) exhibit substantially poorer performance, with HN reaching only 13% and HP at 35%, compared to 45–60% for medium-intensity categories. This performance gap may stem from multiple interacting factors rather than a single cause.

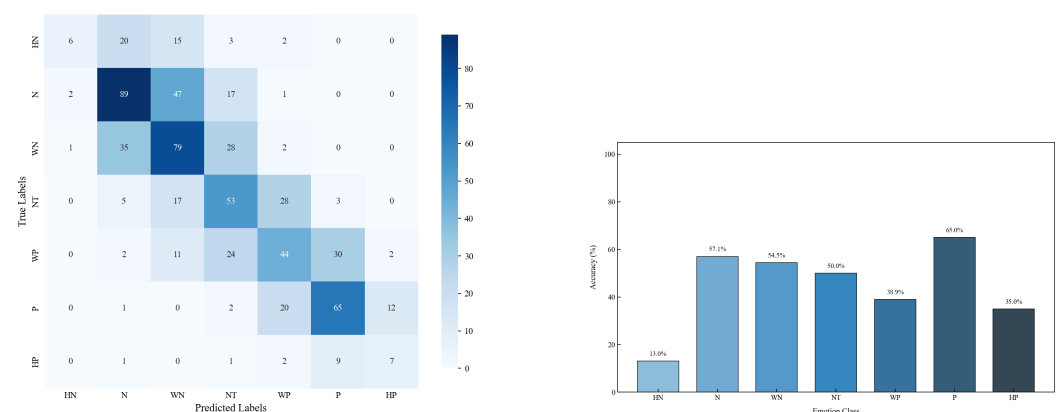


Figure 4. Left: Confusion matrix on MOSI. Right: Corresponding accuracy for each sentiment. HN: Highly Negative; N: Negative; WN: Weakly Negative; NT: Neutral; WP: Weak Positive; P: Positive; HP: Highly Positive.

First, the limited representation of HN and HP samples in the training set likely restricts the model from learning comprehensive feature manifolds for extreme emotional states, resulting in poor generalization. Second, manual inspection of misclassified HN samples revealed a recurring pattern: sarcastic or ironic statements characterized by neutral facial expressions combined with positive or neutral vocal tones are frequently misclassified as neutral or weakly positive. This suggests that extreme emotions are not always expressed through strong unimodal signals but may manifest as inter-modal discrepancies (e.g.,

positive lexical content paired with negative facial expressions). The current decoupling module, while effective at separating shared and specific features, may struggle to explicitly model such contradiction patterns, causing the dominant modality selection to be interfered with by conflicting shared features. Third, the reward-guided weighting mechanism, designed to amplify the most reliable modality, could potentially suppress subtle cross-modal cues necessary for delicate boundary distinctions when faced with ambiguous or conflicting signals.

Despite these observations, RDFD demonstrates that Reward-Guided Dynamic Fusion provides a principled approach to adaptive modality weighting. To address the identified limitations, future research could focus on several promising directions: In the future, strategies such as class-balanced learning (e.g., focal loss, cost-sensitive weighting, or stratified sampling) could be adopted to mitigate the impact of the long-tail distribution; at the same time, specific cross-modal conflict modeling modules could be designed for implicit emotions like sarcasm to detect and utilize the inconsistencies between modalities, especially for complex emotional states such as sarcasm or irony. These approaches may further narrow the gap between dynamic fusion mechanism and robust fine-grained emotion discrimination.

3.4. Ablation Studies

We conducted a series of ablation studies to evaluate the effectiveness of our proposed network, as summarized in Table 5.

The module decoupling MD separates shared and modality-specific features, effectively reducing inter-modal interference and redundancy. Its removal leads to a 1.46% decrease in ACC-7, a 1.26% drop in ACC-2, and a 1.1% reduction in F1-score, primarily due to increased interference and a weakened binary classification ability.

When the MD module is removed, the model is forced to directly fuse all modal features into the same space, resulting in an amplification of semantic conflicts between different modalities. Without separating shared and unique features, visual signals may be diluted by the “neutral” components in text or audio, leading to incorrect predictions of “NT” or “WN”, which are precisely the issues that the design of the MD module aims to alleviate. Moreover, the decline in binary classification performance indicates that even in coarse-grained tasks, the redundancy between modalities weakens the decision boundary. Because the shared features are not effectively extracted, the model has difficulty forming a unified emotional polarity representation.

Table 5. Ablation Study of RDFD on CMU-MOSI and CMU-MOSEI Datasets. w/ Text and w/ Ternary respectively represent the text-dominant fusion and the ternary symmetric fusion.

Method	CMU-MOSI				CMU-MOSEI			
	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE
RDFD (Ours)	46.66	84.76	84.70	0.708	51.75	85.78	85.45	0.538
w/o MD	45.20	83.50	83.60	0.720	49.98	84.43	83.78	0.544
w/o Reward	45.31	83.64	83.75	0.724	50.42	84.73	84.57	0.551
w/ Text	46.39	83.90	84.44	0.713	51.20	85.60	84.79	0.549
w/ Ternary	45.30	84.10	83.80	0.733	50.43	85.00	84.50	0.546
w/o Z_a	45.89	83.98	84.12	0.716	51.03	85.12	84.91	0.545
w/o Z_t	46.12	84.31	84.38	0.713	51.38	85.41	85.18	0.541
w/o Z_u	45.73	84.19	84.26	0.719	51.11	85.29	85.03	0.543
w/o Z_c	46.28	84.52	84.49	0.720	51.49	85.56	85.27	0.540

When the text-based fusion method is replaced with the RDFD-based approach, this model achieves quite good performance in binary classification tasks (for example, the ACC-2 on MOSI is 83.90%, and on MOSEI it is 85.60%). This performance is expected

because textual cues usually can clearly convey the emotional polarity. However, this strategy leads to a slight decrease in ACC-7, F1 score, and MAE. This decline is due to the fact that the model is unable to utilize non-textual emotional signals, such as sound intensity or facial micro-expressions, which are crucial for distinguishing adjacent emotional levels.

The triadic symmetric fusion method treats all three modalities equally, but its performance on various indicators is relatively average. Specifically, ACC-7 decreased by 1.36% on MOSI and by 1.32% on MOSEI compared to RDFD, and the F1 score also decreased (by 0.90% on MOSI and 0.95% on MOSEI). This pattern indicates that forcing all modalities to contribute equally will weaken the influence of the most informative signal in each sample. The triadic fusion may average these signals and produce a too-conservative prediction, so its MAE performance on the MOSEI dataset is slightly higher than that of text dominance. This strategy sacrifices accuracy in capturing subtle emotional nuances. In conclusion, the ablation study confirms that RDFD is not merely a fusion mechanism but a decision-making engine that enables the model to adapt its behavior based on input characteristics, a key advantage over rigid static strategies.

To better understand the effectiveness of our method, we visualize the distribution of the multimodal representations. As shown in Figure 5, we will continue to compare using MD decoupling and the original data distribution. MD can separate different emotions, thereby reducing redundant and conflicting information during the multimodal interaction.

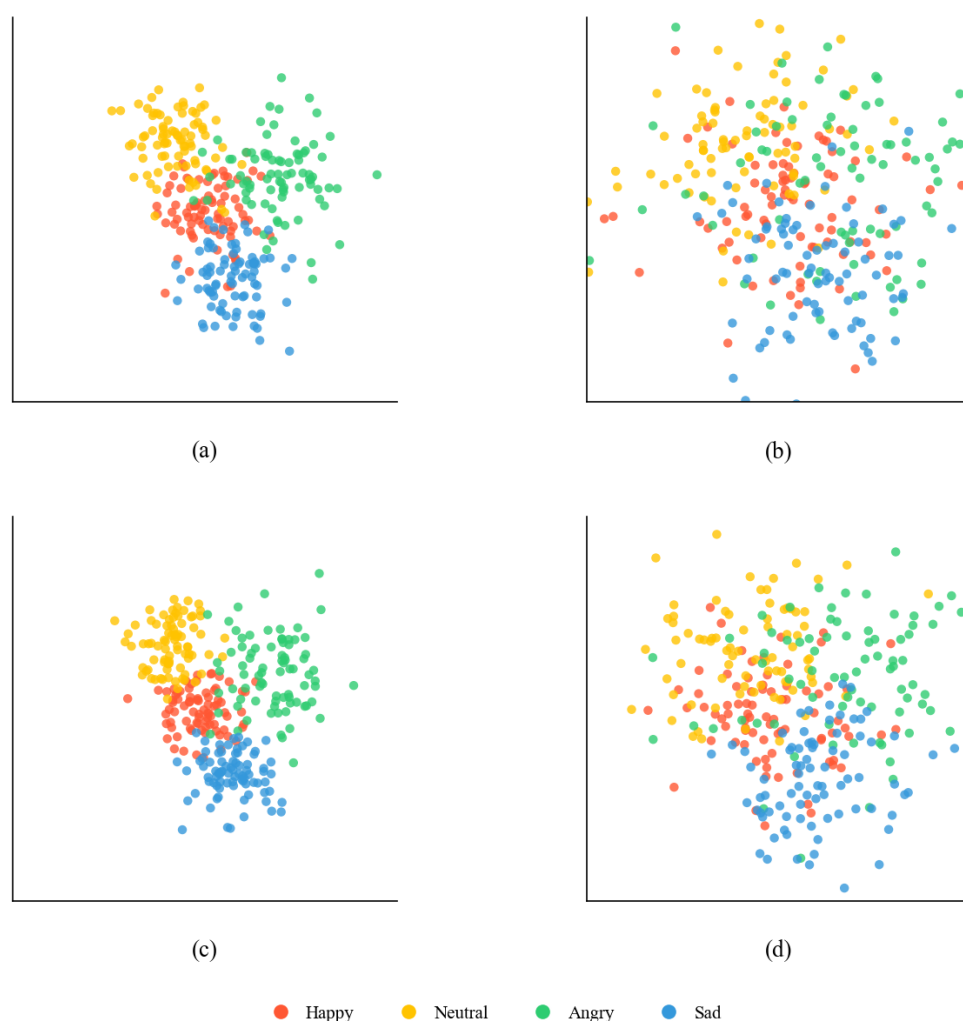


Figure 5. Comparison of decoupling effects: (a) MOSEI with Decoupling, (b) MOSEI without Decoupling, (c) MOSI with Decoupling, and (d) MOSI without Decoupling.

To isolate the contribution of the reward signal, we introduce the variant RDFD w/o Reward, which removes the entire reward mechanism while retaining the reconstruction-based initial weights R_m . In this setting, the dominant modality is selected solely based on feature reconstruction fidelity (i.e., the modality that is easiest to reconstruct is considered dominant), without any task-aware adaptive adjustment.

As shown in Table 5, removing the reward mechanism leads to consistent performance degradation across both datasets. On CMU-MOSI, ACC-2 drops by 1.12% (from 84.76% to 83.64%) and ACC-7 decreases by 1.35%. Similarly, on CMU-MOSEI, ACC-2 declines by 1.05% and ACC-7 by 1.33%. These results confirm that reconstruction fidelity alone—while useful for estimating feature stability—is insufficient for capturing task-relevant modality importance. A modality that is easy to reconstruct may contain redundant or neutral information rather than discriminative emotional cues. The reward signal Z_r provides sample-level, prediction-aware feedback that enables the model to dynamically amplify modalities contributing most to sentiment discrimination, which is essential for the superior performance of RDFD.

To further dissect the role of each component in the composite reward $\tilde{Z}_r$, we evaluate four variants where one term is removed individually. The results reveal distinct functional roles for each component:

Accuracy Reward (Z_a): Removing Z_a causes the largest performance drop among all reward components, ACC-2 83.98% on MOSI, a decrease of 0.78%. This confirms that direct task performance feedback is the most critical anchor for effective modality selection. Without Z_a , the model loses the primary signal for identifying which modality contributes most to correct predictions.

Uncertainty Reward (Z_u): Removing Z_u leads to a notable decline in ACC-7, 45.73% on MOSI, a decrease of 0.93%, which is more pronounced than its impact on ACC-2. This suggests that entropy-based uncertainty regularization plays a key role in fine-grained classification. It helps the model remain cautious on boundary samples where fine distinctions between adjacent sentiment levels are required, directly addressing the ACC-7 limitation discussed in Section 3.3.

Task Loss (Z_t) and Consistency (Z_c): While their individual impacts are relatively modest compared to Z_a and Z_u , removing them still degrades performance. Z_t provides continuous gradient information regarding prediction error magnitude, while Z_c ensures alignment between modality-specific and shared features, preventing semantic drift.

In conclusion, the ablation study demonstrates that the four reward components are complementary. Z_a drives the primary selection, Z_u refines fine-grained boundaries, and Z_t/Z_c stabilize the optimization process, collectively enabling the robust dynamic fusion of RDFD.

3.5. Hyperparameter Sensitivity Analysis

To validate the robustness of our model and determine the optimal configuration, we conduct a comprehensive sensitivity analysis on the reward component weights (α) and loss-balancing coefficients (λ).

Figure 6 illustrates the sensitivity of the model performance to the reward weights. The F1-Score exhibits an inverted-V shape. Conversely, the Mean Absolute Error (MAE) shows a negative correlation, reaching its minimum in the same range. This indicates that while accuracy (α_1), task loss (α_2), uncertainty (α_3), and consistency (α_4) all contribute to performance, their influence must be balanced. Over-emphasizing any single component leads to a degradation in F1-Score and an increase in MAE, confirming the necessity of the composite reward design.

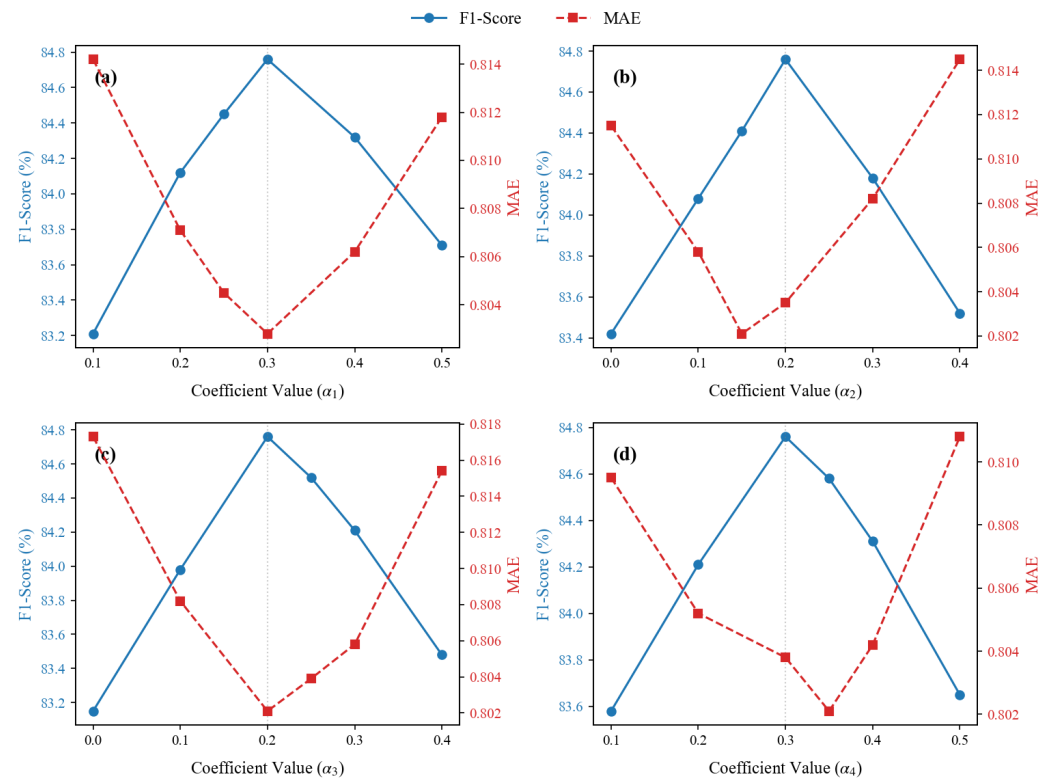


Figure 6. Sensitivity analysis of reward component weights (a–d) $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. The F1-Score (blue solid line) exhibits a unimodal trend, while the MAE (red dashed line) shows an inverse correlation. The vertical dotted line in each subfigure indicates the optimal coefficient value.

Figure 7 presents the sensitivity analysis for the loss coefficients. The orthogonality loss (λ_1) and specific discriminant loss (λ_3) achieve optimal performance around 0.10, while the shared discriminant (λ_2), reconstruction (λ_4), and classification (λ_5) losses peak at 0.20. These results justify our final hyperparameter selection, ensuring a stable trade-off between feature decoupling, reconstruction fidelity, and sentiment classification without overfitting to any single objective.

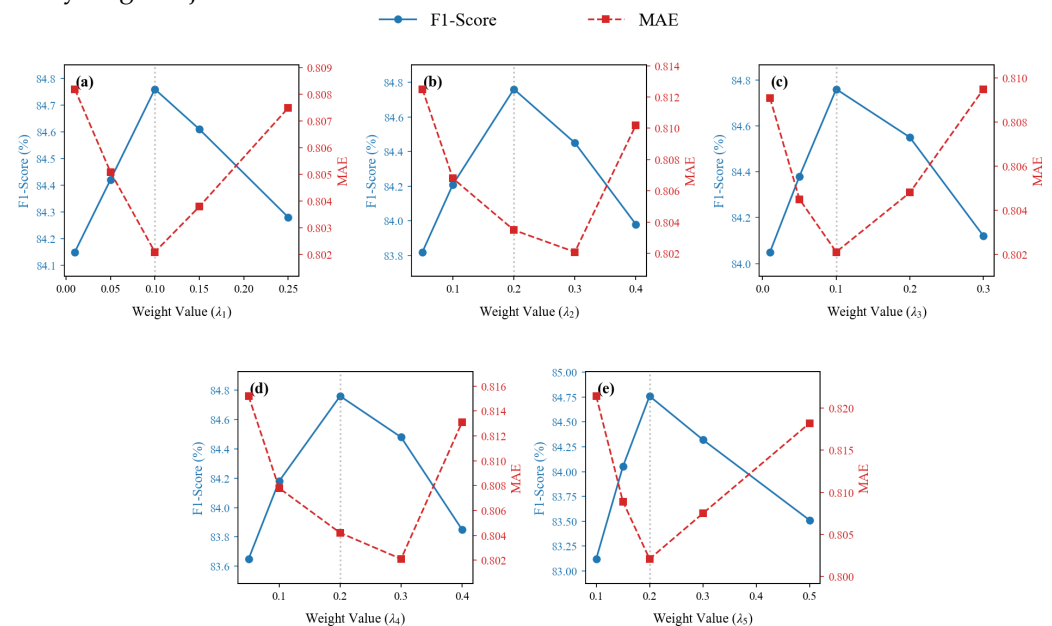


Figure 7. Sensitivity analysis of loss balancing coefficients (a–e) $\lambda_1, \dots, \lambda_5$. The performance curves display a similar unimodal distribution. The vertical dotted line in each subfigure indicates the optimal coefficient value.

4. Conclusions

In this paper, we propose the RDFD method to enhance the performance of Multimodal Sentiment Analysis (MSA). The RDFD framework consists of feature extraction, decoupling, enhancement, and prediction while incorporating a Reward-Guided Dynamic Fusion mechanism. By integrating reward-based indicators, RDFD dynamically adjusts fusion weights, effectively enhancing the model's representation capability. This adaptive fusion strategy contributes to improved performance across both datasets, demonstrating the effectiveness of our approach.

About the main method: After feature decoupling, the method realizes dynamic fusion through three steps: (1) the temporary dominant mode is calculated based on the feature reconstruction ratio; (2) the dominant modes are integrated through the cross-attention mechanism to complete emotion classification; and (3) the reward mechanism of Formula (17) (Section 2.4) is adopted to iteratively adjust the dominant mode selection until the reward value converges to a stable state.

Limitations and future work: Our method only considers the fully modal case. When facing the missing mode, the feature unentanglement and dominant-mode dynamic selection module may be limited, and further improvement is still needed.

Author Contributions: Conceptualization, H.Z.; methodology, H.Z. and Q.Y.; software, Z.G.; validation, Z.G.; formal analysis, H.Z.; investigation, Q.Y.; resources, Y.G.; data curation, Z.G.; writing—original draft preparation, Z.G.; writing—review and editing, L.L. and D.A.; visualization, Z.G.; supervision, Y.G. and L.L.; project administration, S.W.; funding acquisition, Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0110400 and Grant 2021ZD0110404; in part by the National Natural Science Foundation of China under Grant 62271377 and Grant 62201407; and in part by the Key Research and Development Program of Shaanxi Province under Grant 2021ZDLGY01-06, Grant 2022ZDLGY01-12, Grant 2023YBGY244, Grant 2023QCYLL28, Grant 2024GX-ZDCYL-02-08, and Grant 2024GX-ZDCYL-02-17; in part by Xi'an Social Sciences Planning Fund Project under Grant 25FZ63; in part by the Humanities and Social Sciences Research Project of the Shaanxi Provincial Department of Education under Grant 20JK0377.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets in this study were extracted from publicly available sources. These datasets are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Reward-Guided Dynamic Fusion and Modality Decoupling Training (RDFD) Algorithm

This appendix presents the complete training procedure of the revised RDFD framework. Unlike preliminary versions that relied on iterative threshold checks, the updated algorithm implements a single-pass, fully differentiable forward computation. The composite reward $\tilde{Z}_r$ modulates modality weights via gradient scaling during backpropagation, eliminating discrete decision boundaries and ensuring end-to-end optimization. Table A1 summarizes the six key phases aligned with Equations (9)–(25).

Table A1. Overview of Key Phases in the Revised RDFD Algorithm.

Phase	Objective and Mathematical Formulation
1. Decoupling	Extract shared $\mathbf{c}_m$ and specific $\mathbf{s}_m$ features via Equation (1).
2. Initial Weight Estimation	Compute reconstruction fidelity D_m (Equation (10)) and normalize to initial weights R_m (Equation (11)).
3. Composite Reward Computation	Calculate task-aware components Z_a, Z_t, Z_u, Z_c (Equations (11)–(14)), apply batch-wise Z-score normalization, and fuse via learnable coefficients α_i (Equations (15) and (16)).
4. Differentiable Weight Update	Modulate all modality weights in a single step: $R'_m = R_m \cdot (1 + \lambda \tilde{Z}_r)$ (Equation (17)), enabling smooth gradient flow.
5. Cross-Attention Fusion	Derive soft dominance probabilities p_m from R'_m (Equation (19)), construct attention $\mathbf{Q}, \mathbf{K}, \mathbf{V}$, scale values by $\mathbf{R}'$, and compute enhanced features $\mathbf{s}'_m$ (Equations (19)–(22)).
6. Prediction and Optimization	Concatenate $\mathbf{s}'_m \oplus \mathbf{c}'_m$ for sentiment prediction (Equation (23)) and minimize $\mathcal{L}_{\text{RDFD}}$ (Equation (24)) to jointly optimize all modules.

The algorithm dynamically identifies the most informative modality for each batch based on reconstruction fidelity and task-relevant feedback. By replacing hard selection with soft probability weighting and differentiable reward modulation, RDFD ensures that fusion is steered by continuously optimized signals rather than fixed assumptions, thereby enhancing adaptability and gradient stability during training.

Algorithm A1 Reward-Guided Dynamic Fusion and Modality Decoupling Training (RDFD)

Require: Batch of modal features $\{\mathbf{I}_a, \mathbf{I}_v, \mathbf{I}_t\}$, ground truth labels $\mathbf{Y}$

Ensure: Predicted labels $\hat{\mathbf{Y}}$, updated modality weights $\mathbf{R}' = [R'_a, R'_v, R'_t]$

```

1: // Phase 1: Feature Decoupling
2:  $\mathbf{c}_m, \mathbf{s}_m \leftarrow \text{Decoupler}(\mathbf{I}_m; \theta_{\text{dec}}), \quad \forall m \in \{a, v, t\}$ 
3:
4: // Phase 2: Initial Weight Estimation (Modal Refactor)
5:  $\hat{\mathbf{s}}_m \leftarrow D_{\text{sm}}(\mathbf{s}_m; \theta_2)$ 
6:  $D_m \leftarrow \exp(-k \|\hat{\mathbf{s}}_m - \mathbf{s}_m\|^2)$ 
7:  $R_m \leftarrow D_m / \sum_{m' \in \{a, v, t\}} D_{m'}$ 
8:
9: // Phase 3: Composite Reward Computation
10: Compute  $Z_a, Z_t, Z_u, Z_c$  via Equations (13)–(16)
11:  $\tilde{Z}_i \leftarrow (Z_i - \mu_{Z_i}) / (\sigma_{Z_i} + \epsilon), \quad i \in \{a, t, u, c\}$  ▷ Batch-wise Z-score normalization
12:  $\tilde{Z}_r \leftarrow \alpha_1 \tilde{Z}_a - \alpha_2 \tilde{Z}_t - \alpha_3 \tilde{Z}_u + \alpha_4 \tilde{Z}_c$  ▷ Equation (16),  $\alpha_i$  learnable
13:
14: // Phase 4: Differentiable Weight Update
15:  $R'_m \leftarrow R_m \cdot (1 + \lambda \tilde{Z}_r), \quad \forall m$  ▷ Equation (17),  $\lambda$  learnable, single-step
16:
17: // Phase 5: Cross-Attention Fusion
18:  $p_m \leftarrow \frac{\exp(R'_m / \tau)}{\sum_{m'} \exp(R'_{m'} / \tau)}$  ▷ Soft dominance, Equation (18),  $\tau = 0.2$ 
19:  $\mathbf{Q} \leftarrow \sum_m p_m \mathbf{W}_Q \mathbf{s}_m$ 
20:  $\mathbf{K} \leftarrow \mathbf{W}_K[\mathbf{s}_a; \mathbf{s}_v; \mathbf{s}_t], \quad \mathbf{V} \leftarrow \mathbf{W}_V[\mathbf{s}_a; \mathbf{s}_v; \mathbf{s}_t]$ 
21:  $\mathbf{V}' \leftarrow \mathbf{V} \odot \mathbf{R}'$  ▷ Broadcast  $\mathbf{R}'$  to match  $\mathbf{V}$  dimensions
22:  $\mathbf{s}'_m \leftarrow \text{LayerNorm}\left(\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \mathbf{V} \odot \mathbf{R}' + \mathbf{Q}\right)$  ▷ Equation (22)
23:
24: // Phase 6: Prediction & Loss
25:  $\mathbf{c}'_m \leftarrow \text{SelfAttn}(\mathbf{c}_m)$  ▷ Shared enhancement, Equations (5) and (6)
26:  $\hat{\mathbf{Y}} \leftarrow \mathcal{M}(\mathbf{s}'_m \oplus \mathbf{c}'_m; \theta_{\text{cls}})$  ▷ Equation (23)
27:  $\mathcal{L}_{\text{RDFD}} \leftarrow \lambda_1 \mathcal{L}_{\text{ort}} + \lambda_2 \mathcal{L}_c + \lambda_3 \mathcal{L}_s + \lambda_4 \mathcal{L}_d + \lambda_5 \mathcal{L}_{\text{cls}}$  ▷ Equation (25)
28: return  $\hat{\mathbf{Y}}, \mathbf{R}', \mathcal{L}_{\text{RDFD}}$ 

```

References

1. Poria, S.; Hazarika, D.; Majumder, N.; Mihalcea, R. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.* **2020**, *14*, 108–132. [\[CrossRef\]](#)
2. Awal, R.; Ahmadi, S.; Zhang, L.; Agrawal, A. VisMin: Visual Minimal-Change Understanding. *arXiv* **2024**, arXiv:2407.16772.
3. Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 14375–14385.
4. Singh, U.; Abhishek, K.; Azad, H.K. A Survey of Cutting-edge Multimodal Sentiment Analysis. *ACM Comput. Surv.* **2024**, *56*, 1–38. [\[CrossRef\]](#)
5. Pandey, A.; Vishwakarma, D.K. Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey. *Appl. Soft Comput.* **2024**, *152*, 111206. [\[CrossRef\]](#)
6. Nanli, Z.; Ping, Z.; Weiguo, L.; Meng, C. Sentiment analysis: A literature review. In Proceedings of the 2012 International Symposium on Management of Technology (ISMOT), Hangzhou, China, 8–9 November 2012; pp. 572–576.
7. Zhao, W.; Zhao, Y.; Lu, X. CauAIN: Causal Aware Interaction Network for Emotion Recognition in Conversations. In Proceedings of the IJCAI, Vienna, Austria, 23–29 July 2022; pp. 4524–4530.
8. Wang, Q.; Wen, Z.; Ding, K.; Liang, B.; Xu, R. Cross-Domain Sentiment Analysis via Disentangled Representation and Prototypical Learning. *IEEE Trans. Affect. Comput.* **2024**, *16*, 264–276. [\[CrossRef\]](#)
9. Fu, Y.; Huang, B.; Wen, Y.; Zhang, P. FDR-MSA: Enhancing Multimodal Sentiment Analysis through feature disentanglement and reconstruction. *Knowl.-Based Syst.* **2024**, *297*, 111965. [\[CrossRef\]](#)
10. Zeng, Y.; Yan, W.; Mai, S.; Hu, H. Disentanglement Translation Network for multimodal sentiment analysis. *Inf. Fusion* **2024**, *102*, 102031. [\[CrossRef\]](#)
11. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for Multimodal Sentiment Analysis. *arXiv* **2017**, arXiv:1707.07250. [\[CrossRef\]](#)
12. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
13. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [\[CrossRef\]](#)
14. Guo, J.; Song, B.; Zhang, P.; Ma, M.; Luo, W.; Lv, J. Affective video content analysis based on multimodal data fusion in heterogeneous networks. *Inf. Fusion* **2019**, *51*, 224–232. [\[CrossRef\]](#)
15. Qian, S.; Xue, D.; Fang, Q.; Xu, C. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4794–4811. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Lei, Y.; Yang, D.; Li, M.; Wang, S.; Chen, J.; Zhang, L. Text-oriented modality reinforcement network for multimodal sentiment analysis from unaligned multimodal sequences. In *Proceedings of the CAAI International Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 189–200.
17. 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. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 10074–10082.
18. Wang, P.; Zhou, Q.; Wu, Y.; Chen, T.; Hu, J. DLF: Disentangled-Language-Focused Multimodal Sentiment Analysis. *arXiv* **2024**, arXiv:2412.12225. [\[CrossRef\]](#)
19. Zhao, S.; Yang, Z.; Shi, H.; Feng, X.; Meng, L.; Qin, B.; Yan, C.; Tao, J.; Ding, G. SDRS: Sentiment-Aware Disentangled Representation Shifting for Multimodal Sentiment Analysis. *IEEE Trans. Affect. Comput.* **2025**, *16*, 1802–1813. [\[CrossRef\]](#)
20. 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, Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
21. Yang, D.; Huang, S.; Kuang, H.; Du, Y.; Zhang, L. Disentangled representation learning for multimodal emotion recognition. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 1642–1651.
22. Sun, H.; Wang, H.; Liu, J.; Chen, Y.W.; Lin, L. CubeMLP: An MLP-based model for Multimodal Sentiment Analysis and depression estimation. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 3722–3729.
23. Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; Yu, T. Learning language-guided adaptive hyper-modality representation for Multimodal Sentiment Analysis. *arXiv* **2023**, arXiv:2310.05804.
24. Wang, D.; Guo, X.; Tian, Y.; Liu, J.; He, L.; Luo, X. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognit.* **2023**, *136*, 109259. [\[CrossRef\]](#)
25. Lin, R.; Hu, H. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for Multimodal Sentiment Analysis. *arXiv* **2022**, arXiv:2210.14556.

26. Ren, M.; Huang, X.; Liu, J.; Liu, M.; Li, X.; Liu, A.A. Maln: Multimodal adversarial learning network for conversational emotion recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 6965–6980. [[CrossRef](#)]
27. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **2016**, *31*, 82–88. [[CrossRef](#)]
28. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
29. Liang, T.; Lin, G.; Feng, L.; Zhang, Y.; Lv, F. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 8148–8156.
30. Lv, F.; Chen, X.; Huang, Y.; Duan, L.; Lin, G. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 2554–2562.
31. Mao, H.; Yuan, Z.; Xu, H.; Yu, W.; Liu, Y.; Gao, K. M-SENA: An integrated platform for Multimodal Sentiment Analysis. *arXiv* **2022**, arXiv:2203.12441. [[CrossRef](#)]
32. Feng, X.; Lin, Y.; He, L.; Li, Y.; Chang, L.; Zhou, Y. Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis. *arXiv* **2024**, arXiv:2410.04491. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.