Progressive Representation Learning for Multimodal Sentiment Analysis with Incomplete Modalities

Jindi Bao[1], Jianjun Qian[1⋆], Mengkai Yan[2], and Jian Yang[1,3⋆]

[1] PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology
[2] College of Computer Science and Software Engineering, Hohai University
[3] PCA Lab, School of Intelligence Science and Technology, Nanjing University

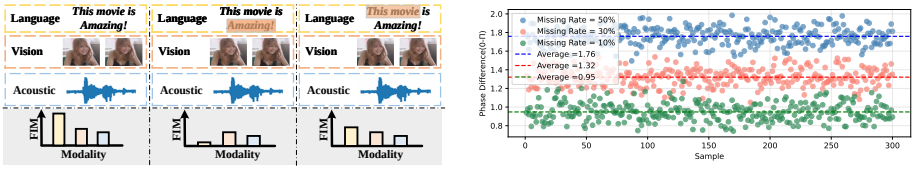
Abstract. Multimodal Sentiment Analysis (MSA) seeks to infer human emotions by integrating textual, acoustic, and visual cues. However, existing approaches often rely on all modalities are completeness, whereas real-world applications frequently encounter noise, hardware failures, or privacy restrictions that result in missing modalities. There exists a significant feature misalignment between incomplete and complete modalities, and directly fusing them may even distort the well-learned representations of the intact modalities. To this end, we propose PRLF, a Progressive Representation Learning Framework designed for MSA under uncertain missing-modality conditions. PRLF introduces an Adaptive Modality Reliability Estimator (AMRE), which dynamically quantifies the reliability of each modality using recognition confidence and Fisher information to determine the dominant modality. In addition, the Progressive Interaction (ProgInteract) module iteratively aligns the other modalities with the dominant one, thereby enhancing cross-modal consistency while suppressing noise. Extensive experiments on CMU-MOSI, CMU-MOSEI, and SIMS verify that PRLF outperforms state-of-the-art methods across both inter- and intra-modality missing scenarios, demonstrating its robustness and generalization capability.

Keywords: Multimodal Sentiment Analysis, Affective Computing

1 Introduction

Multimodal Sentiment Analysis (MSA) has emerged as a crucial research topic in affective computing, aiming to infer human emotions by integrating heterogeneous cues from text, speech, and visual signals [9, 11, 12, 32, 34, 42, 43, 47]. By leveraging the complementarity among different modalities, MSA has achieved remarkable progress [41, 45]. However, most existing approaches rely on an ideal assumption that all modalities are consistently available in both the training and inference stages [24]. In practical scenarios, this assumption rarely holds true, as various factors such as environmental noise, hardware malfunction, data

⋆ Corresponding authors.



(a) Variation of Modality Feature Information under Different Missing-Data Conditions.

(b) Feature Phase Differences under Different Missing Rates.

Fig. 1: Feature information loss and feature phase shift caused by missing modalities.

transmission failure, or privacy restrictions may cause uncertain and incomplete modality inputs [20]. Such imperfect multimodal data often lead to degraded feature representations and unstable emotional understanding, posing a critical challenge for building robust MSA systems [22].

Recently, growing attention has been directed toward addressing the challenge of incomplete modalities in MSA, and existing research can generally be categorized into two main directions: (**i**) Generative approaches, which recover the missing modalities by leveraging the information from the available ones [23, 27, 31, 50, 52]; (**ii**) Distillation-based methods, which transfer knowledge from complete-modality models to the models under missing-modality conditions [15, 17–19, 33]. Although two strategies have made progress, these methods still follow fusion paradigms similar to traditional approaches and fail to account for the importance differences between missing and complete modalities. We use the Fisher Information Matrix (FIM) [1, 25] to evaluate the impact of missing modalities. As shown in Fig. 2, the loss of critical data in a specific modality can lead to a significant reduction in information content, whereas the absence of non-critical data has a relatively minor impact. Therefore, identifying the importance differences among modalities under various missing-data conditions is crucial for optimizing model design and fusion strategies.

Meanwhile, the impact of missing data is often difficult to eliminate, and existing methods generally overlook its effect during multimodal fusion. To investigate the impact of modality missing rates on the extracted features, we selected 300 samples and calculated the phase differences between the features obtained under various missing rates and those extracted from the complete modality, as shown in Fig. 1b. Here, phase difference refers to the angular deviation in the high dimensional feature space. We observed that as the missing rate increases, the phase difference between the extracted features and those obtained from the complete modality also increases. Therefore, the impact of missing data during modality interaction and fusion is manifested as phase shifts in the high-dimensional feature space.

Based on the above discussion, we propose a **P**rogressive **R**epresentation **L**earning **F**ramework (PRLF) for multimodal sentiment analysis under uncertain missing-modality conditions. The framework enables the model to adaptively learn the importance of each modality for every sample under various missing-data conditions, identify the dominant and auxiliary modalities, and fuse them in a progressive manner. In particular, we design an Adaptive Modality Reliability Estimator (AMRE), where a router network dynamically evalu-

ates modality reliability based on recognition confidence and Fisher information, and subsequently leverages the dominant modality to guide cross-modal interaction and fusion. Further considering that missing data disrupts the feature distribution in high-dimensional space, direct cross-modal fusion may distort the representations of the intact modalities. To this end, we design a Progressive Interaction Module (ProgInteract), which performs iterative interactions to gradually align the feature distribution of auxiliary modalities with that of the dominant modality, thereby enhancing emotion information extraction and mitigating noise interference. Specifically, at the early stage of training, the model encounters substantial noise and therefore focuses on extracting intra-modal features while performing only limited cross-modal interactions. As the iteration progresses and the intra-modal representations become more stable, strengthening cross-modal interactions encourages auxiliary modalities to align with the dominant one, ultimately enabling effective multimodal fusion.

Our main contributions are summarized as follows:

- We propose a Progressive Interaction Module (ProgInteract) that iteratively aligns auxiliary modality features with the dominant modality, enabling noise-robust and adaptive cross-modal fusion under missing-data conditions.
- We propose Adaptive Modality Reliability Estimator to evaluate the effectiveness of each modality and adaptively determines the dominant modality.
- We evaluate our method on standard multimodal datasets, including CMU-MOSI, CMU-MOSEI and SIMS, and the results demonstrate that our approach outperforms or matches the state-of-the-art.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) is a task designed to perceive and process heterogeneous data such as language, audio, and visual information, with the goal of understanding and interpreting human emotional states [10, 39, 54]. Current mainstream research primarily focuses on developing more effective modality fusion strategies and interaction mechanisms to further enhance the overall performance of MSA [7,13,21,37,46]. For instance, MAMSA [35] adjusts the contribution of text and image features through adaptive attention and hierarchical fusion to effectively extract and integrate sentiment-related information across modalities. DB-MPCA [16] integrates raw acoustic and visual data into a language model via multimodal pretraining and cross-modal attention, preserving textual dominance for more balanced fusion. DLF [36] disentangles shared and modality-specific features, refines them with geometric constraints, and enhances language representations via language-guided cross-attention for improved sentiment analysis. However, these methods assume the availability of all modalities during inference, making them difficult to apply in real-world scenarios where data from certain modalities may be missing, incomplete, or corrupted.

2.2 Multimodal Sentiment Analysis with Incomplete Data

In multimodal sentiment analysis, the missing modality problem is typically tackled using two main strategies: (1) Generative approaches and (2) Distillation-based approaches. Generative approaches focus on reconstructing absent features and semantic information by modeling the distributions of the observed modalities. For example, DiCMoR [38] using class-specific normalizing flows to align distributions between recovered and true data. MRAN [27] aligns and reconstructs modalities to enhance robustness when the text modality is missing. Distillation-based approaches leverage knowledge from models trained on complete modalities to steer the training of models that must operate with incomplete or missing modalities. For example, UMDF [17] transfers cross-modal knowledge through multi-grained interaction and dynamic feature integration for robust sentiment analysis under uncertain missing modalities. CorrKD [19] leverages contrastive, prototype-based, and consistency distillation to reconstruct missing semantics and enhance robustness under uncertain missing modalities. However, existing methods fail to account for the directional consistency between features from missing modalities and those from complete modalities, making it difficult to suppress noise interference during multimodal interaction and fusion.

3 Method

3.1 Problem Formulation

For a multimodal video segment represented as $\boldsymbol{S} = [\boldsymbol{X}_V, \boldsymbol{X}_A, \boldsymbol{X}_L]$, the three components $\boldsymbol{X}_V \in \mathbb{R}^{T_V \times d_V}$, $\boldsymbol{X}_A \in \mathbb{R}^{T_A \times d_A}$, $\boldsymbol{X}_L \in \mathbb{R}^{T_L \times d_L}$ correspond to the visual, acoustic, and language modalities, typical of multimodal data, respectively. Here, $T_m(\cdot)$ and $d_m(\cdot)$ denote the sequence length and feature dimension of modality $m \in \{\boldsymbol{V}, \boldsymbol{A}, \boldsymbol{L}\}$. To simulate realistic multimodal scenarios, we consider two types of missingness: (i) intra-modality missingness, referring to the absence of certain frame-level features within a modality sequence; and (ii) inter-modality missingness, referring to the complete absence of modalities.

3.2 Overall Framework

Fig. 2 illustrates the workflow of PRLF. PRLF consists of two key components: the Adaptive Modality Reliability Estimator (AMRE) and the Progressive Interaction module (ProgInteract). The AMRE module identifies the dominant modality based on classification confidence and Fisher information. The ProgInteract module gradually aligns the feature distributions of the auxiliary modalities with that of the dominant modality through iterative interactions. The PRLF model takes three modalities as input: visual (V), acoustic (A), and language (L). Before entering the AMRE module, a dedicated encoder $f_m = \varepsilon_m(\boldsymbol{X}_m)$ is employed for each modality to extract its feature representation. The resulting f_m serves as the unimodal feature, which is used both as input to the AMRE module and for the subsequent ProgInteract to enable collaborative learning and feature alignment across modalities.

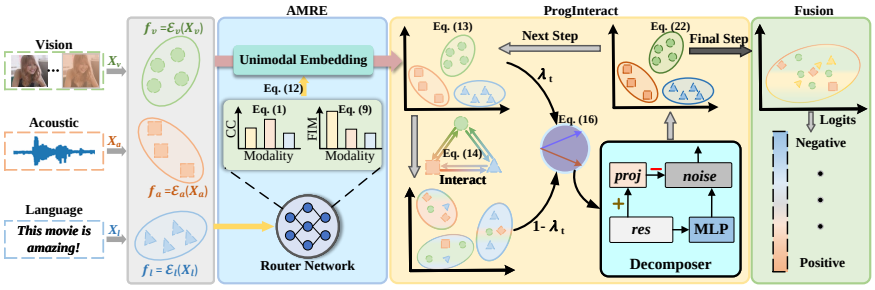


Fig. 2: The structure of PRLF, which consists of two key components: Adaptive Modality Reliability Estimator (AMRE) and Progressive Interaction module (ProInteract).

3.3 Adaptive Modality Reliability Estimator

Confidence-based Modality Importance (CMI). To dynamically assess the importance of each modality for different samples, many multimodal methods compute modality importance weights based on the recognition accuracy and confidence of each unimodal model on the given sample [26,28,53]. Based on this, we design a routing network that effectively assigns an individually independent classification head to each modality. For a given sample, the classification confidence of modality m for the correct class is obtained as $\alpha_m^{(i)} = \hbar_m(f_m^{(i)})$, where $m \in \{\boldsymbol{V}, \boldsymbol{A}, \boldsymbol{L}\}$, $\hbar_m$ is the classification head, i represents sample, and α denotes the confidence for the correct class. The classification confidences of all modalities are then concatenated and normalized for unified representation, as follows:

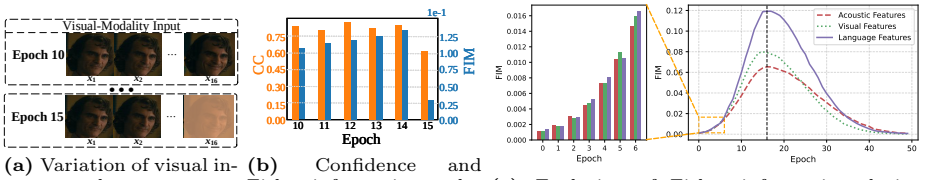
$$\alpha^{(i)} = [\alpha_v^{(i)}, \alpha_a^{(i)}, \alpha_l^{(i)}], \quad \hat{\alpha}^{(i)} = \frac{\alpha^{(i)}}{\left\|\alpha^{(i)}\right\|_1} \tag{1}$$

Each classification head is trained with a cross-entropy loss:

$$\mathcal{L}_{uni} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{m} y^{(i)} \log \sigma(\hbar_m(f_m^{(i)})), \tag{2}$$

where $y^{(i)}$ is the ground-truth label and $\sigma(\cdot)$ denotes the softmax function. However, we found that under missing data conditions, relying solely on classification confidence is insufficient for reliable evaluation. As shown in Fig. 3a, in the visual modality, the model receives complete input data during epochs 10–14, while key frames are missing at epoch 15. Fig. 3b illustrates the variations in classification confidence and Fisher information. During epochs 10–14, the model produces correct predictions with high confidence. Yet, even when key frames are missing in epoch 15, it still reports a high confidence score. This phenomenon may result from the model's memorization of facial features [3,40]. Therefore, relying solely on confidence to identify the dominant modality is unreliable. Notably, the trace of the Fisher Information Matrix reveals a substantial decline in the effective information contained in the input.

Fisher Information-based Modality Importance (FIMI). The trace of the Fisher Information Matrix (FIM) measures the overall sensitivity of model



(a) Variation of visual input completeness across training epochs.

(b) Confidence and Fisher information under missing key frames.

(c) Evolution of Fisher information during training under complete-data inputs.

Fig. 3: Analysis of confidence and Fisher information variations caused by missing key frames in the visual modality.

parameters to the output distribution. It reflects how much effective information the model extracts from the data. In multimodal learning, for a unimodal encoder $\varepsilon_m(\cdot)$ and its corresponding classification head $\hbar_m(\cdot)$, the gradient of the model parameters θ can be expressed as:

$$g_m = \nabla_\theta \big(\hbar_m \circ \varepsilon_m(\boldsymbol{X}_m)\big) \tag{3}$$

Based on this, the trace of the FIM for modality m can be written as:

$$\mathrm{Tr}(F_m^{(i)}) = \mathrm{Tr}\Big(\mathbb{E}_i[g_m^{(i)} {g_m^{(i)}}^\top]\Big) = \mathbb{E}_i[\|g_m^{(i)}\|_2^2] \tag{4}$$

where Tr denotes the trace of a matrix, $\mathbb{E}$ represents the mathematical expectation, $g_m g_m^\top$ denotes the outer product of the gradient vector, and $\|g_m\|_2^2$ represents the squared ℓ_2 norm of the gradient. A larger Tr(FIM) indicates that the model's output is more sensitive to input perturbations, thereby reflecting the amount of effective information contained in that modality. As shown in Fig. 3b, we observe that when key frames are missing, the trace of the Fisher Information Matrix, $\mathrm{Tr}(F_m)$, exhibits a significant decrease. To explain this phenomenon, we provide the following theoretical interpretation.

Assume that the visual modality consists of 16 frames $\mathbf{X}_m = \{x_1, \ldots, x_{16}\}$. The gradient of the model with respect to the parameters can be approximately decomposed as the sum of contributions from individual frames:

$$\nabla_\theta \log p_\theta(y \mid \mathbf{X}m) \approx \sum_{t=1}^{16} g_t, \tag{5}$$

where g_t denotes the gradient contribution of the t-th frame. Under partial-missing conditions, we introduce a binary mask $\delta_t \in \{0, 1\}$ to indicate whether a frame is present. The gradient then becomes:

$$\nabla_\theta \log p_\theta(y \mid \mathbf{X}m) \approx \sum_{t=1}^{16} \delta_t g_t. \tag{6}$$

Substituting this into the definition of the trace of the Fisher information yields:

$$\mathrm{Tr}(F_m) = \mathbb{E}\left[\left\|\textstyle\sum_{t=1}^{16} \delta_t g_t\right\|_2^2\right] \tag{7}$$

If certain frames contain key semantic or emotional cues, their gradient magnitudes are typically much larger,

$$|g_t^{\text{key}}|_2^2 \gg |g_t^{\text{non}}|_2^2. \tag{8}$$

when key frames are removed ($\delta_t = 0$), the dominant contributors to the gradient energy vanish, leading to a substantial reduction in the overall gradient norm and consequently a clear decline in $\text{Tr}(F_m)$. In contrast, removing non-key frames has a relatively minor impact on the Fisher information.

Therefore, we compute the trace of FIM for each modality and form a vector $\beta^{(i)}$ for the i-th sample:

$$\beta^{(i)} = [\text{Tr}(F_v^{(i)}), \text{Tr}(F_a^{(i)}), \text{Tr}(F_l^{(i)})], \quad \hat{\beta}^{(i)} = \frac{\beta^{(i)}}{\left\|\beta^{(i)}\right\|_1} \tag{9}$$

The resulting $\hat{\beta}^{(i)}$ is used as the Fisher information-based modality importance for the i-th sample. However, in the early stages of training, the model exhibits weak gradient responses, resulting in low Fisher information across all modalities. Consequently, it is difficult to provide a reliable estimation of modality importance at this stage, as shown in Fig. 3c At this stage, the model's classification confidence can reflect the relative contribution of each modality. Although the confidence is not yet stable, a modality with relatively higher confidence tends to capture more discriminative patterns in its feature space. Therefore, classification confidence can serve as a complementary and indicative cue for assessing modality importance during the early phase of training.

Modality Importance Fusion Mechanism. Based on the above analysis, we design a modality importance fusion mechanism that allows the classification confidence and Fisher information to complement each other, thereby enabling a more accurate estimation of modality importance. Inspired by [14], we first assess whether FIM can provide a reliable measure of modality importance by evaluating its growth between consecutive training epochs, as follows:

$$\Delta_m^{(t,i)} = \frac{\text{Tr}(F_m^{(t,i)}) - \text{Tr}(F_m^{(t-1,i)})}{\text{Tr}(F_m^{(t,i)})} \tag{10}$$

where t is the training epoch, $m \in \{\boldsymbol{V}, \boldsymbol{A}, \boldsymbol{L}\}$. This ratio measures the increase of the Fisher information in the current epoch relative to the previous epoch. Dynamically compute the fusion weight between classification confidence and Fisher information based on the relative change in Fisher information $\Delta_m^{(t,i)}$:

$$w^{(t,i)} = \sigma\big([\Delta_v^{(t,i)}, \Delta_a^{(t,i)}, \Delta_l^{(t,i)}]\big) \tag{11}$$

where $w^{(t,i)}$ represents the fusion weight for the i-th sample at training epoch t, σ denotes the Sigmoid function. The modality importance derived from classification confidence and that derived from Fisher information are fused according to the dynamically computed weights $w_m^{(t)}$:

$$\mu^{(t,i)} = (1 - w^{(t,i)})\,\hat{\alpha}^{(i)} + w^{(t,i)}\,\hat{\beta}^{(i)} \tag{12}$$

Accordingly, in early training, the low fusion weight makes the fusion rely more on classification confidence, whereas when Fisher information rises significantly, a higher weight shifts the fusion toward Fisher information.

3.4 Progressive Interaction Module

In Fig. 1b, we observe that the impact of missing data can cause deviations in feature directions within the high-dimensional feature space. During the interaction and fusion process, such deviations can negatively affect the final fused features. To mitigate this, we propose a progressive interaction module, which does not perform direct feature fusion. Instead, it iteratively refines multimodal representations, focusing on unimodal feature enhancement in the early iterations and emphasizing cross-modal interactions in the later ones.

Specifically, in each iteration, the three modal features f_m first perform self-refinement to extract their internal discriminative information:

$$f_m^{\text{self}} = f_m + \text{Dropout}\Big(\text{ReLU}(f_m W_1 + b_1)W_2 + b_2\Big) \tag{13}$$

Next, cross-modal interactions are performed across all modalities, with the modality importance μ_m incorporated during the interaction process:

$$f_{m\to n} = \text{softmax}\left(\frac{(\mu_m f_m)(\mu_n f_n)^{\mathbf{T}}}{\sqrt{d}}\right)(\mu_m f_m), \quad n \in \{n_1, n_2\} \tag{14}$$

where $m, n_1, n_2 \in \{\boldsymbol{V}, \boldsymbol{A}, \boldsymbol{L}\}$, $n_1 \neq n_2 \neq m$, d means the dimension of f_m. We concatenate the features $f_c = [f_{m\to n_1}, f_{m\to n_2}]$, and extract the fused features:

$$f_m^{\text{cross}} = \text{softmax}(\frac{f_c f_c^{\mathbf{T}}}{\sqrt{d}})f_c \tag{15}$$

For unimodal features and cross-modal features, We introduce a time-dependent weighting coefficient λ_t to balance the contribution between the unimodal feature f_m^{self} and the cross-modal feature f_m^{cross}:

$$f_m^{\text{fuse},t} = \lambda_t f_m^{\text{self}} + (1-\lambda_t) f_m^{\text{cross}} \tag{16}$$

where, $\lambda_t = 1 - \frac{t}{\max(1,\text{steps}-1)}$, steps > 0, t is the current iteration step, and steps is the total iterations. Based on this, the model focuses on unimodal features in the early iterations and emphasizes cross-modal features in the later stages.

Subsequently, the dominant and auxiliary modality features are determined based on the modality importance μ, and $dom, aux_1, aux_2 \in \{\boldsymbol{V}, \boldsymbol{A}, \boldsymbol{L}\}$. In the high-dimensional feature space, we utilize the phase difference between the dominant and auxiliary modalities to guide the auxiliary one in learning complementary information from the dominant modality. To achieve this, we design a Decomposer module that adaptively models how the auxiliary modality aligns with the dominant one. Specifically, at each iteration step t, the dominant feature $f_{dom}^{fuse,t}$ and the auxiliary feature $f_a^{fuse,t}$ are concatenated to form a joint representation $[f_{dom}^{fuse,t}; f_{aux}^{fuse,t}] \in \mathbb{R}^{2D}$. This representation is fed into a gating network, which predicts a projection weight g_{aux}^t indicating the degree of phase alignment between the two modalities:

$$h_1 = \text{ReLU}\big(W_1[f_{dom}^{fuse,t}; f_{aux}^{fuse,t}] + b_1\big) \in \mathbb{R}^D, \quad g_{aux}^t = \sigma(W_2 h_1 + b_2) \in \mathbb{R}^D \tag{17}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function. Accordingly, the dominant modality feature is projected into the auxiliary space:

$$proj_{aux}^{t} = g_{aux}^{t} \odot f_{dom}^{fuse,t} \tag{18}$$

The residual component of the auxiliary modality, representing information not captured by the projection:

$$res_{aux}^{t} = f_{aux}^{fuse,t} - proj_{aux}^{t} \tag{19}$$

To balance alignment and complementarity, we impose a phase constraint on the orthogonality between the projection and residual terms:

$$\mathcal{L}_{phase}^{t} = \frac{1}{N} \sum_{n \in aux} \mathbb{E}[((proj_{n}^{t})^{\mathbf{T}} res_{n}^{t})^{2}] \tag{20}$$

Minimizing this loss encourages a moderate phase convergence, reducing excessive misalignment while preserving inter-modal complementarity.

However, the residual component may still contain noise. To suppress such noise, a denoising network is applied to estimate the noise within the residual:

$$noise_{aux}^{t} = \text{Dropout}(\text{ReLU}(W_{aux} res_{aux}^{t})) \tag{21}$$

Finally, the cleaned residual and the projection are fused to form the refined auxiliary feature for the next iteration:

$$f_{aux}^{t+1} = proj_{aux}^{t} + \gamma(res_{aux}^{t} - noise_{aux}^{t}), \quad f_{dom}^{t+1} = f_{dom}^{fuse,t} \tag{22}$$

Here, γ serves as a balance coefficient that controls the contribution of the denoised residual information to the auxiliary feature update. In this way, the Decomposer enables each auxiliary modality to dynamically align with the dominant one, learn complementary information, and iteratively refine its representation under noise-suppressed guidance.

3.5 Objective optimization

After the final iteration, the refined representations from all three modalities $f_{dom}^{final}, f_{aux_1}^{final}, f_{aux_2}^{final}$ are integrated to perform the final sentiment prediction:

$$f_{final} = [f_{dom}^{final}, f_{aux_1}^{final}, f_{aux_2}^{final}] \tag{23}$$

The sentiment analysis objective is optimized using the cross-entropy loss:

$$\mathcal{L}_{task} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log\left(\hat{y}_{i,c}\right), \tag{24}$$

where $\hat{y}_{i,c} = \text{Softmax}(W f_{final}^{(i)} + b)_c$ denotes the predicted probability. We integrate the above losses:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \eta_1 \mathcal{L}_{uni} + \eta_2 \mathcal{L}_{phase} \tag{25}$$

Table 1: Performance comparison on CMU-MOSI and CMU-MOSEI datasets under different missing-modality conditions.

Dataset	Models	{l}	{a}	{v}	{l,a}	{l,v}	{a,v}	Avg.	{l,a,v}
MOSI	Self-MM [45]	67.80	40.95	38.52	69.81	74.97	47.12	56.53	84.64
	MISA [9]	81.17	43.51	49.24	80.99	81.49	49.42	64.3	83.36
	TETFN [29]	81.04	39.32	39.32	81.09	80.99	39.32	60.18	82.33
	MMIM [8]	81.17	36.52	31.53	81.78	81.02	37.63	58.28	82.28
	UMDF [19]	82.92	67.80	59.92	**85.63**	84.09	72.98	75.56	83.36
	HRLF [18]	83.36	69.47	**64.59**	83.82	83.56	75.62	76.74	84.15
	CorrKD [19]	81.20	66.52	60.72	83.56	82.41	73.74	74.69	83.94
	EMOE [6]	78.66	65.64	58.1	83.52	80.24	73.69	73.31	85.4
	PRLF	**83.82**	**69.63**	64.05	84.98	**84.13**	**76.03**	**77.02**	**85.78**
MOSEI	Self-MM [45]	71.53	43.57	37.61	75.91	74.62	49.52	58.79	83.69
	MISA [9]	80.55	58.99	58.99	83.76	83.46	58.99	70.79	84.55
	TETFN [29]	81.88	58.99	58.98	83.86	83.55	58.98	71.04	84.83
	MMIM [8]	80.96	59	59.28	81.09	81.29	59.43	70.18	82.06
	UMDF [19]	81.57	67.42	61.57	83.25	82.14	69.48	74.24	82.16
	HRLF [18]	82.05	69.32	**64.90**	82.62	81.09	73.80	75.63	82.93
	CorrKD [19]	80.76	66.09	62.30	81.74	81.28	71.92	74.02	82.16
	EMOE [6]	78.46	65.33	61.54	81.88	81.26	70.57	73.17	85.3
	PRLF	**82.31**	**69.49**	63.67	**84.06**	**83.63**	**74.32**	**76.24**	**85.44**

4 Experiment

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate PRLF and representative MSA methods on three benchmark datasets: CMU-MOSI [48], CMU-MOSEI [49], and SIMS [44]. CMU-MOSI includes 2,199 video clips with acoustic and visual features. CMU-MOSEI consists of 22,856 YouTube review clips with features sampled at 20 Hz and 15 Hz. SIMS is a Chinese multimodal sentiment dataset with 2,281 film and TV segments, providing aligned text, audio, and visual annotations.

Evaluation Metric. We evaluate PRLF using the F1 score on three datasets, and additionally report its ACC-7, ACC-5, ACC-2, and MAE metrics on the same datasets in the supplementary material.

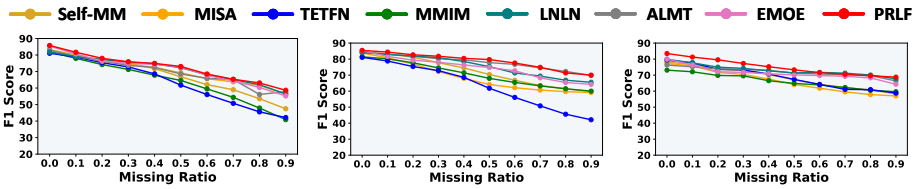
Implementation details. For the CMU-MOSI and CMU-MOSEI datasets, we adopt 300-dimensional GloVe embeddings [30] and 768-dimensional BERT-base-uncased hidden states [5] for the language modality. Visual features are obtained from Facet [2], which provides 35-dimensional facial action unit representations, and acoustic features are extracted using COVAREP [4], yielding 74-dimensional descriptors. the detailed hyper-parameter settings are as follows: the balance coefficient γ is 0.8, the loss weight η_1, η_2 are 0.5,0.1, the iteration step is 4. The missing modality features are substituted with zero vectors. During training, the seed changes with each epoch, resulting in different missing patterns for each sample per round, with all three modalities missing simultaneously and the missing rate kept constant. During testing, we follow the CorrKD [19] setup, and all results are averaged over five random seeds.

4.2 Comparison with State-of-the-art Methods.

Robustness to Inter-modality Missingness. As shown in Table 1, PRLF consistently surpasses existing multimodal fusion methods on both CMU-MOSI

Table 2: Performance comparison on SIMS under different missing conditions

Models	Testing Conditions						
	{l}	{a}	{v}	{l,a}	{l,v}	{a,v}	Avg.
MISA [9]	75.15	56.95	56.82	77.07	75.17	56.78	66.32
Self-MM [45]	67.11	70.36	72.85	73.71	76.76	78.00	73.13
MMIM [8]	73.18	57.70	57.86	73.43	72.96	57.76	65.48
TETFN [29]	79.19	56.82	56.82	79.19	79.16	56.82	68
ALMT [51]	76.17	81.08	**81.91**	75.75	76.36	81.91	78.86
LNLN [50]	79.72	**81.91**	**81.91**	80.11	79.69	81.91	80.86
EMOE [6]	78.66	81.25	80.69	79.54	79.36	81.27	80.12
PRLF	**80.28**	81.74	81.63	**80.57**	**80.36**	**82.58**	**81.19**



(a) MOSI. **(b)** MOSEI. **(c)** SIMS.

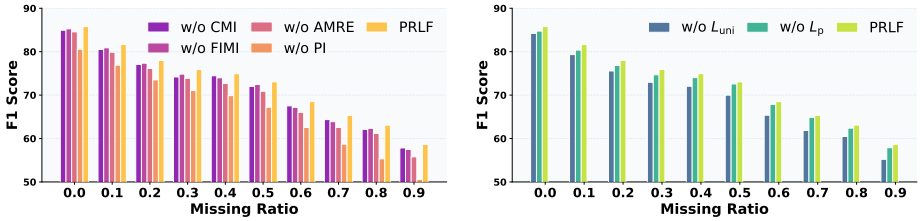
Fig. 4: Comparison results of intra-modality missingness. We report the F1 score.

and CMU-MOSEI datasets across all modality settings. On MOSI, PRLF achieves the best overall accuracy, outperforming HRLF (76.74%) and UMDF (75.56%), while maintaining competitive results under single- and bimodal conditions, demonstrating strong modality-specific representation learning and robust cross-modal interaction. On MOSEI, PRLF further achieves the highest average accuracy (76.24%) and the best full-modality performance (85.44%), confirming its superior adaptability to both complete and missing modality scenarios. Similarly, as shown in Table 2, PRLF also achieves the best or comparable performance under inter-modality missing conditions on the SIMS dataset, with the highest average accuracy of 81.19%. Compared with strong baselines such as LNLN (80.86%) and EMOE (80.12%), PRLF shows consistent improvements, particularly in settings ({l,a}, {l,v}, {a,v}). These results demonstrate that the PRLF framework enhances cross-modal alignment, improves inter-modal complementarity, and maintains strong robustness under missing modality conditions.

Robustness to Intra-modality Missingness. In this subsection, we compare PRLF with several representative methods. We randomly drop modal features at ratios of $p \in \{0, 0.1, 0.2, ..., 0.9\}$ to simulate intra-modal missingness in Fig. 4. PRLF consistently outperforms other methods on all three datasets, with particularly significant advantages under high missing ratios, demonstrating stronger robustness to intra-modal missingness. Although the F1 scores of all methods decrease as missingness increases, PRLF shows the slowest performance degradation, indicating its superior ability to exploit residual information. When the missing ratio reaches 0.9, PRLF still achieves F1 scores of 60 on MOSI and 70 on MOSEI, both substantially higher than TETFN, highlighting its robustness under extreme conditions. The detailed results are reported in supplementary Tables 9 and 10, and Table 11. Fusion-oriented methods such as EMOE perform well at low missing ratios but degrade rapidly as missingness increases, while

Table 3: Iteration Step Ablation results for the intra-modality missingness on MOSI.

Iteration	Missing rate p									
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$Step = 1$	81.14	79.69	75.44	72.25	71.09	69.58	64.71	60.73	59.91	55.37
$Step = 2$	83.08	80.39	76.87	73.88	72.34	70.65	65.82	62.06	61.15	56.39
$Step = 3$	85.21	81.29	**78.06**	75.28	73.96	71.97	67.75	64.62	62.23	57.59
$Step = 4$	**85.78**	**81.65**	77.99	**75.89**	**74.93**	**73.05**	**68.5**	**65.31**	**63.09**	**58.64**
$Step = 5$	84.49	81.08	77.29	75.19	73.78	71.24	67.27	64.16	62.33	57.48



(a) Module Ablation. **(b)** Loss Ablation.

Fig. 5: Ablation for the intra-modality missingness on MOSI.

incomplete-data methods like LNLN are more robust under severe missingness yet still inferior to PRLF. By dynamically estimating modality reliability via FIM and CC, PRLF maintains stable performance across all missing rates. On MOSI, MOSEI, and SIMS, the averaged F1 scores of PRLF, EMOE, and LNLN are (**72.48**, 71.21, 71.25), (**78.82**, 74.60, 75.65), and (**75.04**, 71.25, 73.13), respectively, further confirming PRLF's consistent overall advantage across varying levels of intra-modal missingness.

4.3 Ablation Studies

Iteration Step Ablation on Inter- and Intra-modality Missingness. As shown in Table 3, 4, the experiments on both inter-modal and intra-modal missing scenarios exhibit a consistent trend: as the number of iterations increases from 1 to 4, the model performance steadily improves, reaching its peak at Step = 4. This indicates that multi-step progressive interaction effectively enhances feature alignment and strengthens the representations of incomplete or missing modalities by leveraging guidance from the dominant modality, particularly under high missing rates or partial modality inputs. However, when the number of iterations increases to 5, performance begins to decline, suggesting that excessive interaction steps may impair the model's generalization ability. Therefore, Step = 4 achieves the optimal balance between representational power and stability, confirming the effectiveness and rationality of the proposed progressive fusion strategy for handling diverse missing-modality conditions.

Module Ablation on Inter- and Intra-modality Missingness. As shown in Table 5 and Fig. 5a, experiments on both inter- and intra-modality missing scenarios consistently demonstrate that the CMI, FIMI, ANRE, and PI are indispensable components of the PRLF framework. Removing the PI module leads to the most pronounced performance degradation across all settings, particularly under high intra-modality missing rates. This highlights its crucial role in

Table 4: Iteration Step Ablation results for the inter-modality missingness on MOSI.

Iteration	Testing Conditions							
	{l}	{a}	{v}	{l,a}	{l,v}	{a,v}	Avg.	{l,a,v}
$Step = 1$	79.62	64.33	59.68	80.47	80.06	69.57	72.29	81.14
$Step = 2$	81.49	66.87	62.34	82.65	82.33	72.64	74.72	83.08
$Step = 3$	82.91	68.73	63.11	84.36	**84.27**	75.58	76.49	85.21
$Step = 4$	**83.82**	**69.63**	**64.05**	**84.98**	84.13	**76.03**	**77.11**	**85.78**
$Step = 5$	82.47	68.51	62.77	83.54	82.61	74.88	75.80	84.49

Table 5: Module Ablation for inter-modality missingness on MOSI. Confidence-based Modality Importance (CMI), Fisher Information-based Modality Importance (FIMI), Adaptive Modality Reliability Estimator (AMRE), Progressive Interaction (PI).

Models	Testing Conditions							
	{l}	{a}	{v}	{l,a}	{l,v}	{a,v}	Avg.	{l,a,v}
PRLF	**83.82**	**69.63**	**64.05**	**84.98**	**84.13**	**76.03**	**77.11**	**85.78**
w/o CMI	82.97	68.54	63.21	83.74	83.25	74.98	76.12	85.07
w/o FIMI	83.04	68.77	63.25	83.41	82.99	75.06	76.09	84.88
w/o AMRE	81.92	67.58	61.95	82.36	82.15	73.79	74.96	83.05
w/o PI	78.48	63.53	58.76	79.66	79.01	68.49	71.32	79.33

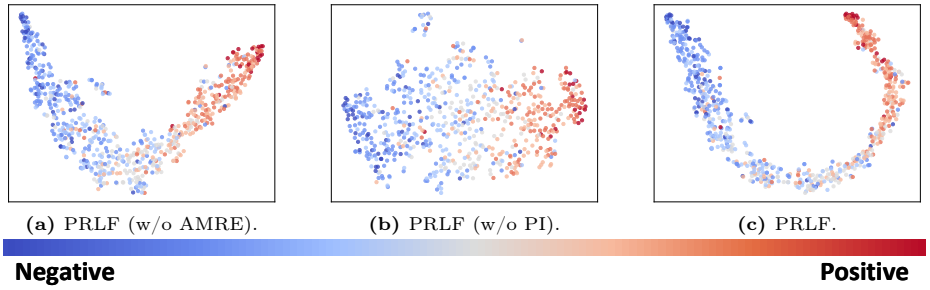
cross-modal alignment and fusion through iterative optimization and dominant-modality guidance. Eliminating the AMRE module results in consistent performance drops, as the model loses the ability to dynamically identify and exploit the most reliable modality for fusion—underscoring the superiority of adaptive weighting over static strategies. The CMI and FIMI serve as complementary mechanisms for assessing modality reliability, with FIMI showing slightly greater importance, especially under degraded input conditions. Notably, the substantial performance gaps observed in bimodal and full-modality settings further confirm that PI and AMRE are essential for optimizing multimodal collaboration and enhancing robustness against both partial and complete modality failures.

Loss Ablation on Inter- and Intra-modality Missingness. As shown in Table 6 and Fig. 5b, removing either the feature uniformity loss ($\mathcal{L}_{uni}$) or the phase consistency loss ($\mathcal{L}_{phase}$) results in consistent performance degradation under both inter- and intra-modality missingness. $\mathcal{L}_{uni}$ encourages stable feature representations across different missing conditions, while $\mathcal{L}_{phase}$ preserves feature directions among modalities. Their combination enables PRLF to better resist noise and retain discriminative structures, especially under high missing ratios.

Visualization of Feature Distribution. As illustrated in Fig. 6, AMRE evaluates modality reliability and promotes more compact feature distributions. The clustering variances with and without AMRE are 1.205 and 1.367, respectively, indicating more pronounced convergence when AMRE is applied. Compared with the complete model, the ablated variants exhibit clear distribution discrepancies. Without AMRE, the clusters become noticeably less compact, reflecting the model's inability to accurately estimate modality reliability and suppress unreliable features. Removing the PI module further results in blurred class boundaries, suggesting insufficient cross-modal alignment. In contrast, the full PRLF framework produces more compact and semantically consistent feature distributions that align well with emotion intensity. These results demonstrate

Table 6: Loss ablation on MOSI (inter-modality missingness).

Models	Testing Conditions							
	{l}	{a}	{v}	{l,a}	{l,v}	{a,v}	Avg.	{l,a,v}
PRLF	**83.82**	**69.63**	**64.05**	**84.98**	**84.13**	**76.03**	**77.11**	**85.78**
w/o $\mathcal{L}_{uni}$	82.64	68.32	62.88	83.07	82.98	74.51	75.73	84.81
w/o $\mathcal{L}_{phase}$	83.07	69.24	63.58	84.75	83.89	75.26	76.63	85.49



(a) PRLF (w/o AMRE). **(b)** PRLF (w/o PI). **(c)** PRLF.

Fig. 6: T-SNE visualization on MOSI. Default: intra-modality missingness ($p = 0.5$).

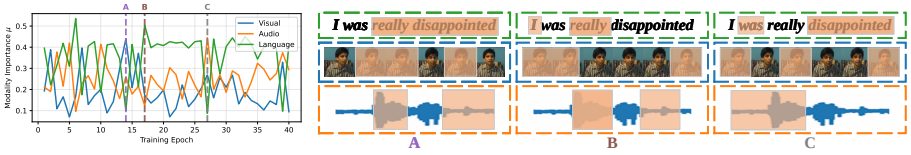


Fig. 7: Modality importance μ changes (left) and different weights for the same sample under varying missing data (right).

that AMRE mitigates noise from unreliable modalities, while PI progressively enhances cross-modal semantic alignment under missing-modality conditions.

Dynamic Modality Importance under Varying Missing Conditions. Fig. 7 visualizes the dynamic evolution of the modality importance weight μ and presents representative examples under different missing conditions. It can be observed that modality reliability varies significantly across different missing scenarios. The proposed method effectively evaluates the validity of each modality in an adaptive manner and accordingly selects the dominant modality to guide the subsequent progressive interaction and feature alignment process.

5 Conclusion

We presented PRLF, a progressive representation learning framework that addresses the challenges of multimodal sentiment analysis under uncertain missing-modality conditions. By integrating an Adaptive Modality Reliability Estimator (AMRE) and a Progressive Interaction (ProgInteract) mechanism, PRLF dynamically evaluates modality reliability and progressively aligns cross-modal representations. This design suppresses noise from incomplete modalities and enhances semantic consistency across views. Experiments on multiple benchmarks demonstrate that PRLF delivers reliable emotion recognition and stable cross-modal fusion under various missing-modality conditions, advancing the robust Multimodal Sentiment Analysis.

Acknowledgements

This work was supported by the Basic Research Program of Jiangsu under Grant No. BK20253028; the National Natural Science Foundation of China (NSFC) under Grant Nos. 62176124, U24A20330, and 62361166670.

References

1. Achille, A., Rovere, M., Soatto, S.: Critical learning periods in deep networks. In: International conference on learning representations (2018)
2. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE winter conference on applications of computer vision (WACV). pp. 1–10. IEEE (2016)
3. Chen, C., Liu, D., Xu, C.: Towards memorization-free diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8425–8434 (2024)
4. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: Covarep—a collaborative voice analysis repository for speech technologies. In: 2014 ieee international conference on acoustics, speech and signal processing (icassp). pp. 960–964. IEEE (2014)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
6. Fang, Y., Huang, W., Wan, G., Su, K., Ye, M.: Emoe: Modality-specific enhanced dynamic emotion experts. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14314–14324 (2025)
7. Guo, Z., Jin, T., Xu, W., Lin, W., Wu, Y.: Bridging the gap for test-time multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 16987–16995 (2025)
8. Han, W., Chen, H., Poria, S.: Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. arXiv preprint arXiv:2109.00412 (2021)
9. Hazarika, D., Zimmermann, R., Poria, S.: Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM international conference on multimedia. pp. 1122–1131 (2020)
10. Hou, J., Omar, N., Tiun, S., Saad, S., He, Q.: Tf-bert: Tensor-based fusion bert for multimodal sentiment analysis. Neural Networks **185**, 107222 (2025)
11. Hu, X., Tai, Y., Zhao, X., Zhao, C., Zhang, Z., Li, J., Zhong, B., Yang, J.: Exploiting multimodal spatial-temporal patterns for video object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3581–3589 (2025)
12. Hu, X., Zhong, B., Liang, Q., Shi, L., Mo, Z., Tai, Y., Yang, J.: Adaptive perception for unified visual multi-modal object tracking. IEEE Transactions on Artificial Intelligence (2025)
13. Huang, C., Chen, J., Huang, Q., Wang, S., Tu, Y., Huang, X.: Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. Information Fusion **114**, 102725 (2025)

14. Huang, C., Wei, Y., Yang, Z., Hu, D.: Adaptive unimodal regulation for balanced multimodal information acquisition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25854–25863 (2025)
15. Jiang, X., He, L., Gao, F., Zhang, K., Li, J., Gao, X.: Boosting modal-specific representations for sentiment analysis with incomplete modalities. IEEE Transactions on Multimedia (2025)
16. Li, M., Zhu, Z., Li, K., Pei, H.: Diversity and balance: multimodal sentiment analysis using multimodal-prefixed and cross-modal attention. IEEE Transactions on Affective Computing (2024)
17. Li, M., Yang, D., Lei, Y., Wang, S., Wang, S., Su, L., Yang, K., Wang, Y., Sun, M., Zhang, L.: A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 10074–10082 (2024)
18. Li, M., Yang, D., Liu, Y., Wang, S., Chen, J., Wang, S., Wei, J., Jiang, Y., Xu, Q., Hou, X., et al.: Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. Advances in Neural Information Processing Systems **37**, 28515–28536 (2024)
19. Li, M., Yang, D., Zhao, X., Wang, S., Wang, Y., Yang, K., Sun, M., Kou, D., Qian, Z., Zhang, L.: Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12458–12468 (2024)
20. Li, S., Zhang, T., Chen, C.P.: Sia-net: Sparse interactive attention network for multimodal emotion recognition. IEEE Transactions on Computational Social Systems **11**(5), 6782–6794 (2024)
21. Li, Y., Lan, X., Chen, H., Lu, K., Jiang, D.: Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications and Applications **20**(9), 1–23 (2025)
22. Li, Y., Ding, H., Lin, Y., Feng, X., Chang, L.: Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. Artificial Intelligence Review **57**(4), 78 (2024)
23. Lian, Z., Chen, L., Sun, L., Liu, B., Tao, J.: Gcnet: Graph completion network for incomplete multimodal learning in conversation. IEEE Transactions on pattern analysis and machine intelligence **45**(7), 8419–8432 (2023)
24. Liang, P.P., Zadeh, A., Morency, L.P.: Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys **56**(10), 1–42 (2024)
25. Liu, J., Yuan, H., Lu, X.M., Wang, X.: Quantum fisher information matrix and multiparameter estimation. Journal of Physics A: Mathematical and Theoretical **53**(2), 023001 (2020)
26. Liu, Z., Cai, L., Yang, W., Liu, J.: Sentiment analysis based on text information enhancement and multimodal feature fusion. Pattern Recognition **156**, 110847 (2024)
27. Luo, W., Xu, M., Lai, H.: Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In: International conference on multimedia modeling. pp. 411–422. Springer (2023)
28. Luo, Y., Liu, W., Sun, Q., Li, S., Li, J., Wu, R., Tang, X.: Triagedmsa: Triaging sentimental disagreement in multimodal sentiment analysis. IEEE Transactions on Affective Computing (2025)
29. Mao, H., Yuan, Z., Xu, H., Yu, W., Liu, Y., Gao, K.: M-sena: An integrated platform for multimodal sentiment analysis. arXiv preprint arXiv:2203.12441 (2022)

30. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
31. Shi, P., Hu, M., Nakagawa, S., Zheng, X., Shi, X., Ren, F.: Text-guided reconstruction network for sentiment analysis with uncertain missing modalities. IEEE Transactions on Affective Computing (2025)
32. Sun, H., Wang, H., Liu, J., Chen, Y.W., Lin, L.: Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In: Proceedings of the 30th ACM international conference on multimedia. pp. 3722–3729 (2022)
33. Sun, N., Zhang, W., Zheng, W., Liu, J., Chai, L., Wu, C.: Robust multimodal sentiment analysis based on adaptive information distillation and adversarial learning. IEEE Transactions on Affective Computing (2025)
34. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the conference. Association for computational linguistics. Meeting. vol. 2019, p. 6558 (2019)
35. Wang, H., Ren, C., Yu, Z.: Multimodal sentiment analysis based on multiple attention. Engineering Applications of Artificial Intelligence **140**, 109731 (2025)
36. Wang, P., Zhou, Q., Wu, Y., Chen, T., Hu, J.: Dlf: Disentangled-language-focused multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 21180–21188 (2025)
37. Wang, W., Ding, L., Shen, L., Luo, Y., Hu, H., Tao, D.: Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 2282–2291 (2024)
38. Wang, Y., Cui, Z., Li, Y.: Distribution-consistent modal recovering for incomplete multimodal learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22025–22034 (2023)
39. Wu, S., He, D., Wang, X., Wang, L., Dang, J.: Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1601–1609 (2025)
40. Yan, M., Qian, J., Bao, J., Yang, J.: Fevermamba: Highlight fever in crowd via mamba for remote fever screening. IEEE Journal of Biomedical and Health Informatics (2025)
41. Yang, D., Li, M., Xiao, D., Liu, Y., Yang, K., Chen, Z., Wang, Y., Zhai, P., Li, K., Zhang, L.: Towards multimodal sentiment analysis debiasing via bias purification. In: European Conference on Computer Vision. pp. 464–481. Springer (2024)
42. Yang, Z., Jiang, Z., Li, X., Zhou, H., Dong, J., Zhang, H., Du, Y.: D^4-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. In: ECCV. pp. 36–52. Springer (2024)
43. Yang, Z., Li, Y., He, S., Li, X., Xu, Y., Dong, J., Du, Y.: Omnivton: Training-free universal virtual try-on. In: ICCV. pp. 16702–16711 (2025)
44. Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., Yang, K.: 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. pp. 3718–3727 (2020)
45. Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 10790–10797 (2021)

46. Yuan, Z., Fang, J., Xu, H., Gao, K.: Multimodal consistency-based teacher for semi-supervised multimodal sentiment analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024)
47. Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.P.: Memory fusion network for multi-view sequential learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
48. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems **31**(6), 82–88 (2016)
49. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: 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). pp. 2236–2246 (2018)
50. Zhang, H., Wang, W., Yu, T.: Towards robust multimodal sentiment analysis with incomplete data. Advances in Neural Information Processing Systems **37**, 55943–55974 (2024)
51. Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., Yu, T.: Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 756–767. Association for Computational Linguistics (2023)
52. Zhu, A., Hu, M., Wang, X., Yang, J., Tang, Y., An, N.: Proxy-driven robust multimodal sentiment analysis with incomplete data. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 22123–22138 (2025)
53. Zhu, L., Zhao, H., Zhu, Z., Zhang, C., Kong, X.: Multimodal sentiment analysis with unimodal label generation and modality decomposition. Information Fusion **116**, 102787 (2025)
54. Zhuang, Y., Zhang, Y., Hu, Z., Zhang, X., Deng, J., Ren, F.: Glomo: Global-local modal fusion for multimodal sentiment analysis. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1800–1809 (2024)