DLF: Disentangled-Language-Focused Multimodal Sentiment Analysis

Pan Wang[1*], **Qiang Zhou**[1], **Yawen Wu**[1], **Tianlong Chen**[2], **Jingtong Hu**[1]

[1]University of Pittsburgh
[2]UNC Chapel Hill
{pan.wang, qiang.zhou, yawen.wu, jthu}@pitt.edu, tianlong@cs.unc.edu

Abstract

Multimodal Sentiment Analysis (MSA) leverages heterogeneous modalities, such as language, vision, and audio, to enhance the understanding of human sentiment. While existing models often focus on extracting shared information across modalities or directly fusing heterogeneous modalities, such approaches can introduce redundancy and conflicts due to equal treatment of all modalities and the mutual transfer of information between modality pairs. To address these issues, we propose a Disentangled-Language-Focused (DLF) multimodal representation learning framework, which incorporates a feature disentanglement module to separate modality-shared and modality-specific information. To further reduce redundancy and enhance language-targeted features, four geometric measures are introduced to refine the disentanglement process. A Language-Focused Attractor (LFA) is further developed to strengthen language representation by leveraging complementary modality-specific information through a language-guided cross-attention mechanism. The framework also employs hierarchical predictions to improve overall accuracy. Extensive experiments on two popular MSA datasets, CMU-MOSI and CMU-MOSEI, demonstrate the significant performance gains achieved by the proposed DLF framework. Comprehensive ablation studies further validate the effectiveness of the feature disentanglement module, language-focused attractor, and hierarchical predictions.

Code — https://github.com/pwang322/DLF.

Introduction

With the rapid development of social media, multimodal interaction has become increasingly popular, which attracts many researchers to transfer uni-modal learning to multi-modal learning tasks (Awal et al. 2024; Guan et al. 2024; Xu, Zhu, and Clifton 2023). One of the most significant subfields is multimodal sentiment analysis (MSA) (Geetha et al. 2024; Yang et al. 2022a). MSA aims to perceive human sentiment through multiple heterogeneous modalities, such as language, vision, and audio, playing a crucial role in many applications including cognitive psychology, scenario understanding, and mental health (Ali and Hughes 2023; Ezzameli and Mahersia 2023; Yang et al. 2023a). Compared with unimodal solutions, MSA often presents a more robust performance by leveraging complementary information from different modalities. How to effectively learn essential representations without redundant and conflicting information from multiple heterogeneous modalities, however, remains an open question in the academic field, especially in multimodal learning communities.

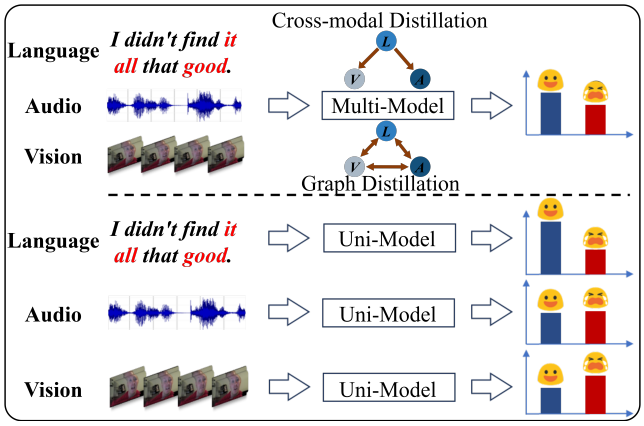

Figure 1: Task pipeline of the Multimodal Sentiment Analysis, and varied performance of different modalities.

In recent years, researchers have shown an increased interest in MSA. Many multimodal models have been proposed to facilitate MSA, and they can be categorized into two groups: representation learning-oriented methods (Guo et al. 2022; Hazarika, Zimmermann, and Poria 2020; Sun et al. 2023; Yang et al. 2022c) and multimodal fusion-oriented methods (Zhang et al. 2023; Huang et al. 2020; Yang et al. 2022b; Tsai et al. 2019; Lv et al. 2021; Rahman et al. 2020). The former primarily aims to acquire an advanced semantic understanding of various modalities enriched with diverse clues of human sentiments, resulting in more powerful human sentiment encoders. Conversely, the latter emphasizes designing sophisticated fusion strategies at various levels, including feature-level, decision-level, and model-level fusion, to derive unified representations from multimodal data. It is worth noting that the fundamental aspect of MSA lies in learning and integrating multimodal representations, where the goal is to accurately process and in-

*Corresponding author
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tegrate various modal inputs to discern sentiments from the underlying data. Although current leading methods in MSA (Hazarika, Zimmermann, and Poria 2020; Tsai et al. 2019; Zadeh et al. 2017; Yu et al. 2021) have shown considerable progress, the inherent disparities across diverse modalities continue to present challenges, complicating the development of stable and effective multimodal representation. For MSA, as shown in Figure 1, existing works and our ablation study (see Table 2) have shown that language, vision, and audio sources contribute differently to the overall prediction performance (Pham et al. 2019; Tsai et al. 2019; Kim and Park 2023; Li et al. 2024a; Lei et al. 2023), which indicates the big distribution gap among different modalities hinders the final performance.

To mitigate the distribution gap among heterogeneous modalities, as shown in Figure 1, knowledge distillation-based methods, such as cross-modal and graph distillation, are introduced to transfer reliable information between different modalities (Gupta, Hoffman, and Malik 2016; Guo et al. 2020; Aslam et al. 2023; Kim and Kang 2022; Hazarika, Zimmermann, and Poria 2020; Li, Wang, and Cui 2023; Tsai et al. 2019). Cross-modal distillation typically leverages the stronger *Language* modality to teach weaker modalities (*Vision and Audio*), while graph distillation completely performs bidirectional information transfer between all modality pairs. However, it is important to note that conventional distillation is inherently asymmetric—it is effective when transferring information from one modality to another, but the benefits of the reciprocal process are unclear. This asymmetry can lead to redundant or even conflicting information in cross-modal and graph distillation, ultimately limiting overall performance.

To this end, we critically reconsider the characteristics illustrated in Figure 1: ***Why focus solely on bridging the gap between different modalities, rather than strategically enhancing the strengths of the dominant one?*** However, directly enhancing the dominant modality while treating all modalities equally and employing bidirectional information transfer across all modality pairs often introduces redundancy and conflicts (Hazarika, Zimmermann, and Poria 2020), thereby reducing overall performance. In contrast, our work strategically leverages a pivotal characteristic of MSA: language has been empirically recognized as the dominant modality (Tsai et al. 2019). Building on this insight, we intend to develop a novel Language-Focused Attractor (LFA), a targeted enhancement scheme designed to transfer complementary information exclusively to the dominant language modality, which consolidates information through pathways such as Video $\rightarrow$ Language, Audio $\rightarrow$ Language, and Language $\rightarrow$ Language, resulting in effectively minimizing redundancy and conflicting information and improving overall MSA accuracy.

To achieve this, we propose a Disentangled-Language-Focused (DLF) multimodal representation learning framework to fully exploit the potential of language-dominant MSA. The framework follows a structured pipeline: feature extraction, disentanglement, enhancement, fusion, and prediction. To specifically address the issues of redundancy and conflicting information to facilitate language-targeted feature enhancement, DLF introduces four geometric measures as regularization terms in the total loss function, effectively refining shared and specific spaces both separately and jointly. Within the modality-specific space, we further develop the LFA to enhance language representation by attracting complementary information from other modalities. This process is guided by a *Language-Query*-based multimodal cross-attention mechanism, ensuring precise and targeted feature enhancement between heterogeneous modality pairs ($X\rightarrow$Language, where X refers to Language, Video, or Audio). Finally, the enhanced shared and specific features are fused, followed by hierarchical predictions to further improve overall prediction accuracy.

Our main contributions can be summarized as follows:

- **Proposed Framework:** In this study, we propose a Disentangled-Language-Focused (DLF) multimodal representation learning framework to promote MSA tasks. The DLF framework presents a structured pipeline: feature extraction, disentanglement, enhancement, fusion, and prediction.
- **Language-Focused Attractor (LFA):** We develop the LFA to fully harness the potential of the dominant language modality within the modality-specific space. The LFA exploits the language-guided multimodal cross-attention mechanisms to achieve a targeted feature enhancement ($X\rightarrow$Language).
- **Hierarchical Predictions:** We devise hierarchical predictions to leverage the pre-fused and post-fused features, improving the total MSA accuracy. Comprehensive ablation studies further validate the effectiveness of each component in the DLF framework.

Related Work

Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) integrates information from diverse modalities, such as language, video, and audio (Ali and Hughes 2023; Ezzameli and Mahersia 2023). Mainstream methods can be categorized into representation learning-oriented (Guo et al. 2022; Sun et al. 2023; Yang et al. 2022c) and fusion-oriented approaches (Zhang et al. 2023; Huang et al. 2020; Tsai et al. 2019). Representation methods like (Guo et al. 2022), enhance cross-modal interactions, while fusion techniques, including Transformer-based model (Huang et al. 2020), focus on combining features effectively. Despite progress, performance disparities among modalities hinder the overall prediction accuracy. To this end, distillation-based strategies were introduced to bridge the gap. For instance, Kim and Kang (2022) proposed cross-modal distillation between textual and auditory modalities to enhance emotion classification granularity. To dynamically adapt to distillation, ternary-symmetric architectures (MulT) (Tsai et al. 2019) or graph distillation units (Zadeh et al. 2018b; Li, Wang, and Cui 2023) were introduced, representing modalities as vertices and their interactions as edges. Recent approaches also leverage large multimodal language models for flexible interactions (Wu et al.

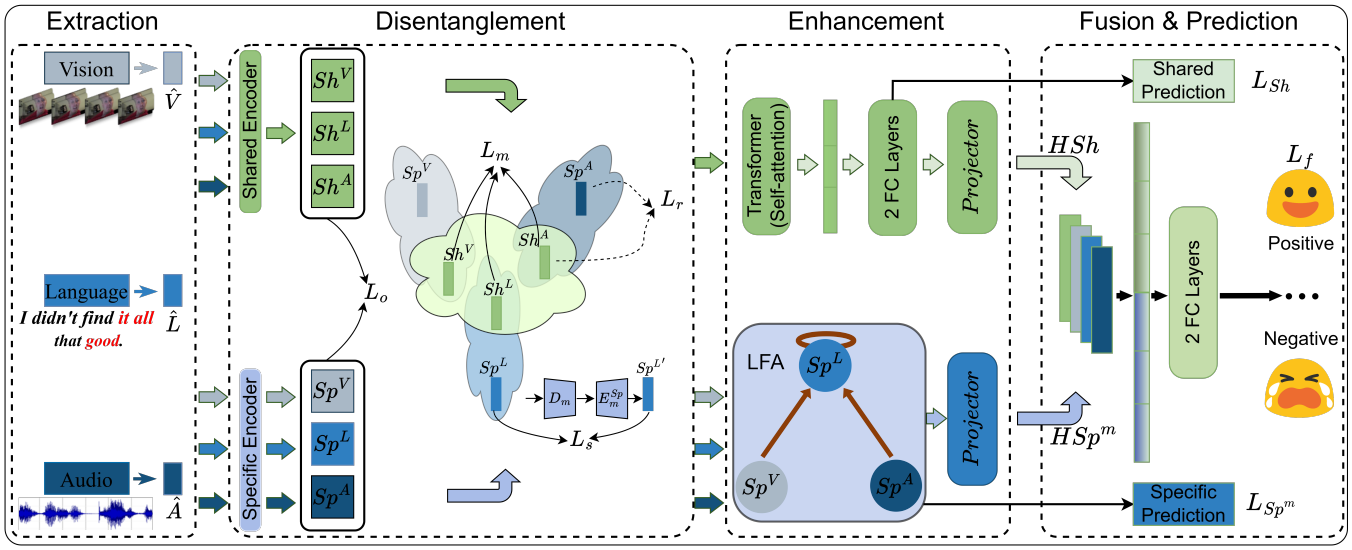

Figure 2: Overview of the proposed DLF framework. The framework follows a pipeline of feature extraction, disentanglement, enhancement, fusion, and prediction, featuring three core components: the feature disentanglement module, the Language-Focused Attractor (LFA), and hierarchical predictions (including shared prediction, specific prediction, and final prediction).

2023) and integrate contextual knowledge to boost predictions (Wang et al. 2024). However, previous paradigms treat different modalities equally, easily causing redundant and conflicting information. Our proposed Language-Focused Attractor (LFA) directs knowledge transfer to the dominant language, mitigating redundant and conflicting information and enhancing overall performance.

Disentangled Multimodal Representation Learning

Disentangled multimodal representation learning aims to separate distinct factors of variation across multiple modalities (e.g., vision, language, audio) into independent subspaces (Yang et al. 2023b; Li et al. 2024b; Yang et al. 2023a). Tsai et al. (2018) introduced the concept of factorized representations for multimodal data, designed to disentangle modality-specific and shared components, successfully isolating audio and visual features in audiovisual speech recognition. Building on this, Hazarika, Zimmermann, and Poria (2020) proposed the MISA framework, which projects each modality into common and private feature spaces, reducing inter-modality disparities while enhancing representation diversity. Further advancements including Yang et al. (2022a,c), employed metric learning and adversarial learning to construct modality-invariant and modality-specific subspaces, significantly improving multimodal fusion. Li, Wang, and Cui (2023) developed a decoupled multimodal distillation (DMD) approach to address distribution gaps between modalities. However, previous methods uniformly treat all modalities regarding the purpose of disentanglement. In our DLF, disentanglement aims to facilitate language-targeted feature enhancement in LFA, thus, we introduce four measures as regularization terms, carefully designed to reinforce disentanglement by jointly and independently optimizing shared and specific spaces.

Proposed Approach

Preliminaries. As shown in Figure 2, the task of MSA aims to predict the sentiment intensity or label of given multimodal inputs. In DLF, three modalities are concurrently considered, such as Language (L), Vision (V), and Audio (A), represented as 2D tensors $\hat{X}_m \in R^{N_m \times d_m}$, where N_m is the sequence length, d_m is the embedding dimension, and $m \in \{L, V, A\}$ means different modalities.

Overview

The framework of the proposed DLF is illustrated in Figure 2. It adopts a structured pipeline comprising feature extraction, disentanglement, enhancement, fusion, and prediction. The framework integrates three core components: the feature disentanglement module, the Language-Focused Attractor (LFA), and hierarchical predictions (shared, specific, and final predictions). DLF decomposes multimodal features into modality-shared and modality-specific spaces to minimize redundancy and conflicts among heterogeneous modalities. To reinforce this decoupling, four geometric measures are incorporated into the total loss as regularization terms. Additionally, the LFA is designed to leverage the dominant language modality by integrating complementary information from other modalities, thereby enhancing language representation. Finally, hierarchical predictions are performed to boost overall MSA performance. The details are as follows:

Feature Disentanglement Module

To reduce redundant and conflicting information, the proposed DLF framework utilizes a shared encoder and three modality-specific encoders to decompose multimodal information into modality-shared and modality-specific feature spaces, denoted as Sh^m and Sp^m, respectively, where

$m \in \{V, L, A\}$. Formally, the shared and specific encoders are defined as:

$$Sh^m = E_m^{Sh}(\hat{X}_m), \tag{1}$$

$$Sp^m = E_m^{Sp}(\hat{X}_m), \tag{2}$$

where E_m^{Sh} and E_m^{Sp} represent the shared and specific encoders, respectively. In this work, both encoders are implemented as cascaded Transformer layers.

For effective disentanglement, DLF incorporates the regularization effect of carefully designed regularization terms. While classical approaches often employ distribution similarity measures such as KL-Divergence along the hidden dimensions (Kim and Mnih 2018), we adopt four geometric measures based on Euclidean distances and cosine similarity due to their intuitive nature and computational efficiency.

After initial disentanglement, DLF concatenates Sh^m and Sp^m for each modality and reconstruct the multimodal input $\hat{X}_m$, resulting $\hat{X}'_m$ by decoding the fused features $[Sh^m \oplus Sp^m]$. This process can be formulated as follows:

$$\hat{X}'_m = D_m([Sh^m \oplus Sp^m]), \tag{3}$$

where D_m is the 1D convolution decoder, and $\oplus$ is the concatenation operation. The discrepancy between the low-level features $\hat{X}_m$ and the generated features $\hat{X}'_m$, called the reconstruction loss L_r, can serve as a regularization term contributing to the feature disentanglement module:

$$L_r = ||\hat{X}_m - \hat{X}'_m||^2. \tag{4}$$

Furthermore, the modality-specific reconstruction process can be formulated as:

$$Sp^{m\prime} = E_m^{Sp}(\hat{X}'_m), \tag{5}$$

where $Sp^{m\prime}$ is estimated modality-specific features from the modality-specific reconstruction process. Naturally, the discrepancy between original modality-specific features Sp^m and the estimated ones $Sp^{m\prime}$ can be regarded as the specific loss L_s:

$$L_s = ||Sp^m - Sp^{m\prime}||^2. \tag{6}$$

Although the reconstruction loss L_r and specific loss L_s contribute to the decoupling process, their effectiveness in achieving robust disentanglement remains limited. This limitation arises from the potential sub-optimal performance of the shared encoder during the initial training phase, especially when compared to the modality-specific encoder. Such a disparity, if left unaddressed, is exacerbated by these loss functions, causing an increasing divergence between the two encoders as training progresses. Therefore, we incorporate a modified triplet loss (Schroff, Kalenichenko, and Philbin 2015) to enhance the performance of the modality-shared encoder. The triplet loss is defined as:

$$L_m = \frac{1}{|T|} \max\left(0, d\left(S, P\right) - d\left(S, N\right) + \mu\right), \tag{7}$$

where S represents a sampled modality in the modality-shared space, P denotes the positive sample corresponding to the representation of the same sentiment across different modalities, and N refers to the negative sample representing distinct sentiments within the same modality. T is the total number of positive and negative samples, $d(\cdot, \cdot)$ computes the cosine similarity between two feature vectors, and μ represents a distance margin.

The aforementioned losses, L_r, L_s, and L_m, regulate the shared and specific features to ensure they focus on their respective objectives. To further refine the decoupling between these two spaces, a soft orthogonality loss, L_o, is introduced to minimize redundancy and conflicts between shared and specific multimodal features. It is defined as:

$$L_o = O(Sh^m, Sp^m), \tag{8}$$

where $O(\cdot, \cdot)$ represents a non-negative counterpart of cosine similarity, promoting orthogonality between the two feature spaces.

Eventually, the four geometric-measure-based regularization terms, addressing both inter- and intra-decoupled spaces, are combined to form the decoupling loss:

$$L_d = \sum_{k \in \{r,s,m,o\}} \lambda_k L_k, \tag{9}$$

where λ_k are weighting coefficients for the individual regularization terms, providing a flexible mechanism to balance their contributions and calibrate the model effectively.

Language-Focused Attractor (LFA)

Unlike conventional feature enhancement methods that aim to bridge modality gaps through cross-modal and graph distillation (Gupta, Hoffman, and Malik 2016; Guo et al. 2020; Aslam et al. 2023; Li, Wang, and Cui 2023), we propose the LFA in the modality-specific space after feature decoupling. The detailed structure of LFA is depicted in Figure 3.

In the LFA, decoupled modality-specific features Sp^m (where $m \in \{L, V, A\}$, representing language, vision, and audio) are first processed through positional embedding and dropout, then fed into Multimodal Transformer layers. The core operation within these layers is the Multimodal Cross-Attention (MCA) mechanism. LFA performs three branches of MCA, including one self-attention and two cross-attention mechanisms, all centered on the language modality as the *Query* (Q_L). This setup allows the language modality to attract complementary information from other modality-specific features Sp^m, where $m \in \{L, V, A\}$. The corresponding *Key*-*Value* pairs are defined as (K_m, V_m). The MCA operation is mathematically expressed as:

$$\begin{aligned} &\text{MCA}(Q_L, K_m, V_m) \\ &= softmax\left(\frac{Q_L K_m^T}{\sqrt{d}}\right) V_m \\ &= softmax\left(\frac{(Sp^L)W_{Q_L}(Sp^m)W_{K_m}^T}{\sqrt{d}}\right)(Sp^m)W_{V_m} \end{aligned} \tag{10}$$

where $m \in \{L, V, A\}$, $softmax$ represents normalized attention score between Q_L and K_m, W_{Q_L} and W_{K_m} are learnable parameters, d indicates the dimension of Q_L and

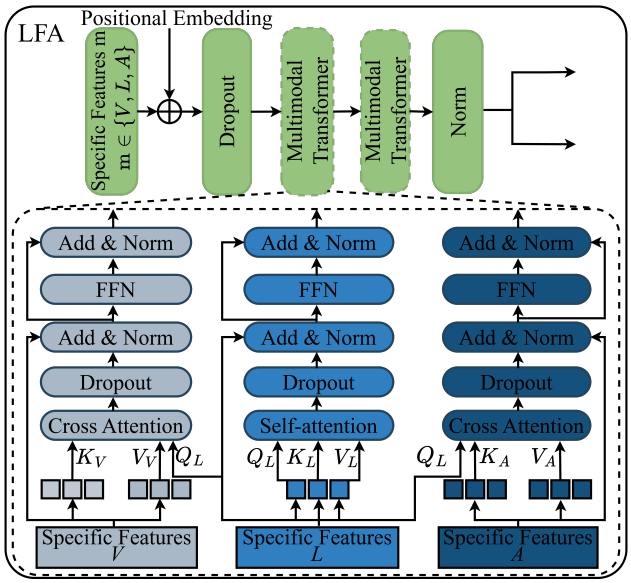

Figure 3: The details of the proposed LFA. The language-focused cross-attention and self-attention achieve targeted feature enhancement: $V \rightarrow L$, $A \rightarrow L$, and $L \rightarrow L$.

K_m. Consequently, the language-focused feature enhancement in the Multimodal Transformer is defined as:

$$\begin{aligned} h_m^{n+1} &= LayerNorm(h_m^n + Drop(\text{MCA}(h_m^n))) \\ h_m^o &= LayerNorm(h_m^{n+1} + FFN(h_m^{n+1})) \end{aligned} \tag{11}$$

where $m \in \{L, V, A\}$, $Drop(\cdot)$ denotes the Dropout operation, and $FFN(\cdot)$ represents a feed-forward module. Here, h_m^n and h_m^{n+1} are the input and intermediate features, respectively, while h_m^o is the output of a Multimodal Transformer layer. As illustrated in Figure 3, cascaded Multimodal Transformers are employed to enhance modality-specific features.

The LFA effectively leverages modality-specific features from three modalities, aligning them with the language modality to strengthen multimodal representation. As shown in Figure 2, the enhanced features are projected into higher-level specific features, denoted as HSp^m ($m \in \{L, V, A\}$), and then integrated with enhanced shared features. Additionally, these features are processed by modality-specific predictors for the specific prediction.

Multimodal Fusion. As illustrated in Figure 2, modality-shared features are first processed through a unified Transformer layer, followed by two fully connected layers, and then projected into higher-level shared features, denoted as HSh. Finally, the multimodal fusion layer combines HSh with HSp^m to form the final multimodal features:

$$\begin{aligned} &F(HSh, HSp^m) \\ &= Concat(HSp^L, HSp^V, HSp^A, HSh), \end{aligned} \tag{12}$$

where $Concat$ denotes the concatenation operation.

Hierarchical Predictions

As shown in Figure 2, a classifier can predict the MSA output $\hat{y}$ after multimodal fusion. The final MSA output loss L_f is defined as:

$$L_f = \frac{1}{N_d} \sum_{n=0}^{N_d} |\hat{y}_n - y_n|, \tag{13}$$

where y_n is the MSA label, N_d is the number of samples. Unlike traditional MSA learning, which only involves a single output loss L_f, the proposed DLF explores hierarchical predictions considering modality-shared loss L_{Sh}, modality-specific loss L_{Sp^m}, and the output loss L_f concurrently. The total MSA learning loss is thus expressed as:

$$L_{MSA} = \sum_{l \in \{f, Sh, Sp^m\}} \beta_l L_l, \tag{14}$$

where $m \in \{L, V, A\}$, β_l are weighting coefficients that control the relative importance of different losses.

Overall Learning Objective. The proposed DLF framework integrates the decoupling loss L_d and the total MSA learning loss L_{MSA} to form the overall learning objective:

$$L_{DLF} = L_d + L_{MSA}. \tag{15}$$

Experiments

Datasets and Evaluation Metrics

We evaluate DLF on two widely used datasets: CMU Multimodal Sentiment Intensity (MOSI) (Zadeh et al. 2016) and CMU Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) (Zadeh et al. 2018b).

MOSI. The MOSI dataset comprises 2,199 monologue video clips, with audio and visual features extracted at 12.5 Hz and 15 Hz, respectively. The dataset is divided into 1,284 training, 229 validation, and 686 test samples.

MOSEI. The MOSEI dataset, significantly larger, consists of 22,856 movie review video clips sourced from YouTube. Features are extracted at 20 Hz for audio and 15 Hz for visual modalities. The dataset is split into 16,326 training samples, 1,871 validation samples, and 4,659 test samples. For both datasets, each video clip is annotated with a sentiment score ranging from -3 to 3, representing a spectrum from highly negative to highly positive sentiment.

Evaluation Metrics. Consistent with established practices in previous studies (Liang et al. 2021; Lv et al. 2021; Mao et al. 2022), the performance of MSA is evaluated using multiple metrics: 7-class accuracy (Acc-7), 5-class accuracy (Acc-5), binary accuracy (Acc-2), F1 score, correlation between model predictions and human annotations (Corr), and mean absolute error (MAE). These metrics collectively offer a comprehensive assessment of DLF's effectiveness across various sentiment analysis tasks.

Implementation Details

In this study, we align our methodology with previous works (Hazarika, Zimmermann, and Poria 2020; Mao et al. 2022) by utilizing the BERT-base-uncased model (Devlin et al. 2018) to extract unimodal linguistic features. This process generates word representations with a 768-dimensional hidden state. For visual data, DLF employs the Facet framework (Baltrušaitis, Robinson, and Morency 2016) to encode each

Method	CMU-MOSI						CMU-MOSEI					
	Acc-7(↑)	Acc-5(↑)	Acc-2(↑)	F1(↑)	Corr(↑)	MAE(↓)	Acc-7(↑)	Acc-5(↑)	Acc-2(↑)	F1(↑)	Corr(↑)	MAE(↓)
TFN*	34.90	39.39†	80.08	80.07	0.698	0.901	50.20	53.10†	82.50	82.10	0.700	0.593
LMF*	33.20	38.13†	82.50	82.40	0.695	0.917	48.00	52.90†	82.00	82.10	0.677	0.623
EF-LSTM†	35.39	40.15	78.48	78.51	0.669	0.949	50.01	51.16	80.79	80.67	0.683	0.601
LF-DNN†	34.52	38.05	78.63	78.63	0.658	0.955	50.83	51.97	82.74	82.52	0.709	0.580
MFN†	35.83	40.47	78.87	78.90	0.670	0.927	51.34	52.76	82.85	82.85	0.718	0.575
Graph-MFN†	34.64	38.63	78.35	78.35	0.649	0.956	51.37	52.69	83.48	83.43	0.713	0.575
MulT	40.00	42.68†	83.00	82.00	0.698	0.871	51.80	54.18†	82.50	82.30	0.703	0.580
PMR	40.60	-	83.60	83.60	-	-	52.50	-	83.60	83.40	-	-
MISA†	41.37	47.08	83.54	83.58	0.778	0.777	52.05	53.63	84.67	84.66	0.752	0.558
MAG-BERT	43.62	-	84.43	84.61	0.781	**0.727**	52.67	-	84.82	84.71	0.755	0.543
DMD**	46.06	-	83.23	83.29	-	0.752	52.78	-	84.62	84.62	-	0.543
DLF (Ours)	**47.08**	**52.33**	**85.06**	**85.04**	**0.781**	0.731	**53.90**	**55.70**	**85.42**	**85.27**	**0.764**	**0.536**

Table 1: Comparision on MOSI and MOSEI. Bold is the best. Note: † represents the result from THUIAR's GitHub page (Thuiar 2024), * represents the result from (Hazarika, Zimmermann, and Poria 2020), - represents the result from the original paper is not provided, and ** represents reproduced results from public code with hyper-parameters provided in the original paper.

video frame, focusing on 35 distinct facial action units as detailed in (Li et al. 2019). For audio processing, we utilize the COVAREP framework (Degottex et al. 2014), which produces 74-dimensional audio features. Our experiments are implemented using the PyTorch framework and executed on one NVIDIA V100 GPU with 32GB of memory. The model is trained with a batch size of 16 and optimized using an initial learning rate of 1e-4. Early stopping with a patience of 10 epochs is applied to ensure convergence.

Main Results

Baselines. We compare the DLF against eleven leading MSA methods on both benchmarks, including **EF-LSTM** (Williams et al. 2018b), **LF-DNN** (Williams et al. 2018a), **TFN** (Zadeh et al. 2017), **LMF** (Liu et al. 2018), **MFN** (Zadeh et al. 2018a), **Graph-MFN** (Zadeh et al. 2018b), **MulT** (Tsai et al. 2019), **PMR** (Lv et al. 2021), **MISA** (Hazarika, Zimmermann, and Poria 2020), **MAG-BERT** (Rahman et al. 2020), and **DMD** (Li, Wang, and Cui 2023).

Performance Comparison. Comparative results, as reported in Table 1, demonstrate that our proposed DLF exhibits superior performance on almost all metrics for both benchmarks. Particularly, we have the following key observations. Compared to decoupled-feature-based MSA methods like MISA (Hazarika, Zimmermann, and Poria 2020), MulT (Tsai et al. 2019), and DMD (Li, Wang, and Cui 2023), the proposed DLF, especially the LFA, captures effective intermodality dynamics and further improves the multimodal representation capability by enhancing the dominant language in the specific subspace. Compared to methods that leverage multimodal transformers to learn cross-modal interactions and fusion such as LMF (Liu et al. 2018), MFN (Zadeh et al. 2018a), and PMR (Lv et al. 2021), our proposed method learns effective multimodal representations in the disentangled subspaces and further facilitates the overall prediction performance using hierarchical predictions which utilizes both pre-fused and post-fused features.

Ablation Study

We conduct extensive ablation studies to thoroughly examine the impact of various modality combinations, different regularization strategies, and critical components including the Feature Disentanglement Module (FDM), Language-Focused Attractor (LFA), and Hierarchical Predictions (HP).

Various Modality Combinations. As shown in Table 2, all analysis are conducted on the MOSI dataset. We first present the performance of each unimodality, it is easy to notice that language modality (L) serves as the dominant one. In the bi-modalities case, we consider both (L,A) and (L,V) pairs in our DLF framework, the results not only demonstrate the performance improvement, especially the fine-grained classification, through two modalities but also showcase that language modality attracts useful information from vision (V) or audio (A) modality by LFA to enhance the multimodal representation capability. Furthermore, the tri-modalities DLF consistently outperforms the bi-modalities DLF on all metrics, which indicates that each modality provides a unique contribution and multimodal learning can effectively improve the MSA performance by reasonably exploiting the information from different modalities.

Different Regularization. We remove each loss to verify the importance of different regularization terms. When removing the soft orthogonality loss L_o, the DLF learns decoupled features under the constraints that focus on separate subspaces. The worst performance suggests the importance of the soft orthogonality loss considering the shared and specific subspaces jointly in the feature disentanglement module. Meanwhile, we also notice that the modified triplet loss L_m in the shared subspace improves the overall performance, which indicates the importance of L_m in learning shared features in the shared subspace. Besides, we observe that both reconstruction loss L_r and specific loss L_s contribute to the model's performance. This is because these two losses ensure feature consistency during disentanglement.

Critical Components. To verify the effectiveness of dif-

Method	Acc-7 (%)	Acc-2 (%)	F1 (%)	MAE(↓)
DLF (Ours)	**47.08**	**85.06**	**85.04**	**0.731**
Different Modalities				
only A	15.31	42.84	26.64	1.453
only V	15.01	43.29	29.73	1.455
only L	45.63	84.45	84.38	0.752
L & A	45.77	83.84	83.88	0.741
L & V	46.65	83.08	83.13	0.745
Different Regularization				
w/o L_r	45.92	84.67	84.59	0.734
w/o L_s	45.36	84.60	84.56	0.740
w/o L_m	45.77	83.99	83.97	0.735
w/o L_o	45.77	83.08	83.16	0.738
Different Components				
w/o FDM	45.92	84.60	84.58	0.739
w/o LFA	42.71	83.84	83.85	0.767
w/o HP	42.42	84.76	84.74	0.761

Table 2: Results of ablation studies on the MOSI benchmark.

ferent components of DLF, we remove each critical component separately. The removal of LFA, replaced by three separate *Query* components like MulT (Tsai et al. 2019), leads to a remarkable decrease in overall MSA performance. This demonstrates that the LFA is a straightforward and effective component, mitigating the potential redundancies or conflicts in traditional cross-attention mechanisms. Moreover, when deactivating the FDM, the performance also becomes inferior to that of DLF, which shows the effectiveness of the designed FDM and further indicates the redundant and conflicting information limits the MSA performance. With a similar phenomenon, subtracting the HP module (which just remains the final output loss function) decreases the overall performance again, revealing the value of both the pre-fused and post-fused features.

Further Analysis

To further study the impact of sentiment granularity on MSA, we present the confusion matrix and corresponding accuracy of each sentiment for the MOSI benchmark. As shown in Figure 4, we observe that most sentiment classes have similar accuracy above 40%. However, "HN" and "HP", especially "HN", have the worst performance, limiting the overall MSA performance. Furthermore, when we dive into the confusion matrix, it can be noticed that the samples for "HN" and "HP" are relatively less than other sentiments, which strongly indicates that the long-tailed distribution of the data limits the overall MSA performance, which can be studied in the future.

To better understand the effectiveness of our method, we visualize the distribution of fused multimodal representations. As depicted in Figure 5, compared to DMD, which is a decoupled-multimodal-distillation strategy, our proposed DLF shows superior performance in separating different

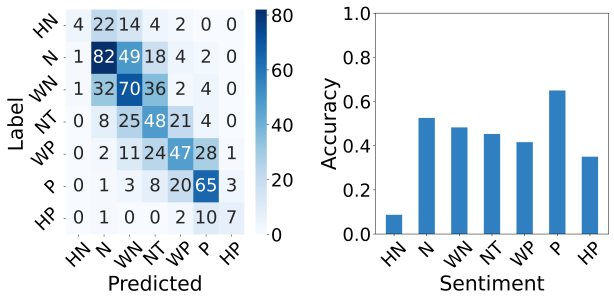

Figure 4: *Left*: Confusion matrix on MOSI. *Right*: Corresponding accuracy for each sentiment. HN: Highly Negative; N: Negative; WN: Weakly Negative; NT: Neutral; WP: Weak Positive; P: Positive; HP: Highly Positive.

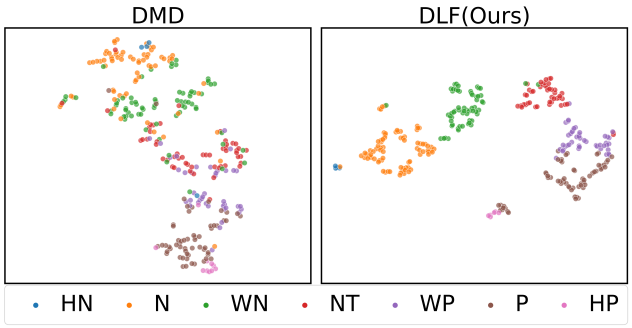

Figure 5: Visualization of the fused multimodal representations. HN: Highly Negative; N: Negative; WN: Weakly Negative; NT: Neutral; WP: Weak Positive; P: Positive; HP: Highly Positive.

sentiments. This is mainly due to that the LFA mitigates redundant and conflicting information during multimodal interaction compared to the interaction between random pairs.

Conclusion

In this paper, we propose the DLF framework to improve the MSA performance. DLF yields powerful multimodal representations by following the pipeline of feature extraction, disentanglement, enhancement, fusion, and prediction, mainly benefiting from the feature disentanglement module, language-focused attractor, and hierarchical predictions. Extensive results verify the superiority of DLF by comparisons with eleven baselines and comprehensive ablation studies.

Broad impacts. (i) This study demonstrates its potential to exploit the imbalanced capabilities of various modalities in multimodal learning, thereby setting a new benchmark in this field. (ii) The proposed LFA facilitates the generalization of our method to other multimodal scenarios by changing the dominant modality. **Limitation and future work.** Our method only considers the scenarios of complete modalities. When facing missing modalities, the feature disentanglement and enhancement modules are potentially limited.

Acknowledgments

This work is supported in part by NIH R01EB033387, NSF CNS-2122320, and NSF CCF-2324937. Tianlong Chen is supported by NIH OT2OD038045-01 and UNC SDSS Seed Grant.

References

Ali, K.; and Hughes, C. E. 2023. A Unified Transformer-based Network for Multimodal Emotion Recognition. *arXiv preprint arXiv:2308.14160*.

Aslam, M. H.; Zeeshan, M. O.; Pedersoli, M.; Koerich, A. L.; Bacon, S.; and Granger, E. 2023. Privileged Knowledge Distillation for Dimensional Emotion Recognition in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3337–3346.

Awal, R.; Ahmadi, S.; Zhang, L.; and Agrawal, A. 2024. VisMin: Visual Minimal-Change Understanding. *arXiv preprint arXiv:2407.16772*.

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, 1–10. IEEE.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, 960–964. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ezzameli, K.; and Mahersia, H. 2023. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 101847.

Geetha, A.; Mala, T.; Priyanka, D.; and Uma, E. 2024. Multimodal Emotion Recognition with deep learning: advancements, challenges, and future directions. *Information Fusion*, 105: 102218.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusion-Bench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385.

Guo, J.; Tang, J.; Dai, W.; Ding, Y.; and Kong, W. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM international conference on multimedia*, 3394–3402.

Guo, Q.; Wang, X.; Wu, Y.; Yu, Z.; Liang, D.; Hu, X.; and Luo, P. 2020. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11020–11029.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2827–2836.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

Huang, J.; Tao, J.; Liu, B.; Lian, Z.; and Niu, M. 2020. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3507–3511. IEEE.

Kim, D.; and Kang, P. 2022. Cross-modal distillation with audio–text fusion for fine-grained emotion classification using BERT and Wav2vec 2.0. *Neurocomputing*, 506: 168–183.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *International conference on machine learning*, 2649–2658. PMLR.

Kim, K.; and Park, S. 2023. AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis. *Information Fusion*, 92: 37–45.

Lei, Y.; Yang, D.; Li, M.; Wang, S.; Chen, J.; and Zhang, L. 2023. Text-oriented modality reinforcement network for multimodal sentiment analysis from unaligned multimodal sequences. In *CAAI International Conference on Artificial Intelligence*, 189–200. Springer.

Li, M.; Yang, D.; Lei, Y.; Wang, S.; Wang, S.; Su, L.; Yang, K.; Wang, Y.; Sun, M.; and Zhang, L. 2024a. A Unified Self-Distillation Framework for Multimodal Sentiment Analysis with Uncertain Missing Modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10074–10082.

Li, M.; Yang, D.; Zhao, X.; Wang, S.; Wang, Y.; Yang, K.; Sun, M.; Kou, D.; Qian, Z.; and Zhang, L. 2024b. Correlation-Decoupled Knowledge Distillation for Multimodal Sentiment Analysis with Incomplete Modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12458–12468.

Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6631–6640.

Li, Y.; Zeng, J.; Shan, S.; and Chen, X. 2019. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer vision and pattern recognition*, 10924–10933.

Liang, T.; Lin, G.; Feng, L.; Zhang, Y.; and Lv, F. 2021. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8148–8156.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Lv, F.; Chen, X.; Huang, Y.; Duan, L.; and Lin, G. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2554–2562.

Mao, H.; Yuan, Z.; Xu, H.; Yu, W.; Liu, Y.; and Gao, K. 2022. M-sena: An integrated platform for multimodal sentiment analysis. *arXiv preprint arXiv:2203.12441*.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6892–6899.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1): 309–325.

Thuiar. 2024. MMSA: A Multi-Modal Sentiment Analysis Toolkit. https://github.com/thuiar/MMSA. Accessed: 2024-08-10.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Wang, W.; Ding, L.; Shen, L.; Luo, Y.; Hu, H.; and Tao, D. 2024. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2282–2291.

Williams, J.; Comanescu, R.; Radu, O.; and Tian, L. 2018a. Dnn multimodal fusion techniques for predicting video sentiment. In *Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML)*, 64–72.

Williams, J.; Kleinegesse, S.; Comanescu, R.; and Radu, O. 2018b. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, 11–19.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12113–12132.

Yang, D.; Chen, Z.; Wang, Y.; Wang, S.; Li, M.; Liu, S.; Zhao, X.; Huang, S.; Dong, Z.; Zhai, P.; et al. 2023a. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19005–19015.

Yang, D.; Huang, S.; Kuang, H.; Du, Y.; and Zhang, L. 2022a. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1642–1651.

Yang, D.; Huang, S.; Liu, Y.; and Zhang, L. 2022b. Contextual and cross-modal interaction for multi-modal speech emotion recognition. *IEEE Signal Processing Letters*, 29: 2093–2097.

Yang, D.; Kuang, H.; Huang, S.; and Zhang, L. 2022c. Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1708–1717.

Yang, K.; Yang, D.; Zhang, J.; Wang, H.; Sun, P.; and Song, L. 2023b. What2comm: Towards communication-efficient collaborative perception via feature decoupling. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7686–7695.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, 10790–10797.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, 5634–5641.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018b. 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*, 2236–2246.

Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; and Yu, T. 2023. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2310.05804*.