Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079330

ARTICLE

Hierarchical Contrastive Representation Learning Guided by Multimodal Feature Decomposition for Multimodal Sentiment Analysis

Hongbin Wang[1,2], **Liusong Li**[1,2] **and Di Jiang**[1,2,*]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China
[2]Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, China

*Corresponding Author: Di Jiang. Email: alexjiang_yn@163.com

Received: 20 January 2026; Accepted: 22 April 2026; Published: 15 June 2026

ABSTRACT: Multimodal sentiment analysis aims to fuse emotional information from data across different modalities to predict human emotional states. Although existing multimodal sentiment analysis methods have made significant progress, the heterogeneity between modalities still leads to an imbalance in feature space distribution, thereby hindering the effective learning and fusion of multimodal representations. In addition, the presence of emotion-irrelevant information in auxiliary modalities is another major factor contributing to differences in feature space distributions. To address this issue, we propose a Hierarchical Contrastive Representation Learning framework with Multimodal Feature Decoupling (HCRL-MFD). To reduce emotion-irrelevant information and optimize feature representations, we introduce a Multimodal Feature Decoupling (MFD) mechanism. This mechanism decouples the multimodal feature space into two subspaces: an emotion-relevant core space and an emotion-irrelevant space, effectively separating key information from redundant signals. This helps alleviate modality heterogeneity and improves the granularity of feature representations in multimodal learning. Meanwhile, we design a Hierarchical Contrastive Representation Learning (HCRL) strategy to further explore intra-modal interactions and cross-sample relationships between different emotional states, aiming to reduce distributional discrepancies across modalities. Finally, we concatenate and fuse the features from each modality for sentiment prediction. We conduct extensive experiments on two publicly available multimodal sentiment analysis datasets, CMU-MOSI and CMU-MOSEI, and the results demonstrate that our model achieves significant performance gains and shows strong competitiveness compared to existing methods.

KEYWORDS: Multimodal sentiment analysis; feature decomposition; contrastive representation learning

1 Introduction

With the rapid development of social media platforms such as Twitter, TikTok, and YouTube, multimodal video data containing text, audio, and visual elements has seen an explosive growth, making multimodal sentiment analysis (MSA) an important and highly focused research area [1,2]. Compared to unimodal approaches, multimodal sentiment analysis typically demonstrates more robustness by fully leveraging the complementary relationships between different modalities, offering significant advantages in enhancing the understanding of human emotions [3]. At the same time, the widespread use of mobile devices not only facilitates the capture of diverse emotional cues from users [4] but also enables the application of multimodal sentiment analysis across various economic and social domains [5], such as healthcare (e.g., as a tool for predicting mental health), education (e.g., understanding student frustration and providing counseling), and criminology (e.g., deception detection). As a result, an increasing number of researchers are delving into this promising and ever-evolving field.

Previous multimodal sentiment analysis methods can generally be categorized into two types: one focuses on representation learning, aiming to extract refined semantic representations rich in diverse emotional cues from each modality, thereby improving emotion understanding and the efficiency of multimodal fusion in relationship modeling [6–10]; the other emphasizes the design of complex fusion mechanisms to obtain high-quality multimodal joint representations [11–14]. Although prior research suggests that fusing complementary emotional information from various modalities helps generate more effective joint representations [15,16], the heterogeneity between modalities inevitably causes imbalances in feature space distributions, making it more difficult to mine complementary information [17]. Additionally, the language modality typically provides higher-quality features compared to visual and audio modalities, thus dominating in multimodal sentiment analysis, while the visual and audio modalities often contain significant amounts of emotion-irrelevant redundant information. For instance, low lighting in videos may obscure facial expressions, and background noise in audio may mask emotional details, which further exacerbates the differences in feature space distributions between modalities [18,19].

To address the above issues, we propose a novel framework termed Hierarchical Contrastive Representation Learning guided by Multimodal Feature Decoupling (HCRL-MFD). It aims to simultaneously alleviate the feature distribution gap caused by modality heterogeneity and the interference of emotion-irrelevant redundant noise in auxiliary modalities. Specifically, we first extract the foundational features from the text, visual, and audio modalities using feature extractors. Then, we introduce three specific decoupling objectives to guide the feature decoupling process, effectively capturing core-sentiment information while explicitly eliminating redundant data. To address the significant feature space distribution differences between modalities, we design a hierarchical contrastive learning mechanism that optimizes the feature space distributions within and across modalities through both intra-modal and cross-modal contrastive learning, thus fully exploring fine-grained emotional cues.

Our main contributions can be summarized as follows:

- We propose the Hierarchical Contrastive Representation Learning guided by Multimodal Feature Decoupling (HCRL-MFD) framework, which integrates four synergistic modules: feature extraction, multimodal feature decoupling, hierarchical contrastive representation learning, and fusion prediction.
- We introduce an Multimodal Feature Decoupling (MFD) mechanism, shifting the learning paradigm from traditional "shared-private" alignment to "sentiment-relevant vs. emotion-irrelevant" purification. Guided by task-specific constraints, it explicitly disentangles the representations of each modality into core-sentiment and redundant subspaces, actively filtering out detrimental intra-modal noise to provide a strictly purified baseline for downstream fusion.
- We design a Hierarchical Contrastive Representation Learning (HCRL) strategy to mitigate feature distribution gaps caused by modality heterogeneity. By jointly optimizing intra-modal and cross-modal contrastive objectives, HCRL strictly aligns the purified feature distributions while enhancing the discrimination of fine-grained emotional intensities.
- We conduct extensive experiments on two public benchmark datasets, CMU-MOSI and CMU-MOSEI. The empirical results demonstrate that our proposed model significantly outperforms recent state-of-the-art baselines and achieves superior performance.

2 Related Works

In this subsection, we briefly review the MSA approach as well as research advances in contrastive learning in this area.

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis aims to interpret emotions conveyed through different channels such as text, audio, and visual cues, integrating both linguistic and non-linguistic information (including visual and auditory clues) to better understand human emotions. It is widely applied in human-computer interaction and affective computing fields [20].

Zhuang et al. [21] proposed GLoMo to resolve granularity mismatch by modeling correlations between decomposed global and local representations via a Global-Local Interaction Module. Wang et al. [22] proposed leveraging large vision-language models to generate contextual world knowledge for multimodal sentiment analysis, complemented by a training-free contextual fusion mechanism designed to mitigate noise and handle hard samples. Tsai et al. [23] employed directional cross-modal attention to achieve implicit alignment and capture long-range dependencies. Yu et al. [24] utilized self-supervised unimodal label generation and a weight adjustment strategy to jointly learn modality consistency and differences. Han et al. [8] maximized mutual information between fusion results and unimodal inputs to preserve task-relevant information. Zhu et al. [11] enhanced text representations via emotional knowledge graphs and aligned multimodal features using text-guided attention. Huang et al. [25] improved model reliability through causal-aware text debiasing and counterfactual cross-modal attention mechanisms. Zhou et al. [26] proposed DPDF-LQ, leveraging complementary parallel paths to effectively balance global dependencies and fine-grained features.

However, although existing methods have made progress in improving the accuracy of multimodal sentiment analysis, most of them fail to fully consider the heterogeneity between modalities and differences in feature space distributions. This leads to potential imbalance in multimodal representation learning, limiting the effective fusion of information and the performance of models.

2.2 Disentangled Multimodal Representation Learning

Disentangled multimodal representation learning aims to explicitly model the shared and modality-specific representations before fusion, by separating the different factors of variation from each modality into independent subspaces. This approach helps reduce the gap between modalities by extracting both shared and unique latent factors across different modalities. By explicitly modeling and disentangling desired informational attributes, disentangled learning aligns semantically related concepts across modalities, thereby effectively mitigating the modality gap caused by heterogeneity. Moreover, by providing more structured and explicit representations of the latent factors, disentangled learning further facilitates the effective fusion of multimodal information.

Tsai et al. [27] factorized multimodal data into discriminative and generative factors to distinctively learn joint and specific information. Zeng et al. [28] utilized adversarial learning to capture modality-invariant embeddings and achieved shared-private distribution matching to facilitate fusion. Hazarika et al. [29] projected modalities into invariant and specific subspaces to alleviate distribution differences and enhance feature representation. Yang et al. [30] reduced inter-modal gaps by designing shared and private encoders to explicitly disentangle modality-specific and shared representations. Wang et al. [31] proposed geometric metric regularization for feature decoupling and a Language-Focused Attractor to extract complementary cues for hierarchical prediction.

However, previous methods typically project modalities into two shared and private feature subspaces for feature decoupling, overlooking the large amount of emotion-irrelevant redundant information in auxiliary modalities. In contrast, our proposed Multimodal Feature Decoupling (MFD) mechanism aims

to decouple the core-sentient information from redundant information within the features, removing the redundant data while providing cleaner features for subsequent contrastive representation learning.

2.3 Contrastive Learning

In recent years, contrastive learning has attracted significant attention and achieved groundbreaking progress in the field of multimodal sentiment analysis. Existing methods can be broadly categorized into two main approaches: self-supervised contrastive learning [32–37] and supervised contrastive learning [38–40], with the core difference lying in whether label information is utilized to construct positive and negative sample pairs. Against this research background, scholars have continuously advanced the application boundaries of contrastive learning in multimodal sentiment analysis.

Mai et al. [41] pioneered the HyCon framework to capture inter-sample and inter-class relationships via cross-modal interactions, thereby reducing modality discrepancies. Yang et al. [7] developed ConFEDE, utilizing a unified contrastive loss for modality disentanglement and supervised learning to simultaneously capture consistency and divergence. Yu et al. [42] proposed ConKI, leveraging knowledge injection and hierarchical contrastive learning to collaboratively optimize specific and general representations. Yang et al. [43] introduced an emotion intensity-guided framework that projects data into a unified space for the joint learning of shared and specific features. Liu et al. [44] employed contrastive learning on heterogeneous sample pairs to extract modality-invariant features, aligning representations to guide missing data reconstruction.

However, existing contrastive learning methods often overlook the inherent heterogeneity between modalities, especially the differences in distribution of different modalities within the feature space. Such differences exacerbate the issue of information misalignment between modalities and hinder effective multimodal representation learning. Therefore, we propose the Hierarchical Contrastive Representation Learning (HCRL) method, aiming to optimize the feature space of each modality through joint intra-modal and inter-modal contrast, thereby reducing distribution differences between modalities.

3 Methodology

3.1 Problem Definition

The core goal of Multimodal Sentiment Analysis (MSA) is to achieve emotion prediction by effectively extracting and fusing multimodal information such as text, visual, and audio data. We use BERT [45], COVAREP [46], and OpenFace [47], respectively to obtain raw feature embedding sequences from the three modalities of data, and we denote these input sequences as $I_m \in \mathbb{R}^{L_m \times d_m}$. Here, L_m denotes the input sequence length of each modality, d_m corresponds to the embedding dimension of the feature vector for each modality, and $m \in \{T, V, A\}$. Consistent with previous works [7,11,24,43], we formulate MSA as a regression task to ensure fair comparison.

3.2 Overall Architecture

The overall workflow of the Hierarchical Contrastive Representation Learning with Multimodal Feature Decoupling (HCRL-MFD) framework is illustrated in Fig. 1. The model consists of four main modules: the feature extraction module, the Multimodal Feature Decoupling (MFD) module, the Hierarchical Contrastive Representation Learning (HCRL) module, and the fusion and prediction module. The core components are the MFD and HCRL modules.

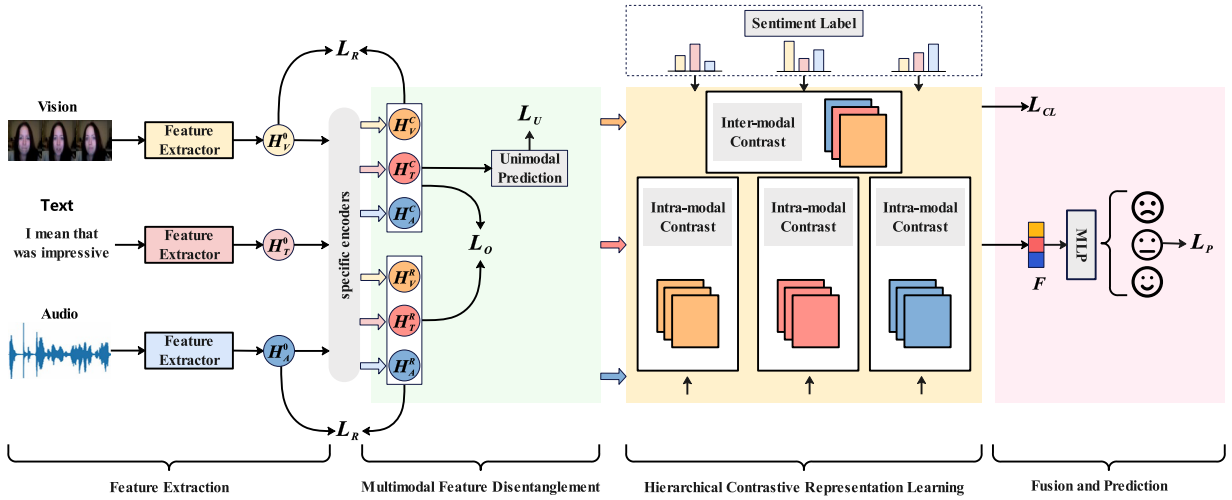


Figure 1: Overall architecture of the HCRL-MFD model. The framework consists of four main components: Feature Extraction (FE), Multimodal Feature Decomposition (MFD), Hierarchical Contrastive Representation Learning (HCRL), and Fusion and Prediction (FP), which are designed to adequately capture efficient multimodal representations.

Specifically, in the MFD module, to ensure the effective and rigorous decoupling of feature information, we design three distinct decoupling objectives. To avoid any terminological ambiguity, these objectives consist of two spatial geometric constraints to guarantee information conservation and mutual independence, alongside a task-specific semantic constraint to provide accurate semantic orientation. Meanwhile, the HCRL module is designed to enhance the model's ability to capture fine-grained emotional information and optimize the spatial distribution of features across modalities. Finally, the fused features are concatenated and input into an MLP classifier to complete the final emotion prediction.

3.3 Feature Extraction

We begin by encoding the input features of each modality using their respective feature extractors to obtain the basic modality representations. Specifically, for the text modality, we use a pre-trained BERT to encode the input sentence, generating the basic textual representation as follows:

$$H_T^0 = Bert\left(I_T; \theta_T^{Bert}\right) \in \mathbb{R}^{L_T \times d_T} \tag{1}$$

For the visual and audio modalities, it is important to note that their extracted inputs are inherently frame-level time-series data. To effectively model their temporal dynamics, we first apply standard Positional Encodings to the sequences to retain their strict sequential order. Subsequently, we utilize two independent Transformer [48] encoders to process these sequences. This design empowers the self-attention mechanism to explicitly capture the global temporal dynamics and complex contextual interactions across different time steps. The basic visual and audio representations are captured as follows:

$$H_V^0 = Transformer\left(I_V; \theta_V^{Transformer}\right) \in \mathbb{R}^{L_V \times d_V} \tag{2}$$

$$H_A^0 = Transformer\left(I_A; \theta_A^{Transformer}\right) \in \mathbb{R}^{L_A \times d_A} \tag{3}$$

where d_m denotes the output feature dimension of the respective modality encoder. In our experiments, we explicitly set these dimensions to $d_T = 768$, $d_V = 128$, and $d_A = 64$.

3.4 Multimodal Feature Decomposition

In Multimodal Sentiment Analysis (MSA), the audio and visual modalities often contain more noise compared to the text modality, which significantly limits the representational power and performance of multimodal models. To mitigate this issue, we propose a Multimodal Feature Decoupling (MFD) mechanism, with the core goal of extracting the emotionally relevant core components from the foundational features of each modality while gradually removing emotion-irrelevant redundant information.

Specifically, for each modality, we input its foundational feature H_m^0 into an core-sentient encoder E_m^C and a non-emotional redundant encoder E_m^R, respectively obtaining the core-sentient feature H_m^C and the emotion-irrelevant feature H_m^R. Formally, the two specific encoders are defined as follows:

$$H_m^C = E_m^C(H_m^0, \theta_C) \tag{4}$$

$$H_m^R = E_m^R(H_m^0, \theta_R) \tag{5}$$

where θ_C and θ_R represent the parameters of the core-sentient encoder and the emotion-irrelevant encoder, respectively. In this work, both encoders are implemented as cascaded Transformer layers.

To achieve effective decoupling of features and reduce emotion-irrelevant and conflicting information, we design three specific loss functions as decoupling objectives. To avoid ambiguity in terminology, it is important to clarify that the first two losses explicitly serve as spatial geometric constraints to guide feature decoupling, while the third (unimodal prediction loss) acts as a task-driven semantic constraint.

First, to ensure that no fundamental information is lost during the decoupling process and to logically validate the structural relationship between the features, we introduce a reconstruction loss $\mathcal{L}_R$ as the primary spatial geometric constraint. As the foundational feature H_m^0 inherently contains both the sentiment signal and the redundant noise, the original signal should be mathematically reconstructible from the decoupled core-sentiment and redundant components. We use the Mean Squared Error (MSE) to measure this reconstruction penalty:

$$\mathcal{L}_R = \frac{1}{N}\sum_{i=1}^{N}\left\|H_m^0 - (H_m^C + H_m^R)\right\|_2^2 \tag{6}$$

This structural constraint ensures that the explicit decoupling process retains all original information, forcing the respective encoders to strictly separate—rather than arbitrarily discard—the input signals into distinct representational components.

Secondly, as another geometric constraint, considering the randomness in parameter initialization during the early stages of training, the H_m^C and H_m^R extracted by the encoders often exhibit unavoidable information overlap. To address this, we introduce an orthogonal constraint loss $\mathcal{L}_O$ to enhance their independence in the feature space and avoid cross-information interference. This loss is defined by minimizing the Frobenius norm of the inner product between the two, as follows:

$$\mathcal{L}_O = \left\|(H_m^C)^T H_m^R\right\|_F^2 \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This constraint forces H_m^C and H_m^R to be as orthogonal as possible in the feature space, allowing them to focus on different semantic subspaces, thereby improving the reliability and robustness of the decoupled representations.

Finally, to ensure that the decoupled core-sentiment features intrinsically possess strong emotion-discriminative capabilities, we introduce a task-driven semantic constraint: the unimodal prediction loss $\mathcal{L}_U$.

Recognizing that the text, audio, and visual modalities exhibit fundamentally different data distributions, representational structures, and optimization spaces, forcing them through a shared-weight classifier would inevitably cause cross-modal interference and hinder modality-specific optimization.

Therefore, we explicitly allocate independent, parameter-specific Multi-Layer Perceptron (MLP) classifiers for each modality. The decoupled core-sentiment feature H_m^C of each modality is fed into its uniquely corresponding predictor MLP_m to generate the unimodal sentiment prediction $\hat{y}_m$:

$$\hat{y}_m = MLP_m(H_m^C) \tag{8}$$

Subsequently, we compute the unimodal prediction loss $\mathcal{L}_U$ by calculating the Mean Squared Error (MSE) between these modality-specific predictions and the ground-truth sentiment labels y:

$$\mathcal{L}_U = \frac{1}{3} \sum_{m \in \{T,V,A\}} \mathrm{MSE}(\hat{y}_m, y) \tag{9}$$

By deploying completely unshared, modality-specific classifiers (MLP_T, MLP_A, MLP_V), this mechanism strictly forces the Multimodal Feature Decoupling (MFD) module to isolate the purest core-sentiment signals within each distinct modality space, effectively eliminating optimization bottlenecks and inter-modal conflicts. Since this loss term is directly derived from the emotion prediction task, it acts as a robust semantic constraint, forcing H_m^C to possess strong emotion-discriminative power and enhancing its semantic effectiveness for the downstream multimodal fusion.

Finally, we combine the three aforementioned loss terms with weighted summation to form the final feature decoupling loss:

$$\mathcal{L}_D = \sum_{k \in \{R,O,U\}} \lambda_k \mathcal{L}_k \tag{10}$$

where λ_k are weighting coefficients that control the contribution of each regularization term.

3.5 Hierarchical Contrastive Representation Learning

In multimodal sentiment analysis tasks, there exist significant differences in emotional expression among different samples, and the various modalities naturally exhibit inconsistencies in representation forms and expressive capabilities. This combination of modality heterogeneity and sentiment variability leads to difficulties in aligning features across modalities.

To alleviate this issue and achieve more robust cross-modal feature modeling, we propose a hierarchical contrastive learning strategy, aiming to enhance the consistency and discriminative power of multimodal representations through intra-modal and inter-modal contrastive learning at two levels.

Specifically, we first construct an initial set of positive and negative sample pairs based on the true sentiment labels of the samples. For any sample x_a in a training batch B, if the difference in their labels is smaller than a predefined threshold δ, they are treated as candidate positive pairs; otherwise, as negative pairs. The formal definition is as follows:

$$\mathcal{P}^a = \{(x_a, x_b) \,||y_a - y_b| \le \delta\}, a \ne b \tag{11}$$

$$\mathcal{N}^a = \{(x_a, x_c) \,||y_a - y_c| > \delta\}, a \ne c \tag{12}$$

Subsequently, for each candidate pair, we further compute the intra-modal and inter-modal cosine similarity based on their feature representations, in order to reflect the distance relationships in both intra-

and inter-modal feature spaces. Next, we rank the candidate positive and negative pairs according to the computed cosine similarity scores.

For each modality $m \in \{T, V, A\}$, the intra-modal similarity is defined as:

$$sim_{intra} = \frac{(H_m^a)^\top H_m^b}{\|H_m^a\| \cdot \|H_m^b\|} \tag{13}$$

The corresponding intra-modal similarity ranking matrix is defined as:

$$S_{\text{intra}} = \left\{\text{sim}_{\text{intra}}^1, \text{sim}_{\text{intra}}^2, \cdots, \text{sim}_{\text{intra}}^n\right\} \tag{14}$$

In contrast, the inter-modal similarity and its corresponding ranking matrix are defined as:

$$\text{sim}_{\text{inter}} = \frac{(H_{m_1}^a)^\top H_{m_2}^b}{\|H_{m_1}^a\| \cdot \|H_{m_2}^b\|}, \quad m_1 \neq m_2 \tag{15}$$

$$S_{\text{inter}} = \left\{\text{sim}_{\text{inter}}^1, \text{sim}_{\text{inter}}^2, \cdots, \text{sim}_{\text{inter}}^n\right\} \tag{16}$$

Finally, from the obtained intra-modal and inter-modal ranking matrices, we correspondingly select positive and negative sample pairs. Specifically, we choose the top-k sample pairs with the highest similarity scores from the ranking matrices as positive pairs, and the bottom-k sample pairs with the lowest similarity scores as negative pairs.

This selection strategy is theoretically inspired by established contrastive learning paradigms in recent multimodal sentiment analysis literature, aiming to effectively strike a balance between providing sufficient contrastive signals and avoiding the introduction of overly noisy, less-informative pairs. In our framework, k serves as a tunable hyperparameter rather than a fixed architectural constraint. Based on comprehensive empirical verification and ablation studies (detailed in Section 4.6.5), we optimally set $k = 2$ for our main experiments. The specific strategies for intra-modal and inter-modal sample selection are illustrated in Fig. 2.

(1) Intra-modal pairs

Positive pairs:

$$\mathcal{P}_{intra}^a = \left\{\left(T^a, T^b\right), \left(V^a, V^b\right), \left(A^a, A^b\right)\right\} \in S_{intra} \tag{17}$$

Negative pairs:

$$\mathcal{N}_{intra}^a = \left\{\left(T^a, T^c\right), \left(V^a, V^c\right), \left(A^a, A^c\right)\right\} \in S_{intra} \tag{18}$$

(2) Inter-modal pairs

Positive pairs:

$$\begin{aligned}\mathcal{P}_{inter}^a = \big\{&(T^a, V^b),\ (T^a, A^b),\ (V^a, T^b),\\ &(V^a, A^b),\ (A^a, T^b),\ (A^a, V^b)\big\}\\ &\cup \big\{(T^a, V^a),\ (T^a, A^a),\ (V^a, A^a)\big\} \in S_{inter}\end{aligned} \tag{19}$$

Negative pairs:

$$\begin{aligned}\mathcal{N}_{inter}^a = \big\{&(T^a, V^c),\ (T^a, A^c),\ (V^a, T^c),\\ &(V^a, A^c),\ (A^a, T^c),\ (A^a, V^c)\big\} \in S_{inter}\end{aligned} \tag{20}$$

To jointly optimize intra-modal and inter-modal representational consistency, we construct a unified contrastive loss function:

$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\mathrm{sim}(\alpha,\beta)/\tau\right)}{\sum_{(\alpha,\gamma)}\exp\left(\mathrm{sim}(\alpha,\gamma)/\tau\right)}, \quad \text{where } (\alpha,\beta)\in P^a, (\alpha,\gamma)\in P^a\cup N^a \tag{21}$$

where N is the number of samples, τ is the temperature coefficient, and $sim(\cdot,\cdot)$ denotes cosine similarity.

This contrastive loss is designed to simultaneously incorporate both intra-modal and inter-modal positive and negative pairs during training. It enhances the consistency of core-sentiment features within a single modality while balancing the feature distributions across modalities. Moreover, it significantly amplifies fine-grained emotional differences between samples, enabling the model to more precisely capture and distinguish subtle variations in emotional intensity.

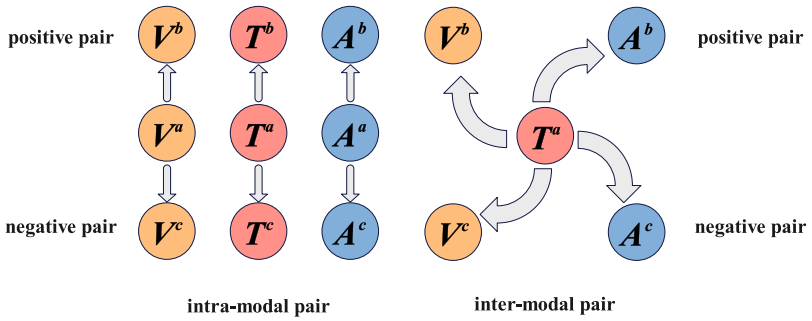


Figure 2: The strategy of selecting positive and negative pairs from intra-modal and inter-modal similarity ranking matrices.

3.6 Fusion and Prediction

Unlike traditional multimodal approaches that rely on highly complex, parameter-intensive fusion mechanisms (e.g., Cross-Attention or Tensor Fusion) to model inter-modal dynamics, our framework adopts a straightforward concatenation strategy followed by a Multi-Layer Perceptron (MLP). This design choice is strictly grounded in the core philosophy of our methodology: resolving multimodal heterogeneity and noise during the upstream *representation learning* stage rather than the downstream *fusion* stage.

Because our Multimodal Feature Decomposition (MFD) and Hierarchical Contrastive Representation Learning (HCRL) modules have already rigorously purified modality-specific noises and actively aligned the core-sentiment features in a unified latent space, the resulting representations possess high semantic consistency. Consequently, simple concatenation is highly sufficient to decode the sentiment signals. Forcing these already well-aligned representations through complex fusion blocks would introduce redundant computational overhead and increase the risk of overfitting (as empirically verified in Section 4.6.6).

Specifically, we concatenate the core-sentiment features of all modalities into a unified feature vector F^C and feed it into an MLP classifier to predict the sentiment score $\hat{y}$:

$$F^C = \mathrm{Concat}\left(H_T^C, H_V^C, H_A^C\right) \tag{22}$$

$$\hat{y} = \mathrm{MLP}\left(F^C\right) \tag{23}$$

For the Multimodal Sentiment Analysis (MSA) task, we adopt the standard Mean Absolute Error (MAE) loss as the optimization objective for prediction:

$$\mathcal{L}_P = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{24}$$

where N is the number of samples, y_i denotes the ground-truth multimodal sentiment score, and $\hat{y}_i$ denotes the predicted multimodal sentiment score.

Overall Learning Objective

Our proposed HCRL-MFD framework is trained end-to-end with the following multi-task loss:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_D + \mathcal{L}_{CL} + \mathcal{L}_P \tag{25}$$

where $\mathcal{L}_D$, $\mathcal{L}_{CL}$, and $\mathcal{L}_P$ represent the feature decomposition loss, hierarchical contrastive loss, and multimodal prediction loss, respectively.

4 Experimental

4.1 Datasets

We evaluated our approach on two widely adopted benchmarks for multimodal sentiment analysis: CMU-MOSI [49] and CMU-MOSEI [50]. Table 1 summarizes their key statistics.

Table 1: CMU-MOSI and CMU-MOSEI statistics.

Dataset	Train	Valid	Test
CMU-MOSI	1284	229	686
CMU-MOSEI	16326	1871	4659
All	17610	2100	5345

The CMU-MOSI dataset comprises 93 YouTube videos featuring 89 speakers who discuss various English-language topics. Each video is transcribed and annotated at the opinion level with sentiment scores ranging from –3 (strongly negative) to +3 (strongly positive).

CMU-MOSEI is currently the largest benchmark for multimodal sentiment and emotion recognition, containing 23,453 annotated clips from 1000 speakers across 250 topics sourced from platforms like YouTube. Every clip includes manual transcripts alongside visual and audio tracks, with sentiment and emotion labeled on a –3 to +3 polarity scale.

4.2 Evaluation Metrics

Following previous works [7,11,24,43], we report the experimental results of the model on regression and classification tasks separately.

Regression Metrics:

MAE: the average absolute difference between predicted and ground-truth sentiment scores; lower MAE denotes more accurate sentiment estimation.

Corr: measures the linear correlation between predicted and true scores, ranging from –1 to +1; values closer to ±1 indicate stronger predictive alignment.

Classification Metrics:

ACC-7: the proportion of correctly classified samples when sentiment scores (–3 to +3) are divided into seven discrete classes; higher ACC-7 reflects finer granularity performance.

ACC-2: the accuracy on a two-class split—negative (scores –3 to –1) vs. positive (+1 to +3); higher ACC-2 indicates superior polarity discrimination.

F1: the harmonic mean of precision and recall across classes, weighted by support; higher F1 captures balanced performance, especially under class imbalance.

Among these, ACC-2 and F1 are particularly indicative of a model's ability to distinguish overall sentiment polarity. In general, improved performance is reflected by higher ACC-2, F1, and Corr values, alongside a lower MAE.

4.3 Baseline

In order to fully validate the performance of our model, we compare our experimental results with the following state-of-the-art techniques in MSA tasks.

Self-MM [24] effectively improves the performance of multimodal learning by automatically generating stable single-peak labels and adjusting task weights based on label differences through a self-supervised multitask learning strategy.

MMIM [8] reduces information loss by hierarchically maximising mutual information between single-peak representations and fusion results, combined with neural network and Gaussian mixture model estimation.

HyCon [41] proposes a hybrid contrastive learning framework for three-modal representations, aiming to fully exploit cross-modal interactions, model inter-sample and inter-class relationships, and bridge the gap between different modalities.

ConFEDE [7] proposes to split each modality into similarity and difference features, construct cross-feature contrastives using the similarity features of the text as anchors, and use supplementary information from other modalities to improve prediction performance.

TETFN [12] is a text-enhanced Transformer fusion network that captures context information by pre-training visual features, LSTM with TCN, and utilises text-oriented multicentre attention with an inter-modal Transformer to achieve inter-modal consistency and dissimilarity retention.

SIMSUF [14] proposed a multimodal fusion method centred on dominant modal complementation, where the dominant modality is identified by estimating the interdependencies between modalities and complementing the other modalities with its features to enhance the representation.

DTN [28] unifies modal distributions and reduces redundant information by de-entangling cross-modal inputs and employing a two-step translation and relaxation reconstruction approach, while retaining task-relevant specific information.

CLGSI [43] fuses multimodal representations through sentiment intensity guided contrastive learning to select and weight positive and negative pairs, combined with fine-grained knowledge mechanisms to extract common and specific features.

DLF [31] mitigates modal interference by decomposing representations into modality-shared and modality-specific subspaces via geometric regularization, and employs a Language Focused Attractor to capture complementary non-verbal cues.

DEVA [51] enriches representations by translating audio-visual content into textual emotional descriptions via large multimodal models, integrating them with the original text through parameter-efficient fine-tuning and cross-modal attention.

TC-SIFN-MCM [52] tackles intra-modal noise using a Modality Calibrating Module, and employs a text-centric sparse interaction network to guide audio-visual representations while reducing inter-modal redundancy.

TCTR [53] addresses the issue of missing modalities through a text-guided contrastive learning framework and a token-level reconstruction network, supplementing incomplete information in both feature and semantic spaces.

4.4 Experimental Setup

We implemented our model using the PyTorch framework on an NVIDIA RTX 4090 GPU, with CUDA 12.4, PyTorch 2.0.1, and Python 3.8. The network was optimized using the Adam optimizer [54]. To ensure fair comparisons with other baseline methods, we strictly followed the experimental settings used in the latest competitive and state-of-the-art approaches. Specifically, the initial learning rate for BERT was set to 1e–5. On the CMU-MOSI dataset, the batch size was set to 32 and the model was trained for 30 epochs; on the CMU-MOSEI dataset, the batch size was set to 64 and the model was trained for 50 epochs.

4.5 Performance Comparison

Our experimental results on the CMU-MOSI and CMU-MOSEI datasets, as shown in Table 2, demonstrate the effectiveness of the proposed HCRL-MFD framework in comparison with several mainstream baseline methods. Overall, HCRL-MFD consistently outperforms competing approaches across nearly all evaluation metrics. Specifically, we observe the following:

Table 2: Performance comparison on CMU-MOSI and CMU-MOSEI datasets. Metrics labeled ↑ indicate higher performance with higher values, and metrics labeled ↓ indicate better performance with lower values. The best values are marked in bold.

Models	CMU-MOSI					CMU-MOSEI				
	ACC-2↑	F1↑	ACC-7↑	MAE↓	Corr↑	ACC-2↑	F1↑	ACC-7↑	MAE↓	Corr↑
Self-MM (2021)	82.54/84.77	82.68/84.91	45.79	0.712	0.795	82.68/84.96	82.95/84.93	53.46	0.529	0.767
MMIM (2021)	84.14/86.06	84.00/85.98	46.65	0.700	0.800	82.24/85.57	82.66/85.94	54.24	0.526	0.772
HyCon (2022)	–/85.2	–/85.1	46.6	0.713	0.790	–/85.4	–/85.6	52.8	0.601	0.776
ConFEDE (2023)	84.17/85.52	84.13/85.52	42.27	0.742	0.784	81.65/85.82	82.17/85.83	54.86	0.522	0.780
TETFN (2023)	84.05/86.10	83.83/86.07	–	0.717	0.800	84.25/85.18	84.18/85.27	–	0.551	0.748
SIMSUF (2024)	–/86.08	–/85.98	45.72	0.709	0.802	–/86.23	–/86.12	53.68	0.529	0.772
DTN (2024)	–/86.2	–/86.2	48.1	0.714	0.807	–/86.3	–/86.3	52.5	0.579	**0.788**
CLGSI (2024)	83.97/86.43	83.63/86.25	47.96	0.703	0.790	84.01/86.32	84.21/86.18	54.56	0.532	0.763
DLF (2025)	–/85.06	–/85.04	47.08	0.731	0.781	–/85.42	–/85.27	53.90	0.536	0.764
DEVA (2025)	84.40/86.29	84.48/86.30	46.32	0.730	0.787	83.26/86.13	82.93/86.21	**55.32**	0.541	0.769
TCTR (2026)	83.75/86.11	84.04/86.31	44.53	0.751	0.785	84.01/85.77	84.41/85.77	53.88	0.532	0.766
TC-SIFN-MCM (2026)	83.47/85.31	83.42/85.31	45.42	0.717	0.796	83.42/86.07	83.70/85.96	53.34	0.542	0.765
Ours (HCRL-MFD)	**84.69/86.81**	**84.51/86.95**	**48.31**	**0.691**	**0.814**	**84.71/86.79**	**84.91/86.72**	54.96	**0.517**	0.771

First, compared to feature fusion-based MSA methods such as Self-MM, MMIM, and TETFN, HCRL-MFD significantly enhances multimodal representation by introducing a hierarchical contrastive representation learning mechanism that effectively optimizes the distribution of modality-specific feature spaces. Second, in comparison with existing contrastive learning-based MSA methods, HCRL-MFD

incorporates an multimodal feature decoupling strategy that provides more fine-grained multimodal representations. This enables the model to effectively disentangle sentiment-relevant core information from redundant or irrelevant components within the feature subspace.

Finally, when evaluated against the most recent state-of-the-art models from 2026, including TCTR (which utilizes a token-level reconstruction network) and TC-SIFN-MCM (which employs a modality calibrating module), our HCRL-MFD maintains highly competitive and often superior performance. For instance, our model achieves the lowest MAE and the highest correlation on CMU-MOSI, further validating that rigorously purifying and aligning multimodal features prior to fusion yields more robust sentiment recognition accuracy than complex late-fusion calibration.

4.6 Ablation Experiments and Analyses

In this section, we conduct ablation studies to verify the effectiveness and importance of the proposed modules, including the Multimodal Feature Disentanglement (MFD) module and the Hierarchical Contrastive Representation Learning (HCRL) module. All ablation experiments are performed on the MOSEI dataset.

4.6.1 Impact of Key Modules

To validate the effectiveness of each module, we conducted experiments by removing the MFD module and the HCRL module, respectively. The results are shown in Table 3. Removing either module leads to a performance decline to varying degrees. Without the MFD module, the model's ability to learn multimodal representations significantly deteriorates due to the inability to effectively eliminate noise and irrelevant emotional information. Meanwhile, removing the HCRL module also causes a significant performance drop, further demonstrating the effectiveness of HCRL and indicating that imbalanced feature space distributions limit the model's performance.

Table 3: Ablation results of key modules.

Model	ACC-2	ACC-7	MAE	Corr
w/o MFD	81.64/83.12	49.98	0.566	0.737
w/o HCRL	82.14/84.04	51.82	0.552	0.743
Ours	**84.71/86.79**	**54.96**	**0.517**	**0.771**

Note: The best results of the complete model are marked in bold.

4.6.2 Impact of Multimodal Feature Decomposition on Different Modalities

To verify the effectiveness of the MFD module on each modality, we conducted experiments with feature disentanglement applied to only a single modality and to only two modalities, respectively. The experimental results are shown in Table 4. Compared to disentangling features in just one modality, performing disentanglement on two modalities significantly improved the model's performance. Additionally, we found that models applying MFD to the auxiliary modalities achieved performance closest to that of the full HCRL-MFD framework. This indicates that, compared to the textual modality, auxiliary modalities contain more redundant information unrelated to sentiment. Since the textual modality inherently carries abundant direct sentiment information, the model can still effectively extract certain features even without feature disentanglement. This further confirms the dominant role of the textual modality in multimodal sentiment analysis (MSA).

Table 4: Ablation results with different modality combinations.

Method	ACC-7	ACC-2	F1	MAE
Only T	51.84	82.12/84.76	82.26/84.79	0.557
Only V	53.01	83.14/85.41	83.27/85.36	0.546
Only A	52.85	83.13/85.21	83.08/85.08	0.541
T & V	53.45	83.72/85.76	83.77/85.87	0.538
T & A	53.47	83.81/85.85	83.85/85.86	0.531
V & A	54.11	84.21/86.13	84.15/86.08	0.527
Ours	**54.96**	**84.71/86.79**	**84.91/86.72**	**0.517**

Note: The best results of the complete model are marked in bold.

4.6.3 Impact of Different Decoupling Objectives

To verify the necessity of our specifically designed decoupling objectives, we conducted an ablation study by removing each constraint separately. The results are shown in Table 5. When the Reconstruction Loss $\mathcal{L}_R$ is removed, the model's performance significantly declines. This indicates that without this primary spatial geometric constraint to ensure information conservation, the feature decomposition process suffers from severe information loss, leading the encoders to arbitrarily discard input signals rather than strictly separating them. Similarly, removing the Orthogonal Loss $\mathcal{L}_O$ causes the core-sentiment and redundant features to remain entangled in the latent space. Without this spatial constraint to guarantee mutual independence, the features cannot be well separated, thereby degrading the purity of the decoupled emotional representations. Finally, although $\mathcal{L}_R$ and $\mathcal{L}_O$ theoretically establish a valid orthogonal feature space, removing the Unimodal Prediction Loss $\mathcal{L}_U$ removes the crucial task-driven semantic constraint. Consequently, the decoupled features lack clear emotional orientation, resulting in insufficient optimization of the emotional expression for each modality and leading to further performance drops.

Table 5: Ablation results of different decoupling objectives.

Method	ACC-7	ACC-2	F1	MAE
w/o $\mathcal{L}_R$	52.54	82.81/84.51	82.92/84.52	0.541
w/o $\mathcal{L}_O$	52.76	83.04/84.99	83.13/84.93	0.545
w/o $\mathcal{L}_U$	53.31	83.16/85.12	83.29/5.09	0.538
Ours	**54.96**	**84.71/86.79**	**84.91/86.72**	**0.517**

Note: The best results of the complete model are marked in bold.

4.6.4 Impact of Hierarchical Contrastive Representation Learning

To verify the effectiveness of the hierarchical contrastive representation learning, we conducted ablation experiments by removing the hierarchical contrastive learning, intra-modal contrastive learning, and inter-modal contrastive learning modules. The experimental results are shown in Table 6. "w/o CL" denotes the removal of the contrastive learning module, "intra" indicates using only intra-modal contrastive learning, and "inter" indicates using only inter-modal contrastive learning. The results demonstrate that all variants perform worse than the original model. Intra-modal contrastive learning improves the purity of auxiliary modal information by optimizing the feature space within each modality, showing better performance on

binary classification tasks. In contrast, inter-modal contrastive learning achieves better results on multi-class classification tasks, indicating that cross-modal interactions can more effectively integrate multi-source sentiment cues, thereby enhancing the fine-grained recognition of complex emotional states.

Table 6: Ablation results of hierarchical contrastive representation learning (HCRL).

Method	ACC-7	ACC-2	F1	MAE
w/o CL	51.82	82.14/84.04	82.28/84.11	0.552
Inter	53.58	82.74/84.96	82.81/84.83	0.531
Intra	52.97	83.13/85.69	83.04/85.66	0.537
Ours	**54.96**	**84.71/86.79**	**84.91/86.72**	**0.517**

Note: The best results of the complete model are marked in bold.

4.6.5 Impact of Contrastive Pair Selection Size

To justify our selection strategy for constructing contrastive pairs, we conduct an ablation study on the hyperparameter k, which dictates the number of top-k most similar and bottom-k least similar pairs selected. Specifically, we evaluate the impact of different selection sizes by varying $k \in \{1, 2, 3, 5\}$ within our HCRL-MFD framework. The detailed results are presented in Table 7.

Table 7: Ablation results of different contrastive pair selection sizes (k).

Size	ACC-7	ACC-2	F1	MAE
$k = 1$	52.91	82.94/85.65	82.97/85.47	0.541
$k = 2$ (Ours)	**54.96**	**84.71/86.79**	**84.91/86.72**	**0.517**
$k = 3$	54.07	84.13/86.69	84.34/86.51	0.527
$k = 5$	51.83	81.08/84.11	81.04/84.06	0.558

Note: The best results of the complete model are marked in bold.

As observed in Table 7, setting $k = 2$ achieves the optimal performance across all evaluation metrics. When $k = 1$, the contrastive signals are excessively sparse, failing to provide sufficient guidance for effective representation clustering and resulting in suboptimal alignment. Conversely, as k increases to 3 or 5, the model inevitably begins to incorporate pairs with marginal similarity or dissimilarity. This introduces detrimental noise into the contrastive objective, misguiding the representation learning process and leading to a continuous degradation in performance. Consequently, $k = 2$ serves as the optimal threshold, striking the best balance between providing abundant contrastive signals and filtering out less-informative, noisy pairs.

4.6.6 Impact of Different Fusion Mechanisms

To empirically justify our design choice of employing a simple concatenation followed by an MLP for multimodal fusion, we conduct an ablation study comparing it against several widely-adopted advanced fusion mechanisms. Specifically, we replace our concatenation layer with Tensor Fusion Network (TFN), Low-rank Multimodal Fusion (LMF), and Cross-Attention, while keeping the upstream representation learning modules unchanged. The detailed results are presented in Table 8.

As observed in Table 8, substituting the simple concatenation with complex advanced fusion mechanisms does not yield performance improvements; rather, it leads to a noticeable degradation across all evaluation metrics. We attribute this phenomenon to the robust representational power of our upstream

modules. Through Multimodal Feature Decomposition (MFD) and Hierarchical Contrastive Representation Learning (HCRL), the multimodal features have already been rigorously purified and strictly aligned within a unified semantic space. Consequently, a straightforward concatenation is highly capable of decoding the sentiment signals. Forcing these already well-aligned representations through parameter-heavy blocks like Cross-Attention or TFN introduces redundant computational complexity and increases the risk of overfitting, thereby deteriorating the final predictive performance.

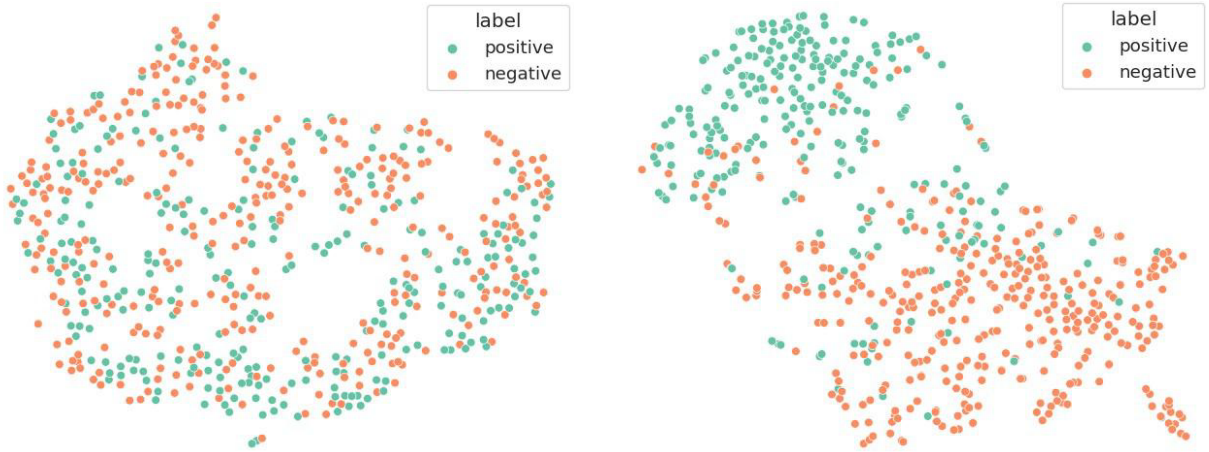
Table 8: Ablation results of different fusion mechanisms.

Method	ACC-7	ACC-2	F1	MAE
TFN	51.15	81.53/83.82	81.44/83.75	0.552
LMF	52.23	81.98/85.30	82.26/85.63	0.543
Cross-Attention	53.79	82.92/86.07	83.26/86.23	0.535
Concatenation (Ours)	**54.96**	**84.71/86.79**	**84.91/86.72**	**0.517**

Note: The best results of the complete model are marked in bold.

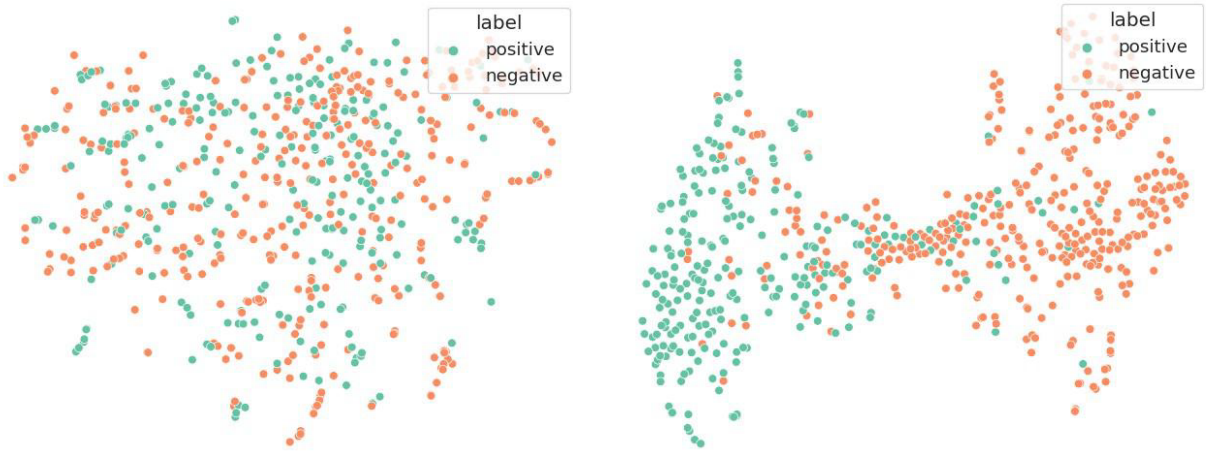
4.7 Visualisation

To further verify the effectiveness of HCRL-MFD, we conducted visualization analyses on audio and visual features from the CMU-MOSI test set, as shown in Figs. 3 and 4. Specifically, Figs. 3a and 4a respectively display the distribution of basic features for the audio and visual modalities, while Figs. 3b and 4b correspondingly show the changes in the distribution of features from the two modalities after processing with HCRL-MFD. It can be observed from the figures that the original audio and visual features are in a highly scattered state, with significant overlap between different emotion categories and a lack of clear clustering boundaries. In contrast, after applying HCRL-MFD, redundant information within modalities is effectively eliminated, and each emotion category gradually forms clustering structures with clearer boundaries and more distinct separation in the feature space. This further verifies the effectiveness of our method in enhancing the discriminability of multimodal representations.



(a) Basic audio feature distribution **(b)** Feature distribution after HCRL-MFD processing

Figure 3: T-SNE visualization of audio feature space.



(a) Basic vision feature distribution **(b)** Feature distribution after HCRL-MFD processing

Figure 4: T-SNE visualization of vision feature space.

4.8 Computational Efficiency Analysis

To evaluate the practical deployment feasibility of our proposed model, we analyze its computational complexity in terms of the number of trainable parameters (Params) and the average inference time per sample.

A key architectural advantage of our HCRL-MFD framework lies in its structural asymmetry between the training and inference phases. The Hierarchical Contrastive Representation Learning (HCRL) module, which computes similarities across massive sample pairs to refine the representation space, is strictly a training-time regularization mechanism. During the inference phase, this entire contrastive module is safely decoupled and discarded. The model only executes the standard feature extraction and the lightweight Multimodal Feature Decoupling (MFD) modules.

As demonstrated in Table 9, our proposed HCRL-MFD framework contains approximately 124.4 M parameters. Compared to the standard BERT-base backbone (~110 M), our framework introduces only a marginal increase in parameters (primarily from the decoupled projection layers and sentiment classifiers), while achieving significant performance improvements. Furthermore, evaluated on the complete test set, the average inference latency is 265.3 ms per sample. This indicates that our model maintains high computational efficiency and avoids the heavy computational overhead typically associated with complex fusion mechanisms, making it well-suited for practical multimodal sentiment analysis applications.

Table 9: Computational efficiency analysis of the proposed framework during the inference phase.

Model	Params (M)	Inference Time (ms/sample)
Standard BERT-base	~ 110.0	–
HCRL-MFD (Ours)	124.4	265.3

4.9 Case Study and Error Analysis

To provide a more comprehensive and intuitive evaluation of the proposed HCRL-MFD framework, we conduct a qualitative case study comparing typical successful predictions and a failure instance from the

CMU-MOSI test set. The visual representations of the selected cases, including their text contents, ground truths, and model predictions, are illustrated in Fig. 5.

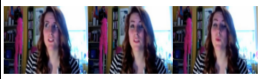
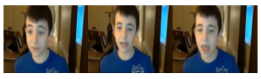
Cases	Sample1	Sample2	Sample3
Vision			
Text	I REALLY THINK THAT I REALLY LIKE THE MIRANDAS CHARACTER	AND WHICH HURTS THE FILM DRASTICALLY	HI IM PRETTY I HAVE A GIANT SMILE IM SUPPOSED TO KNOW THINGSUM WALK OF SCREEN
Audio			
Ground Truth	1.75	-2.2	-0.8
Output	1.9	-2.3	2.4

Figure 5: Qualitative case study of the proposed framework, including two typical success cases (Cases 1 & 2) and a failure case involving complex pragmatics (Case 3).

Analysis of Success Cases (Cases 1 & 2): For typical expressions where the sentiment is consistent across modalities, our framework demonstrates exceptionally accurate predictions. As shown in the success cases of Fig. 5, the literal meaning of the text strongly aligns with the acoustic and visual cues. Our Multimodal Feature Decomposition (MFD) module successfully purifies the core-sentiment features by explicitly stripping away emotion-irrelevant background noise. Subsequently, the Hierarchical Contrastive Representation Learning (HCRL) actively aligns these consistent features in the latent space. Consequently, the model accurately captures both strong positive (+1.90) and strong negative (−2.30) polarities with minimal deviation from the ground truth.

Error Analysis of the Failure Case (Case 3): While highly effective for typical expressions, our framework encounters limitations when dealing with complex pragmatics such as sarcasm. In Case 3 of Fig. 5, the text contains explicitly positive sentiment words (*"pretty", "giant smile"*). However, the true sentiment is strongly negative (−0.80), conveyed entirely through the mocking acoustic tone and dismissive visual gestures. In this scenario, the modalities are *intentionally* misaligned to create a rhetorical sarcastic effect.

Because our framework is explicitly constrained to extract consistent features across modalities and orthogonalize conflicting signals as "emotion-irrelevant noise", it misinterprets this deliberate semantic contradiction. The rigid spatial decoupling constraint forces the model to treat the crucial mocking tone in the audio and vision as modality-specific redundant noise because it sharply contradicts the strong positive semantics of the text. By suppressing these "noisy" but essential pragmatic cues, the model relies heavily on the literal text, outputting an incorrect highly positive score (+2.40).

This qualitative finding highlights an important boundary condition: While strict feature alignment and decoupling are highly effective for reducing actual noise, they face challenges when modalities interact in an adversarial manner. Future work will explore conflict-aware dynamic routing mechanisms to selectively preserve critical contradictory signals in such complex pragmatic contexts.

5 Conclusions and Future Work

In this paper, we proposed a novel multimodal sentiment analysis framework termed HCRL-MFD, which leverages Multimodal Feature Decoupling (MFD) to guide Hierarchical Contrastive Representation Learning (HCRL). Its primary goal is to simultaneously suppress emotion-irrelevant interference within auxiliary modalities and alleviate the uneven feature-space distributions caused by cross-modal heterogeneity. Extensive experiments on the CMU-MOSI and CMU-MOSEI benchmark datasets confirm the effectiveness and superiority of our approach over state-of-the-art baselines.

Despite its competitive performance, our framework has several limitations that point to important directions for future work. First, our method currently assumes the presence of complete modalities; in real-world scenarios where modality dropout occurs, the alignment in both feature decoupling and contrastive learning mechanisms may be severely compromised. Second, qualitative analysis of failure cases indicates that the explicit decoupling module struggles under extreme conditions. For instance, when audio or visual signals are heavily corrupted by severe environmental noise, or when the sentiment expression relies on implicit sarcasm (where linguistic semantics completely contradict the acoustic and visual cues), the decoupling encoders may fail to accurately isolate the core-sentiment features. Enhancing the robustness of the decoupling module against such extreme noise and complex linguistic phenomena remains a key priority.

Finally, while our Transformer-based modality encoders equipped with positional encodings effectively capture global temporal interactions, they currently lack explicit mechanisms for modeling strictly localized, fine-grained temporal continuity within the audio and visual streams. Exploring hybrid temporal modeling architectures to capture both global and local temporal dynamics will be another major focus of our future research.

Acknowledgement: This work was supported by the Nation Natural Science Foundation of China under Grant, the Yunnan Natural Science Funds under Grant.

Funding Statement: This work was supported by the Nation Natural Science Foundation of China under Grant 62466029, the Yunnan Natural Science Funds under Grant 202201AT070157.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Hongbin Wang and Di Jiang; methodology, Hongbin Wang; software, Hongbin Wang and Liusong Li; validation, Liusong Li; formal analysis, Hongbin Wang; investigation, Hongbin Wang and Liusong Li; resources, Hongbin Wang and Di Jiang; data curation, Liusong Li; writing—original draft preparation, Hongbin Wang; writing—review and editing, Di Jiang; visualization, Liusong Li; supervision, Di Jiang; project administration, Di Jiang; funding acquisition, Di Jiang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in a public repository.

Ethics Approval: Not applicable.

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

References

1. Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A. Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Inf Fusion. 2023;91:424–44.
2. Huan R, Zhong G, Chen P, Liang R. MulDeF: a model-agnostic debiasing framework for robust multimodal sentiment analysis. IEEE Trans Multimed. 2024;27:2304–19.

3. Zhang H, Wang Y, Yin G, Liu K, Liu Y, Yu T. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 756–67.
4. Angelou M, Solachidis V, Vretos N, Daras P. Graph-based multimodal fusion with metric learning for multimodal classification. Pattern Recognit. 2019;95(9):296–307. doi:10.1016/j.patcog.2019.06.013.
5. Wang Y, Qiu S, Li D, Du C, Lu BL, He H. Multi-modal domain adaptation variational autoencoder for EEG-based emotion recognition. IEEE/CAA J Autom Sin. 2022;9(9):1612–26. doi:10.1109/jas.2022.105515.
6. Zhao X, Li X, Jiang R, Tang B. Decoupled cross-attribute correlation network for multimodal sentiment analysis. Inf Fusion. 2025;117(5):102897. doi:10.1016/j.inffus.2024.102897.
7. Yang J, Yu Y, Niu D, Guo W, Xu Y. ConFEDE: contrastive feature decomposition for multimodal sentiment analysis. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, ON, Canada: Association for Computational Linguistics; 2023. p. 7617–30.
8. Han W, Chen H, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Moens MF, Huang X, Specia L, Yih SW, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 9180–92.
9. Lin R, Hu H. Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis. IEEE Trans Multimed. 2024;26(86):2740–55. doi:10.1109/tmm.2023.3303711.
10. Wang D, Liu S, Wang Q, Tian Y, He L, Gao X. Cross-modal enhancement network for multimodal sentiment analysis. IEEE Trans Multimed. 2023;25:4909–21.
11. Zhu C, Chen M, Zhang S, Sun C, Liang H, Liu Y, et al. SKEAFN: sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis. Inf Fusion. 2023;100:101958.
12. 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.
13. Fu Z, Liu F, Xu Q, Qi J, Fu X, Zhou A, et al. NHFNet: a non-homogeneous fusion network for multimodal sentiment analysis. In: Proceedings of the IEEE International Conference on Multimedia Expo (ICME); 2022 Jul 18–22; Taipei, Taiwan. p. 1–6.
14. Huang J, Ji Y, Qin Z, Yang Y, Shen HT. Dominant single-modal supplementary fusion (SIMSUF) for multimodal sentiment analysis. IEEE Trans Multimed. 2023;26(11):8383–94. doi:10.1109/tmm.2023.3344358.
15. Hu G, Lin TE, Zhao Y, Lu G, Wu Y, Li Y. UniMSE: towards unified multimodal sentiment analysis and emotion recognition. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 7837–51.
16. Gan C, Liu X, Tang Y, Yu X, Zhu Q, Jain DK. Enhanced multimodal sentiment analysis via integrated spatial position encoding and fusion embedding. Comput Mater Contin. 2025;85(3):5399. doi:10.32604/cmc.2025.068126.
17. Mai S, Hu H, Xing S. Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion. Proc AAAI Conf Artif Intell. 2020;34:164–72.
18. Li Z, Huang Z, Pan Y, Yu J, Liu W, Chen H, et al. Hierarchical denoising representation disentanglement and dual-channel cross-modal-context interaction for multimodal sentiment analysis. Expert Syst Appl. 2024;252(5):124236. doi:10.1016/j.eswa.2024.124236.
19. Lin R, Hu H. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 511–23.
20. Morency L, Mihalcea R, Doshi P. Towards multimodal sentiment analysis: harvesting opinions from the web. In: Proceedings of the 13th International Conference On Multimodal Interfaces; 2011 Nov 14–18; Alicante, Spain. p. 169–76.

21. Zhuang Y, Zhang Y, Hu Z, Zhang X, Deng J, Ren F. GLoMo: global-local modal fusion for multimodal sentiment analysis. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, VIC, Australia. p. 1800–9.
22. Wang W, Ding L, Shen L, Luo Y, Hu H, Tao D. Wisdom: improving multimodal sentiment analysis by fusing contextual world knowledge. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, VIC, Australia. p. 2282–91.
23. Tsai Y, Bai S, Liang P, Kolter J, Morency L, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. 6558 p.
24. Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proc AAAI Conf Artif Intell. 2021;35(12):10790–7. doi:10.1609/aaai.v35i12.17289.
25. Huang C, Chen J, Huang Q, Wang S, Tu Y, Huang X. AtCAF: attention-based causality-aware fusion network for multimodal sentiment analysis. Inf Fusion. 2025;114(6):102725. doi:10.1016/j.inffus.2024.102725.
26. Zhou M, Yang L, Wu T, Yang D, Zhang X. Dual-path dynamic fusion with learnable query for multimodal sentiment analysis. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics; 2025. p. 11355–65.
27. Tsai YHH, Liang PP, Zadeh A, Morency LP, Salakhutdinov R. Learning factorized multimodal representations. arXiv:180606176. 2018.
28. Zeng Y, Yan W, Mai S, Hu H. Disentanglement translation network for multimodal sentiment analysis. Inf Fusion. 2024;102(11):102031. doi:10.1016/j.inffus.2023.102031.
29. Hazarika D, Zimmermann R, Poria S. Misa: modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the MM '20: The 28th ACM International Conference on Multimedia; 2020 Oct 12–16; Seattle, WA, USA. p. 1122–31.
30. 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; 2022 Oct 10–14; Lisboa, Portugal. p. 1642–51.
31. Wang P, Zhou Q, Wu Y, Chen T, Hu J. DLF: disentangled-language-focused multimodal sentiment analysis. Proc AAAI Conf Artif Intell. 2025;39(20):21180–8. doi:10.1609/aaai.v39i20.35416.
32. Akbari H, Yuan L, Qian R, Chuang W, Chang S, Cui C, et al. Vatt: transformers for multimodal self-supervised learning from raw video, audio and text. Adv Neural Inf Process Syst. 2021;34:24206–21.
33. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. Int Conf Mach Learn. 2020;119:1597–607.
34. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 9729–38.
35. You Y, Chen T, Sui Y, Chen T, Wang Z, Shen Y. Graph contrastive learning with augmentations. Adv Neural Inf Process Syst. 2020;33:5812–23.
36. Tao L, Wang X, Yamasaki T. Self-supervised video representation learning using inter-intra contrastive framework. In: Proceedings of the 28th ACM international conference on multimedia. MM '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 2193–201.
37. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 8748–63.
38. Lin Z, Liang B, Long Y, Dang Y, Yang M, Zhang M, et al. Modeling intra-and inter-modal relations: hierarchical graph contrastive learning for multimodal sentiment analysis. In: Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea: International Committee on Computational Linguistics; 2022. p. 7124–35.
39. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. Adv Neural Inf Process Syst. 2020;33:18661–73.

40. Zha K, Cao P, Son J, Yang Y, Katabi D. Rank-n-contrast: learning continuous representations for regression. Adv Neural Inf Process Syst. 2023;36:17882–903. doi:10.52202/075280-0786.
41. Mai S, Zeng Y, Zheng S, Hu H. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Trans Affect Comput. 2022;14(3):2276–89. doi:10.1109/taffc.2022.3172360.
42. Yu Y, Zhao M, Qi S, Sun F, Wang B, Guo W, et al. ConKI: contrastive knowledge injection for multimodal sentiment analysis. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Findings of the association for computational linguistics: ACL 2023; 2023 Jul 9–14; Toronto, ON, Canada. p. 13610–24.
43. Yang Y, Dong X, Qiang Y. CLGSI: a multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. Mexico City, Mexico: Association for Computational Linguistics; 2024. p. 2099–110.
44. Liu R, Zuo H, Lian Z, Schuller BW, Li H. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. IEEE Trans Affect Comput. 2024;15(4):1856–73. doi:10.1109/taffc.2024.3378570.
45. Devlin J, Chang MW, Lee K, Tutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, MN, USA: Association for Computational Linguistics; 2019.
46. Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP—a collaborative voice analysis repository for speech technologies. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP); 2014 May 4–9; Florence, Italy. p. 960–4.
47. Baltrušaitis T, Robinson P, Morency LP. Openface: an open source facial behavior analysis toolkit. In: Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV); 2016 Mar 7–10; Lake Placid, NY, USA. p. 1–10.
48. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30. doi:10.65215/ctdc8e75.
49. Zadeh A, Zellers R, Pincus E, Morency LP. Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell Syst. 2016;31(6):82–8. doi:10.1109/mis.2016.94.
50. Bagher Zadeh A, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Gurevych I, Miyao Y, editors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, ON, Australia: Association for Computational Linguistics; 2018. p. 2236–46.
51. Wu S, He D, Wang X, Wang L, Dang J. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. Proc AAAI Conf Artif Intell. 2025;39(2):1601–9. doi:10.1609/aaai.v39i2.32152.
52. Zhou H, Liu J, Li X, Liu Y, He H. Text-centric sparse interaction fusion network with a modality calibrating module for multimodal sentiment analysis. IEEE Trans Affect Comput. 2026. doi:10.1109/TAFFC.2026.3658336.
53. Yang Z, He Q, Yu M, Du N, Lu Y. TCTR: text-guided contrastive learning with token-level reconstruction network for missing modalities in multimodal sentiment analysis. Inf Fusion. 2026;126:103571.
54. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv:1412.6980. 2014.