Feature Decomposition-Augmentation Network for Multimodal Sentiment Analysis

Dapeng Yin [1,2], Bingxuan Hou [1,2], Mengna Gao [1,2], Shuyue Zhu [1,2]
Junli Wang [1,2]*

[1] Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education, Shanghai 201804, China.
[2] National (Province-Ministry Joint) Collaborative Innovation Center for Financial Network Security, Tongji University, Shanghai 201804, China.
{2432122, 2432023, 2432121, 2432272}@tongji.edu.cn
{junliwang}@tongji.edu.cn

Abstract

Multimodal sentiment analysis identifies human emotional tendencies by analyzing text, visual, and auditory modalities. In most studies, the textual modality is usually considered to contain the most emotional information and is regarded as the dominant modality. Existing methods mostly map auxiliary modalities into a semantic space close to the dominant modality, which overly relies on the dominant modality. In this work, we propose a Feature Decomposition-Augmentation (FeaDA) framework, which aims to elevate the role of auxiliary modalities in multimodal data fusion. We first design a projector to decompose auxiliary modalities into partial features, which contain features for emotion judgment, and then utilize these decomposed features to guide the fusion process with KL loss, thereby enhancing the status of auxiliary modality fusion. To verify the effectiveness of our method, we conducted experiments on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets. The experimental results show that our FeaDA framework outperforms mutilmodal sentiment analysis methods of the same type in main metrics. Our code is available at https://github.com/PowerLittleYin/FeaDA-main.

1 Introduction

With the advancement of multimodal learning technologies, substantial progress has been made in multimodal representation understanding (Fukui et al., 2016; Tan and Bansal, 2019; Radford et al., 2021a) and multimodal fusion (Liu et al., 2018; Mai et al., 2020). Modern social-media platforms like TikTok and Twitter have multiplied the ways we express emotion: a single post and its replies can now weave together videos and text into a single, hybrid utterance, leading to a growing interest

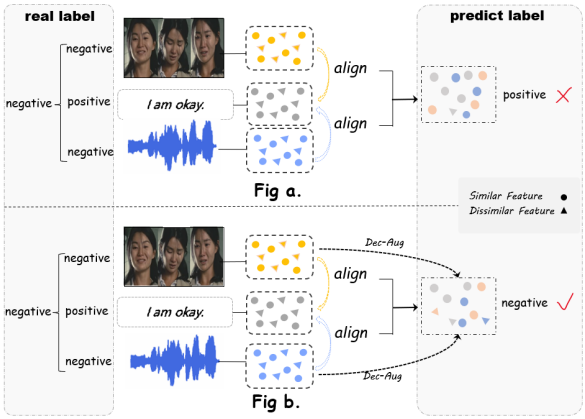

Figure 1: Previous Method **a** vs. Our Method **b**

in multimodal sentiment analysis (Tsai et al., 2019; Hazarika et al., 2020; Wu et al., 2024; Zhang et al., 2025). Compared with single-modal approaches, multimodal methods can capture emotional tendencies more comprehensively and dramatically improve the accuracy of sentiment analysis. Consequently, multimodal sentiment analysis has become a critical tool for understanding multimedia content. It is also widely applied in real-world scenarios such as election polling, public-opinion monitoring of news events, and customer satisfaction evaluation for products and services.

Current multimodal sentiment analysis primarily focuses on the analysis of three modalities: text, video, and audio. Drawing on previous research (He et al., 2025; Chen et al., 2025), text is usually identified as the main modality in multimodal sentiment analysis, as it generally contains a richer array of emotional information. Meanwhile, video and audio are regarded as auxiliary modalities: although they typically carry less rich emotional information than text, they can still provide valuable cues for determining sentiment orientation. Recent research has mainly focused on two directions—multimodal data fusion and multimodal feature understanding. The former is pur-

*Corresponding author. This work was supported by the National Key Research and Development Program of China under Grant 2022YFB4501704.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 86–98

sued through sophisticated attention-based fusion networks (Tsai et al., 2019; Zhang et al., 2022; Sun and Tian, 2025), while the latter has recently leveraged contrastive learning to devise specialized alignment networks that enhance the comprehension of multimodal features, such as ConFEDE (Yang et al., 2023) and QUEST (Song et al., 2025) .

There is an inherent flaw in multimodal learning, when jointly optimizing heterogeneous features, the model tends to concentrate its capacity on the primary modality while under-utilising or even discarding cues from auxiliary modalities (Fan et al., 2023; Lu et al., 2024; Zong et al., 2024; Yang et al., 2025). Consequently, the fused representation may only partially capture the subject's affective state, resulting in incomplete or erroneous sentiment predictions. As illustrated in Figure 1, when cross-modal semantics are inconsistent, accurate inference becomes unattainable unless the fine-grained signals residing in the auxiliary streams are sufficiently amplified. In response to this limitation, recent studies have dedicated considerable effort to feature-level understanding (Wang et al., 2023b; Li and Liu, 2025). However, most of them still condition the learning process on the dominant modality, leaving auxiliary cues vulnerable to suppression during fusion. Thus, when semantic inconsistency arises, task-critical features in the subordinate modality remain under-represented. Despite the maturity of contemporary multimodal architectures, a systematic approach to auxiliary-modality feature augmentation is still missing.

Therefore, we propose a novel framework for enhancing the features of the auxiliary modalities based on feature decomposition——**Fea**ture **D**ecomposition-**A**ugmentation (**FeaDA**). The main contributions of our framework are as follows:

- We propose an effective feature enhancement framework. For the auxiliary modalities, we introduce a feature decomposition module that can extract the parts of the auxiliary modalities that contain more emotional information.
- During the interaction between the main and auxiliary modalities, we employ an efficient feature augmentation method to compensate for the weaker position of the auxiliary modalities in multimodal interactions.
- Our proposed method achieves overall excellent performance on the public datasets CMU-MOSI, CMU-MOSEI, and CH-SIMS, demonstrating the effectiveness of our approach.

2 Related Work

2.1 Multimodal Sentiment Analysis

In recent years, with the development of multimodal learning, researchers have increasingly diversified their studies on multimodal sentiment analysis. Initially, methods for fusing multimodal sentiment features were explored. Early researchers employed straightforward techniques such as direct concatenation and weighted summation of features from different modalities to achieve preliminary multimodal feature fusion (Boulahia et al., 2021; Tsai et al., 2019). More recently, fusion methods utilizing transformer architectures or attention mechanisms have been proposed (Zhang et al., 2022; Sun and Tian, 2025).

Another significant area of research is multimodal representation understanding. The contrastive feature decomposition framework (Yang et al., 2023) leverages contrastive learning mechanisms to decompose features from each modality into similarity features and dissimilarity features. Gradient modulation in multimodal sentiment analysis is another important direction to understand representation. To address the issue of modality imbalance, researchers have proposed a series of gradient modulation-based methods. For instance, classifier-guided gradient modulation (Peng et al., 2022; Guo et al., 2025) introduces a classifier to evaluate the utilization of each modality and adaptively adjusts the gradient magnitude of the encoder based on this assessment.

2.2 Multimodal Learning with Contrastive Learning

Multimodal learning aims to integrate data from multiple modalities to achieve a more comprehensive understanding and representation of features. Contrastive learning has recently demonstrated significant potential in multimodal learning. In 2021, the CLIP model (Radford et al., 2021b) proposed by OpenAI marked an important milestone in the field of multimodal contrastive learning. The contrastive learning objective employed by CLIP enables learning from weakly supervised web-scale data with only pair-wise relationships but no explicit labels. The same year, a global-local representations framework (Ma et al., 2021) is proposed, which based on global and local perspec-

tives, where the model can learn spatially localized correspondences between audio and visual signals through contrastive learning. Unlike existing contrastive learning methods that focus on maximizing the mutual information between two views while ignoring unique information, the QUEST (Song et al., 2025) framework advances towards learning more disentangled representations, where shared and unique factors are effectively separated. Similarly, ConFEDE (Yang et al., 2023) introduced a contrastive feature decomposition framework that also aims to disentangle similar and dissimilar features across modalities while utilizing multiple aspects of features from each modality.

Despite learning more comprehensive features, these methods still face the modality-bias problem in multimodal sentiment analysis. This paper propose a decomposition-augmentation method to alleviate the modality-bias issues caused by over-reliance on dominant modalities.

3 Methodology

In this section, we present the overall architecture of the proposed FeaDA, with the pipeline of FeaDA illustrated in Figure 2.

3.1 Feature Embedding

Let the input be denoted as $Input = [T, V, A]$, where $T \in R^{S_t \times d_t}, V \in R^{S_v \times d_v}, A \in R^{S_a \times d_a}$. S denotes the sequence length per modality, and d the feature dimensionality. After encoding, we obtain the unified feature representation $F = [f^t, f^v, f^a]$. For the textual modality, we employ BERT as the encoder. For vision and audio modality, we follow the same feature extraction as previous work (Tsai et al., 2019). Then, we can obtain f^t, f^v, f^a as follow:

$$f^t = BERT(T), \tag{1}$$

$$f^v = VEncoder(V), \tag{2}$$

$$f^a = AEncoder(A). \tag{3}$$

3.2 Feature Association with Prompt

Since the encoded modalities mentioned above lack interaction with each other, in order to effectively determine which parts of the features from each modality are more important for judging emotional tendency, this paper employs cross-modal attention to generate interaction information between different modalities. Next, we will introduce the role of the Feaure Association module shown in Figure 2.

We observe that, in other multimodal feature-understanding tasks (Zhou et al., 2022; Liang et al., 2024), soft prompts have been employed to enhance image understanding. In order to make the modality features interact more fully, we add the auxiliary modality features to the learnable soft prompt P^m initialized to all zeros, where $m \in \{t, v, a\}$, representing the text, video and audio modalities respectively. The resrepresentation can be described as follow:

$$f_v^{'} = [f^v, P^v], \tag{4}$$

$$f_a^{'} = [f^a, P^a], \tag{5}$$

where $[\cdot]$ denotes vector concatenation. We employ cross-modal attention to generate interaction information between different modalities. Let the weight matrices of each modality be W^m. Let the product of the dominant modality feature and its corresponding weight matrix, that is, $f^t \cdot W^t$ as the Query Q. Let the products of the auxiliary modality features and their corresponding weight matrices, that is the, audio feature $f_a^{'} \cdot W^a$ be the Key K^a and Value V^a respectively. Taking the interaction between the text and audio modalities as an example:

$$f^{at} = softmax(\frac{QK^a}{\sqrt{d_k}})V^a. \tag{6}$$

where d_k represents the dimension of the Key K^a.

3.3 Feature Decomposition-Augmentation

After the the feature association stage, although the textual modality has interacted separately with the audio and visual modalities, the dominant role of the textual modality may lead to an over-reliance on textual features in the resulting fused representations. This dominance can consequently cause information loss in the visual and audio modalities during cross-modal interactions. To address this issue, we propose a Decomposition-Augmentation mechanism for the post-interaction text-visual feature f^{vt} and text-audio feature f^{at}, aiming to mitigate the imbalance in cross-modal representation.

3.3.1 Modality Selective Decomposition

Traditional approaches typically align visual details with textual attributes to generate pseudo-representations for feature enhancement (Zhao et al., 2025). However, such methods fail to capture the dynamic temporal patterns in videos or spectral

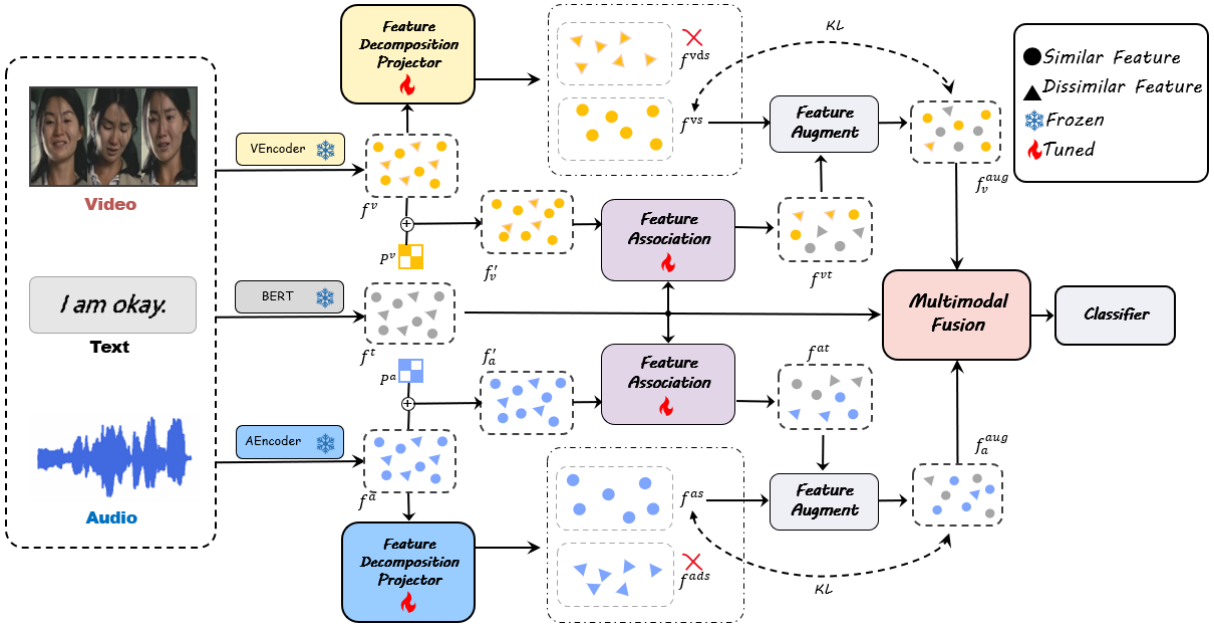

Figure 2: The architecture of FeaDA. Initially, features from each modality are extracted from the input, which follows the previous work (Tsai et al., 2019). First, vision and audio features are concatenated with the soft prompt, the correlated features are generated with the Feature Association module. Then, the features of the video and audio modalities are decomposed with the Feature Decomposition Projector to select features for the next augmentation state. Following this, the decomposed features are utilized to enhance the correlated text-video/audio features with Feature Augment module. Finally, the enhanced text-video/audio features are concatenated with the textual features for the ultimate sentiment classification.

nuances in audio, as these modalities cannot be sufficiently summarized by static textual descriptors. We consider the method of feature reuse, but not all features are what we need, and reusing features that do not carry sentiment information will obviously increase noise, which will hinder our analysis of sentiment.

Inspired by ConFEDE (Yang et al., 2023), we leverage contrastive learning's discriminative power to address this limitation. In our proposed method, we decompose features into similarity-preserving and dissimilarity-preserving components. Unlike ConFEDE, our decomposition is modality-selective: we exclusively apply it to visual and audio features, retaining only the similarity-preserving subset. This design is motivated by following consideration: prior work (Xu et al., 2017) demonstrates that dissimilarity components can degrade model performance by introducing noisy or conflicting information. Our retention of similarity features ensures cleaner cross-modal interactions while maintaining discriminative ability.

Next, we introduce the components of feature decomposition, taking visual features as an example; audio features can be handled in a similar way. The encoded video feature f^v is processed through a modality-specific projector, which decomposes it into similarity-preserving features f^{vs} and dissimilarity-preserving features f^{vds}. The projector consists of layer normalization for feature stabilization, a linear layer with Tanh activation and dropout for regularization. To train the projector, we employ a dual-loss objective comprising: unimodal prediction loss $\mathcal{L}_{uni}$ and contrastive decomposition loss $\mathcal{L}_{con}$.

Unimodal Loss The unimodal loss ensures each modality independently extracts its emotion-discriminative features, preserving modality-specific emotional cues, so that capture modality-unique emotional patterns. As previously described, contrastive decomposition loss enforces the anchor-based decomposition to make sure that features aligned with the textual anchor are attracted and dissimilar features are repelled. When predicting the unimodal label $\hat{u}$, we concatenate auxiliary modalities' the similar and dissimilar features and pass them through an MLP with a ReLU activation function as the classifier. Then, the unimodal prediction loss $\mathcal{L}_{uni}$ is computed using the Mean Squared Error (MSE). The specific process

is as follows:

$$\hat{u} = MLP(f^t, [f^{vs}, f^{as}], [f^{vds}, f^{ads}]), \quad (7)$$

$$u = [y_m, y_m, y_m, y_t, y_v, y_a], \quad (8)$$

$$\mathcal{L}_{uni} = MSE(\hat{u}, u), \quad (9)$$

where y_m represents the ground-truth multimodal label, while y_t, y_v, and y_a denote the ground-truth unimodal labels for text, vision, and audio modalities respectively. $MSE(\cdot)$ stands for calculating the Mean Squared Error loss.

Contrastive Decomposition Loss Unlike traditional contrastive learning methods (Radford et al., 2021b), this paper adopts a feature pair perspective (Yang et al., 2023) and designs Algorithm 1 to compute this loss. To divide positive and negative feature pairs, for any sample i, other samples will be selected to construct two sets according to the cosine similarity with i: a similar sample set Nei^i and an outlier sample set Out^i. The concrete computation is given in Lines 1–5 of Algorithm 1.

We divide positive and negative feature pairs from two perspectives. First, given a single sample, in Lines 6-9 of Algorithm 1: treat similar features across different modalities as intra-sample positive feature pairs P_0^i; we treat similar versus dissimilar features within the same modality or across modalities as intra-sample negative pairs N_0^i. Second, between samples, in Lines 10-12 of Algorithm 1: if sample j comes from the Nei^i, we treat the similar features in the same modality between sample (i, j) as inter-sample positive pairs P_1^i; if sample j comes from the Out^i, we treat the similar features in the same modality between sample (i, j) as inter-sample negative pairs N_1^i. We then merge these sets to obtain the positive and negative feature pairs respectively:

$$P^i = P_0^i \cup P_1^i, \quad (10)$$

$$N^i = N_0^i \cup N_1^i. \quad (11)$$

Subsequently, we calculate the contrastive loss $\mathcal{L}_{con}^i$ for sample i through the NT-Xent contrastive learning framework:

$$\ell_{con}^i = \sum_{(a,p)\in P^i} -log \frac{exp(\frac{cos(a,p)}{\tau})}{\sum_{(a,k)\in P^i \cup N^i} exp(\frac{cos(a,k)}{\tau})}, \quad (12)$$

here, (a, p) and (a, k) denote a pair of decomposed feature vectors, $cos(\cdot)$ denotes the calculation of cosine similarity. Therefore, the contrastive decomposition loss $\mathcal{L}_{con}$ can be expressed as:

$$\mathcal{L}_{con} = \frac{1}{n} \sum_{i=1}^{n} \ell_{con}^i, \quad (13)$$

where n is the number of samples in a batch.

3.3.2 Feature Independence Augmentation

After the video and audio features have undergone feature decomposition, the trained projector outputs the feature components f^{vs} and f^{as}. Next, we use these two components to enhance the text-video feature f^{tv} and the text-audio feature f^{ta} independence to overcome reliance on the textual modality after the feature association stage.

To enhance the features, we first directly augment the features after the feature association stage using the similar features. We notice a star operation (Ma et al., 2024), which is essentially matrix element-wise product, can capture more subtle data differences. We reduce the star operation from a matrix operation to a vector one, which is the dot product of vectors, to augment the common features in the auxiliary modality and those after the feature association stage. For instance, to get the augmented text-visual feature f_v^{aug} :

$$f_v^{aug} = f^{vt} \odot f^{vs}. \quad (14)$$

In the second phase, we address the weaker role of the auxiliary modality during the feature association stage by using KL Loss. To improve the auxiliary modality's interaction with the dominant modality, we employ KL loss to guide the interaction between the dominant and auxiliary modalities with the auxiliary modality's similar features:

$$\mathcal{L}_v = KL(f^{vs}, f_v^{aug}), \quad (15)$$

$$\mathcal{L}_a = KL(f^{as}, f_a^{aug}), \quad (16)$$

where $KL(\cdot)$ represents the calculation of KL loss.

3.4 Final loss function

We use an MLP with three layers as the classifier. The predicted output $\hat{y}$ and the prediction loss are expressed as follows:

$$\hat{y} = MLP(f^t, f_v^{aug}, f_a^{aug}). \quad (17)$$

The predict loss is:

$$\mathcal{L}_{pred} = MSE(y, \hat{y}). \quad (18)$$

Algorithm 1: Feature Divide Algorithm

Input: Dataset D, multimodal features f^t, f^v, f^a, multimodal labels y^m.
Output: positive pairs P^i, negative pairs N^i

for $i \in D$ **do**
  $S_0^i \leftarrow$ samples that share the same y^m as sample i, and sorted by cosine similarity;
  $S_1^i \leftarrow$ samples that have different y^m with sample i, and sorted by cosine similarity;
  $Nei^i \leftarrow$ Randomly select samples with high cosine similarity from the samples in S_0^i;
  $Out^i \leftarrow$ Randomly select samples from samples with high similarity in S_1^i and samples with low similarity in S_1^i ;
for $i, j \in D$ **do**
  if $i == j$ **then**
    $P_0^i \leftarrow$ Similar features across different modalities;
    $N_0^i \leftarrow$ Dissimilar features of both same modality and different modalities;
  else
    $P_1^i \leftarrow$ The same modality features across different samples i, j, where$j \in Nei^i$;
    $N_1^i \leftarrow$ The same modality features across different samples i, j, where $j \in Out^i$;
  $P^i = P_0^i \cup P_1^i$;
  $N^i = N_0^i \cup N_1^i$;

The final loss $\mathcal{L}$ can be expressed as:

$$\mathcal{L} = \mathcal{L}_{pred} + \alpha\mathcal{L}_{uni} + \beta\mathcal{L}_{con} + \gamma(\mathcal{L}_a + \mathcal{L}_v), \quad (19)$$

where α, β and γ are pre-defined hyperparameters. The training objective is to minimize the final loss $\mathcal{L}$.

4 Experiments

In this section, we introduce the details of our experiments, including the datasets, evaluation metrics, baselines and relevant settings.

4.1 Datasets

In our work, we conducted evaluations on three publicly available multimodal sentiment analysis datasets: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020). The information of the datasets is listed in Table 1.

	train	validation	test
CMU-MOSI	1284	229	686
CMU-MOSEI	16326	1871	4659
CH-SIMS	1368	456	457

Table 1: Dataset Information

4.2 Experimental Setting

To demonstrate the performance of our FeaDA network, we select state-of-the-art baselines from recent multimodal sentiment analysis research for comparison. TFN (Poria et al., 2017) , LMF (Liu et al., 2018) , MulT (Tsai et al., 2019) , MISA (Hazarika et al., 2020) , Self-MM (Yu et al., 2021) , FDMER (Yang et al., 2022) , ConFEDE (Yang et al., 2023) , SFTTR (Sun and Tian, 2025) . The first four baselines are early fusion methods, and the last four are methods of the same kind as ours, employing representation-learning approaches that decompose features, which makes the effectiveness of our method even more directly evident. Additionally, to investigate the differences from text-centric approaches, we also included ALMT (Zhang et al., 2023) and TETFN (Wang et al., 2023a) for comparison.

All experiments were conducted on a single Tesla V100-SXM2 GPU, to mitigate device-related discrepancies, we reproduced Self-MM, ConFEDE, and SFTTR on our hardware following the original authors' replication guidelines. For CMU-MOSEI and CMU-MOSI, we follow Tsai et al. (2019) to extract features. For CH-SIMS, we follow Yu et al. (2020) to extract features. We employ a two-stage training pipeline: in the first stage we train the encoders, and in the second we frozen the encoders and train our multimodal framework. In CMU-MOSI and CMU-MOSEI we use BERT as the text encoder, and in CH-SIMS we use BERT-Chinese as the text encoder. For vision and audio encoder, we use transformer encoders. In the second stage,we train our FeaDA with the above trained encoders. Through multiple experiments, we set α = 0.02, β = 0.03 and γ = 0.01, the procedure for selecting the

Model	CMU-MOSI					CMU-MOSEI				
	Acc-2 ↑	Acc-7 ↑	F1 ↑	MAE ↓	Corr ↑	Acc-2 ↑	Acc-7 ↑	F1 ↑	MAE ↓	Corr ↑
TFN†	-/80.8	34.9	-/80.7	0.907	0.698	78.50/81.89	51.60	78.96/81.74	0.573	0.714
LMF†	-/82.5	33.2	-/82.4	0.917	0.695	80.54/83.48	51.59	80.94/83.36	0.576	0.717
Mult†	-/83.0	40.0	-/82.8	0.871	0.698	81.15/84.63	82.84	81.56/84.52	0.559	0.733
MISA†	81.8/83.4	42.3	81.7/83.6	0.783	0.776	83.6/85.5	52.2	83.8/85.3	0.555	0.756
Self-MM	81.38/83.87	**46.51**	81.57/84.12	0.713	0.790	82.89/85.0	52.17	82.78/84.57	**0.528**	0.755
FDMER†	-/84.6	44.1	-/84.7	0.724	0.788	-/86.1	**54.1**	-/85.8	0.536	0.741
TETFN	83.24/85.37	-	83.01/85.33	**0.708**	**0.798**	84.12/85.18	-	84.18/85.27	0.551	0.748
ALMT	83.08/85.41	-	83.11/85.42	0.722	0.791	**84.66**/84.49	-	**85.13**/85.16	0.609	0.776
ConFEDE	81.53/85.06	44.75	83.61/85.09	0.726	0.795	81.13/85.69	52.35	81.68/85.67	0.539	0.769
SFTTR	82.94/84.60	46.50	82.92/84.63	0.709	0.795	81.69/**86.16**	53.53	82.25/**86.18**	0.530	**0.776**
FeaDA	**83.97/86.13**	44.31	**83.90/86.12**	0.723	0.796	84.25/85.47	53.49	84.22/85.16	0.548	0.771

Table 2: Results on CMU-MOSI and CMU-MOSEI. † means the result from Sun and Tian (2025). "Acc" is Accuracy, "F1" is F1 Score. The "/" in Acc-2 (Zadeh et al., 2017) corresponds to "negative/non-negative" and the "/" in F1 (Tsai et al., 2019) corresponds to "negative/positive", "-" indicates that the method does not take this metric into account. "MAE" is Mean Absolute Error, "Corr" is Pearson Correlation.

Model	CH-SIMS				
	Acc-2 ↑	Acc-3 ↑	F1 ↑	MAE ↓	Corr ↑
TFN†	78.38	65.12	78.62	0.432	0.591
LMF†	77.77	64.68	77.88	0.441	0.576
MulT†	78.56	65.12	79.66	0.453	0.561
MISA†	76.54	-	76.59	0.447	0.563
Self-MM	78.07	65.08	78.27	0.431	0.601
ConFEDE	78.34	67.83	78.72	0.395	0.640
SFTTR	78.56	**68.49**	78.82	**0.380**	**0.645**
FeaDA	**79.43**	65.43	**79.43**	0.417	0.595

Table 3: Results on CH-SIMS.

hyperparameters is provided in the appendix A.5.

4.3 Result Comparison

Table 2 and Table 3 compare our method with the state-of-the-art approaches on three datasets. We highlight the best-performing values in bold.

Table 2 reports our results on CMU-MOSI and CMU-MOSEI. On both classification and regression metrics, our method is either superior to or on par with most baselines. For coarse-grained classification (Acc-2 and F1) we achieve the best performance: In CMU-MOSI, Acc-2 improves the SFTTR by about 1.5%, and F1 by about 1.6%, demonstrating our model's stronger understanding of sentiment polarity. On fine-grained metrics (Acc-7) and regression metrics we do not outperform the baseline, but the performance remains broadly comparable.

Table 3 also presents results on the more challenging CH-SIMS dataset, where the overall trend is consistent with MOSI and MOSEI. For coarse-grained classification we again outperform best, confirming that our method can effectively extract affective cues and grasp sentiment polarity across diverse scenarios. Although we lag slightly on fine-grained and regression metrics, the overall performance remains excellent.

4.4 Ablation Study and Error Analysis

We conducted five ablation studies on CMU-MOSI to validate the contribution of each component and to explain why some metrics fall short of expectations. The results are reported in Table 4, where "-uni" removes the unimodal loss, "-con" removes the contrastive-decomposition loss, "-KL" removes the KL loss, "-P" removes the soft prompts. Consistent with the main experiments, our full model achieves the best Acc-2 and F1 scores.

Notably, when the soft prompts are removed, Corr reach the highest values, and Acc-7 also shows a certain degree of improvement. We hypothesize that soft prompts tend to over-emphasize similar features during training, thereby suppressing fine-grained cues, which explains the observed boost in fine-grained and regression metrics once they are ablated. Similarly, removing the KL loss also yields better fine-grained classification and regression results, which aligns with our expectation. The KL loss is intended to let similarity features guide the auxiliary modality during fusion. While this strengthens the auxiliary modality's role, it simultaneously causes the loss of unique dissimilar features, which may in fact carry affective information. Since sentiment analysis is ultimately a classification-centric task, we nonetheless regard

the full model as the most effective configuration.

Model	CMU-MOSI				
	Acc-2 ↑	Acc-7 ↑	F1 ↑	MAE ↓	Corr ↑
-uni	83.67/85.98	43.00	83.57/85.95	0.735	0.794
-con	82.94/84.60	39.50	82.90/84.61	1.793	0.790
-KL	82.27/84.4	**50.98**	82.40/84.12	**0.559**	0.755
-P	83.82/85.82	48.10	83.76/85.83	0.694	**0.808**
FeaDA	**83.97/86.13**	44.31	**83.90/86.12**	0.723	0.796

Table 4: Ablation Study on CMU-MOSI.

4.5 Visualization

Figure 3 shows the 2D visualization of the features on the CH-SIMS dataset using the t-SNE method. We mainly focus on the distribution of the red, yellow, and blue points in the figure. Without any optimization method, it can be seen that the similar features of the three modalities still remain independently distributed and do not interact with each other. After applying our FeaDA method, the similar features of the three modalities become associated. Unlike ConFEDE (Yang et al., 2023) and SFTTR (Sun and Tian, 2025), whose interactions among similar features are highly intensive, our method maintains a certain degree of independence while allowing interaction. The comparisons with prior methods confirms that the auxiliary modality has indeed gained prominence without becoming overly reliant on the textual modality; the learned features preserve the desired independence during interaction. Similarity-driven features form a compact cluster, which aligns with our strong performance on the binary task.

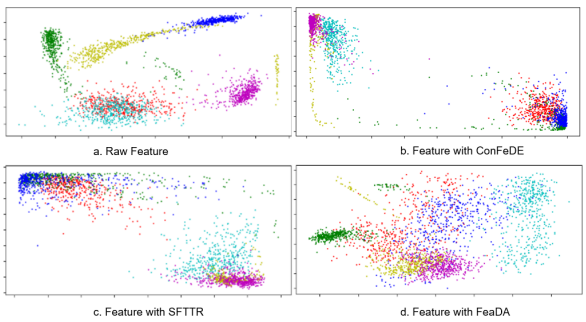

Figure 3: The figure a shows the visualization of feature distribution without optimization, the figure b shows the feature distribution using ConFEDE, the figure c shows the feature distribution using SFTTR, the figure d shows the feature distribution using FeaDA. The colors red, green, bule represent similar features of Text, Vision, Audio, and cyan, yellow and magenta represent dissimilar features of Text, Vision, Audio.

4.6 Discussion

Here we analyze why our model degrades on fine-grained classification and regression metrics. We argue there are two main reasons. According to the t-SNT visualization, the auxiliary modality is indeed strengthened and becomes less dependent on the dominant one. However, the features we exploit to enhance the auxiliary modality are similarity-based, which causes their weights to dominate the auxiliary feature space. In some scenarios, fine-grained classification and regression may hinge on features that carry distinctive emotional cues; under such conditions, our similarity-oriented feature augmentation inadvertently dilutes these unique signals, leading to performance decline.

Moreover, compared to SFTTR, our approach demonstrates a marked superiority in model complexity. While maintaining accuracy on par with, or slightly better than SFTTR, it reduces the number of trainable parameters by approximately 60%. Since SFTTR uses the intricate sequential cross modality fusion network during feature fusion, which incurs substantial computational overhead. In contrast, we forgo any elaborate fusion architecture in the final stage and instead employ a simple concatenation operation, yet still preserve the accuracy required for emotion analysis.

5 Conclusion

In this paper, we present FeaDA, a feature decomposition and augmentation framework for multimodal sentiment analysis. FeaDA comprises three collaborative components: a feature-decomposition projection module, a feature-fusion module, and a feature augmentation module. Together, they elevate the role of the auxiliary modality during fusion. At the final stage we deliberately adopt the simplest strategy, direct concatenation, to avoid overshadowing the augmentation effects with an overly sophisticated fusion network. Experimental results show that we effectively elevated the status of auxiliary modalities, mitigating modal bias in multimodal sentiment analysis to a certain extent. While the independence of auxiliary modalities is strengthened, their ability to discern sentiment polarity remains uncompromised. Future work will seek more appropriate features to reinforce the auxiliary modality, aiming to improve fine-grained sentiment analysis.

Limitations

Although FeaDA advances several metrics on mainstream benchmarks, two limitations remain. First, to suppress noise while strengthening the auxiliary modality, we rely on similarity-based features for augmentation; this inadvertently causes the model to overlook unique yet affective cues inherent in the auxiliary modality. Second, being grounded in contrastive learning, our approach is highly sensitive to batch size: larger batches yield better results but demand prohibitive hardware resources. The details between batch-size and metrics are shown in Appendix A.4.

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Said Yacine Boulahia, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121.

Miao Chen, Jin Liu, Xingye Li, Yaohui Zhang, Hongze Liu, Jiajia Jiao, and Huihua He. 2025. Dnmcn: Dual-stage normalization based modality-collaborative fusion network for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–17.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20029–20038.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas. Association for Computational Linguistics.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. 2025. Classifier-guided gradient modulation for enhanced multimodal learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Xiaojiang He, Yanjie Fang, Nan Xiang, Zuhe Li, Qiufeng Wang, Chenguang Yang, Hao Wang, and Yushan Pan. 2025. Text-guided multi-level interaction and multi-scale spatial-memory fusion for multimodal sentiment analysis. *Neurocomputing*, 626:129532.

Tianyi Li and Daming Liu. 2025. MPID: A modality-preserving and interaction-driven fusion network for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4313–4322, Abu Dhabi, UAE. Association for Computational Linguistics.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12373–12387, Bangkok, Thailand. Association for Computational Linguistics.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Qiwen Lu, Shengbo Chen, and Xiaoke Zhu. 2024. Collaborative modality fusion for mitigating language bias in visual question answering. *Journal of Imaging*, 10(3).

Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. 2021. Contrastive learning of global-local video representations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Xu Ma, Xiyang Dai, Yue Bai, Yizhou Wang, and Yun Fu. 2024. Rewrite the stars. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5694–5703.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):164–172.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8237.

Soujanya Poria, E. Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Qi Song, Tianxiang Gong, Shiqi Gao, Haoyi Zhou, and Jianxin Li. 2025. Quest: quadruple multimodal contrastive learning with constraints and self-penalization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Kaiwei Sun and Mi Tian. 2025. Sequential fusion of text-close and text-far representations for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 40–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. 2023a. Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136:109259.

Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023b. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 216–226, Cham. Springer Nature Switzerland.

Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. 2024. Multimodal multi-loss fusion network for sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3588–3602, Mexico City, Mexico. Association for Computational Linguistics.

Yong Xu, Qiang Huang, Wenwu Wang, Peter Foster, Siddharth Sigtia, Philip J. B. Jackson, and Mark D. Plumbley. 2017. Unsupervised feature learning based on deep models for environmental audio tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1230–1241.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1642–1651, New York, NY, USA. Association for Computing Machinery.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630, Toronto, Canada. Association for Computational Linguistics.

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. 2025. Learning to rebalance multimodal optimization by adaptively masking subnetworks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4553–4566.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767, Singapore. Association for Computational Linguistics.

Xiangmin Zhang, Wei Wei, and Shihao Zou. 2025. Modal feature optimization network with prompt for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4611–4621, Abu Dhabi, UAE. Association for Computational Linguistics.

Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 9100–9108. AAAI Press.

Zheyu Zhao, Zhongqing Wang, Shichen Li, Hongling Wang, and Guodong Zhou. 2025. Bridging modality gap for effective multimodal sentiment analysis in fashion-related social media. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1813–1823, Abu Dhabi, UAE. Association for Computational Linguistics.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *Int. J. Comput. Vision*, 130(9):2337–2348.

Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, and Ken Zheng. 2024. Balancing multimodal learning via online logit modulation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 5753–5761. International Joint Conferences on Artificial Intelligence Organization. Main Track.

A Appendix

A.1 Baseline Details

1)TFN (Poria et al., 2017) Jointly models intra-modality dynamics within each modality and inter-modality dynamics across modalities, enabling a richer understanding of a speaker's affective orientation in videos.

2)LMF (Liu et al., 2018) Factorizes the high-order weight tensor into modality-specific low-rank factors, eliminating the need to explicitly construct high-dimensional tensors and thus mitigating the exponential growth in computational cost and parameter count that plagues tensor-based multimodal fusion.

3)MulT (Tsai et al., 2019) Introduces a novel cross-modal attention mechanism that can directly process unaligned multimodal language sequences.

4)MISA (Hazarika et al., 2020) Decomposes each modality's features into modality-invariant and modality-specific subspace representations, then fuses these representations for prediction.

5)Self-MM (Yu et al., 2021) Leverages self-supervision to automatically generate unimodal labels, guiding the model to learn both inter-modal consistency and inter-modal differences without any human annotation cost.

6)FEMER (Yang et al., 2022) Explicitly disentangles each modality's representation into shared and private components, combining adversarial training with attentive fusion to address the heterogeneity challenge in multimodal emotion recognition.

7)ConFEDE (Yang et al., 2023) Unifies contrastive learning to decompose each modality's features into similar and dissimilar parts, using the text modality's similar features as anchors to jointly perform cross-sample contrastive learning and within-sample modality decomposition.

8)SFTTR (Sun and Tian, 2025) Adopts a contrastive decomposition and sequential fusion strategy that thoroughly excavates the commonalities and discrepancies between textual and audio/visual modalities.

9)ALMT (Zhang et al., 2023) Introduces an adaptive late-fusion mechanism with modality-specific transformers that dynamically re-weights textual and acoustic/visual streams, enabling fine-grained alignment of cross-modal emotional cues while preserving unimodal discriminability.

10)TETFN (Wang et al., 2023a) Devises a temporal-enhanced Transformer fusion network that hierarchically models text–audio–visual interactions via time-aware cross-modal attention, jointly optimizing synchronous alignment and asynchronous discrepancy mining for robust multimodal sentiment detection.

A.2 Datasets Details

CMU-MOSI (Zadeh et al., 2016).

This is an English-language dataset, comprising a collection of 2199 opinion video clips. The dataset is divided into 1284 training samples, 229 vali-

dation samples, and 686 testing samples. Each opinion video is annotated with sentiment values within the range of [-3, 3]. Lower annotation values indicate more negative sentiment, while higher values indicate more positive sentiment. The dataset is meticulously annotated by humans, including labels for subjectivity, sentiment intensity, visual features annotated on a per-frame and per-opinion basis, and audio features annotated on a per-millisecond basis.

CMU-MOSEI (Bagher Zadeh et al., 2018).

This dataset can be considered an extended version of CMU-MOSI. The CMU-MOSEI dataset is larger in scale, containing 22,852 annotated video clips from 1,000 different speakers. It covers a broader range of topics and emotional expressions. The dataset is divided into 16,326 training samples, 1,871 validation samples, and 4,659 testing samples.

CH-SIMS (Yu et al., 2020).

This is a Chinese-language dataset, consisting of 60 original videos and 2,281 video segments. Each video segment is annotated with sentiment values within the range of [-1, 1], and is accompanied by unimodal sentiment labels. Lower annotation values indicate more negative sentiment, while higher values indicate more positive sentiment. The dataset is divided into 1,368 training samples, 456 validation samples, and 457 testing samples.

A.3 Evaluation Metrics

Since sentiment data is inherently both continuous and discrete in nature, and the complexity of multimodal data as well as the demands of practical applications require models to handle both sentiment categories and sentiment intensity simultaneously. Therefore, it is necessary to evaluate the performance of multimodal sentiment analysis on both classification and regression. For CMU-MOSI and CMU-MOSEI, in the classification, we select the accuracy of 2-class prediction(Acc-2), 7-class prediction(Acc-7), and F1-score as the classification evaluation metrics. For CH-SIMS, in the classification, we select the accuracy of 2-class prediction(Acc-2), 3-class prediction(Acc-3), and F1-score as the classification evaluation metrics. To maintain consistency with prior work, we calculate Acc-2 and F1-Score in two ways for CMU-MOSI and CMU-MOSEI. For Acc-2 (Zadeh et al., 2017), we set negative/non-negative evaluations. For F1-Score (Tsai et al., 2019), we set negative/positive evaluations. For regression, we evaluate using Mean Absolute Error (MAE) and Pearson correlation (Corr).

A.4 Analysis of Batch Size

As we employed contrastive learning in the feature decomposition module, this section examines how batch size affects model performance. We select batch sizes of {2, 4, 8, 16, 32} to observe the trends of various metrics under different settings. Figure A4 illustrates the trends of all metrics across different batch sizes. It can be observed that performance improves as the batch size increases.

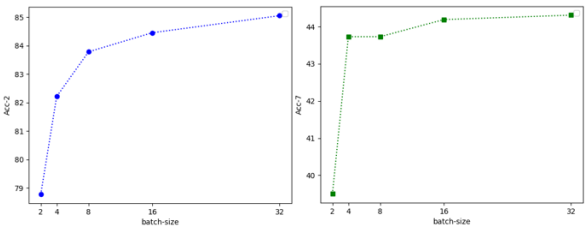

Figure A4: The trends of each metric under varying batch sizes, blue denotes Acc-2, and green denotes Acc-7.

A.5 Hyperparameter Setting

We set x $\in$ {0.01, 0.02, 0.03, 0.04, 0.05, 0.07, 0.09}, the results are shown in Figure A5. Notably, the choice of hyperparameter has a significant impact on model performance. When testing α, we fix β and γ; likewise, when testing β and γ, we fix the other two hyperparameters, respectively. We select the value of each hyperparameter that yields the best F1 score, therefore, we select α = 0.02, β = 0.03, and γ = 0.01.

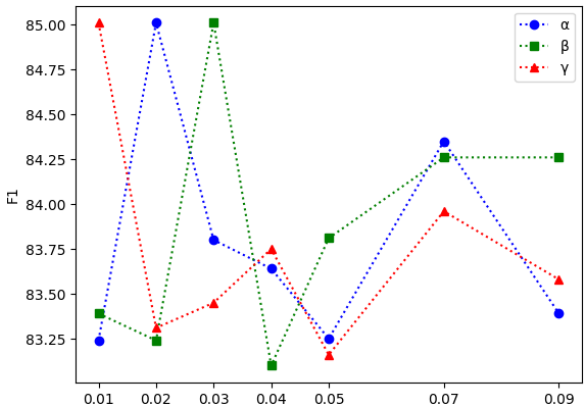

Figure A5: Experimental results with different hyperparameter settings. The x-axis represents the values of the hyperparameter, and the y-axis denotes the average F1 score.