Sample-specific Modality Diagnosis and Cross-modal Enhancement for Incomplete Multimodal Representations

Junsong Chen[1], Jiyuan Liu[1*], Suyuan Liu[1], Wei Zhang[1], Ao Li[2*], En Zhu[1*], Xinwang Liu[1]

[1]National University of Defense Technology
[2]Harbin University of Science and Technology
{chenjunsong257, liujiyuan13, suyuanliu, zhangwei23, enzhu, xinwangliu}@nudt.edu.cn, ao.li@hrbust.edu.cn

Abstract

In multimodal sentiment analysis, modality missingness and quality degradation are common. Existing methods often rely on batch-level modality generation, generation but neglect sample-level missingness, hence their flexibility is limited severely in real-world scenarios. To address this, Sample-specific Modality Diagnosis and Cross-modal Enhancement for Incomplete Multimodal Representations (SMCIR) is proposed. Specifically, The Dynamic Multi-feature Fusion Detector (DMFD) is presented, which detects missingness and severity at the sample-level using indicators such as information entropy, modality similarity, and mutual information. Unlike batch-based methods, the DMFD provides fine-grained detection and adaptive responses, improving sensitivity to modality disturbances. Meanwhile, the Context-aware Modality Completion Generator (CMCG) is developed to restore missing modalities through context-guided reconstruction using multiscale feature fusion and cross-modal attention. In this way, the proposed CMCG method can avoid redundancy and inconsistency, enhancing the consistency and discriminativity of the fused representation. In CMCG, the text modality serves as a stable guide to improve context consistency. Experiments on the CMU-MOSI and CMU-MOSEI datasets show that SMCIR outperforms existing full-modal and non-recovery-based methods, well validating its efficacy and superiority in multimodal learning.

Code — https://github.com/js257/SMCIR

Introduction

With the rapid development of multimodal learning, research is shifting from single-modality modeling (Yang et al. 2024b,c) to efficient multimodal fusion (Yang et al. 2024a; Huang et al. 2025). Multimodal learning leverages data from multiple modalities, such as text, vision, and audio, to improve model performance and robustness across tasks. Current research (Zhu et al. 2024; Zhuang et al. 2024; Huang et al. 2024) focuses on modeling modality dependencies and designing cross-modal fusion mechanisms for more accurate task inference. To address modality missingness in multimodal data, various modality completion and feature recovery methods have been proposed. However, they face two main limitations:

First, most existing approaches detect missing modalities at the batch level rather than on a sample-by-sample basis. They often rely on predefined missing patterns (Li et al. 2024b; Wang, Li, and Cui 2023) or prior missing labels (Ke et al. 2025; Ma et al. 2021a), limiting their flexibility in handling sample-specific modality uncertainty. In addition, some GNN-based methods (Lian et al. 2023; Wang et al. 2024) assume a uniform missing pattern across all samples in a batch—for example, if the text modality is available, the visual and audio modalities are regarded as missing for all. Such assumptions overlook heterogeneous missingness, reducing both fusion quality and downstream performance. Second, recent non-recovery strategies—such as graph-based models (Lian et al. 2023), Bayesian inference (Ma et al. 2021b), and prompt learning frameworks (Guo, Jin, and Zhao 2024)—enhance robustness to incomplete modalities. However, they generally rely on static modeling and lack dynamic awareness of modality state changes, limiting their ability to capture complementary relationships and collaborative dynamics among modalities under real-world degradation or missingness.

To address these issues, we propose the explicit Dynamic Multi-Feature Fusion Detector (DMFD), which directly identifies the missing modality type and severity for each sample without prior labels or rules. Compared to methods using masks or global inference, the DMFD offers several advantages: **(i)** it detects the modality missingness state for each sample, improving accuracy; **(ii)** it performs personalized processing based on the modality missingness state, enhancing flexibility in complex environments; **(iii)** it incorporates statistical metrics, such as entropy, modality mutual information, and similarity, improving robustness and interpretability. This provides a more flexible and generalizable solution for handling modality missingness in multimodal tasks. Additionally, the Context-aware Modality Completion Generator (CMCG) is presented, utilizing multi-scale feature fusion and refined attention to enhance traditional recovery methods. Local and global contextual relationships between modality features are refined to ensure accurate and consistent recovery. Unlike traditional global recovery strategies (Le et al. 2025), context learning is employed for adaptive modality generation, with recovery treated as semantic reconstruction. This allows for more accurate recovery of ambiguous or partially missing modal-

*Corresponding authors.

ities, avoiding redundancy and significantly enhancing generalization and robustness in multimodal tasks. The contributions can be summarized as follows:

- We design the DMFD to automatically identify the missing modality type for each sample, without prior guidance or assumptions about missing patterns, effectively filling the gap in sample-level modality detection.
- To address partially missing or ambiguous modalities, we propose the CMCG, which combines multi-scale feature modeling and refined attention. This approach refines local and global context for adaptive recovery, avoiding redundancy and inconsistency in traditional methods, thus improving accuracy and consistency in modality reconstruction.
- Ablation experiments and visualizations on two benchmark datasets evaluate the performance of the SMCIR against various state-of-the-art methods in scenarios with full modalities, partial missingness, and others.

Related Work

Incomplete Modality Recognition

Modality missingness is a key challenge in multimodal learning, weakening model performance and introducing inconsistencies that impact task outcomes (Wu et al. 2024). Existing research typically simulates random missingness, where modality data is lost due to noise, sensor failures, or environmental changes. Some studies (Wang et al. 2023b; Liu et al. 2024; Sun et al. 2024b) adopted varying missing rates, patterns, and dropout strategies to assess model robustness, but predefined rules limit real-world generalization. Missingness detection, which identifies missing modalities and their severity, better aligns with real-world needs. However, research in multimodal emotion analysis remains limited. Detection methods include mask-based, similarity-based, traditional machine learning, and deep generative models. For instance, Zhang (Zhang et al. 2022) proposed a task-guided modality-adaptive similarity metric for inferring missing modalities based on relationships between non-missing modalities. Zhang (Zhang et al. 2025) constructed modality missingness masks to identify missing locations. Traditional machine learning methods, including BP neural networks and K-means, are also applied (Yan et al. 2021). Recently, deep generative models such as autoencoders (Tran et al. 2017) and GANs (Wu et al. 2024) have gained popularity due to their strong modeling capabilities.

Incomplete Modality Recovery

Existing research mainly focuses on reconstruction-based, distillation enhancement, and adversarial generative methods. Reconstruction-based methods use autoencoders to model the relationship between available and missing modalities, reconstructing missing features. For example, Li (Li et al. 2024a) adopted stacked residual autoencoders to approximate complete modality information, and Tao (Tao et al. 2025) adopted encoder-decoder structures for latent representations of missing modalities. However, these methods may lose individual variation by fully replacing missing features. To preserve semantic consistency, knowledge distillation transfers knowledge from complete data to scenarios with missing modalities, aiding cross-modal representation learning (Li et al. 2024c; Wang et al. 2023c). However, they heavily rely on supervision signals from complete modalities and paired samples, limiting adaptability in complex or incomplete scenarios. Cross-modal generative methods have gained attention, using related available modalities to complete missing ones (Li et al. 2024d; Kang et al. 2024). However, these methods struggle with ambiguous, partially lost, or misaligned missing modalities in real-world scenarios. Relying solely on generative methods may reduce accuracy and consistency, making it crucial to fully utilize available modality information to ensure quality and accuracy.

Methodology

Preliminaries. The goal of the multimodal sentiment analysis task is to predict sentiment polarity (such as positive, negative, or neutral) or sentiment intensity (such as valence and arousal along a continuous dimension) based on given multimodal inputs. Three modalities are considered: language (L), vision (V), and audio (A). These modalities are represented as two-dimensional tensors $X_m \in \mathbb{R}^{N_m \times d_m}$, where N_m is the sequence length, d_m is the embedding dimension, and $m \in \{L, V, A\}$ represents the different modalities.

Model Overview. The proposed SMCIR, as shown in Figure 1, aims to address the issue of modality missingness in multimodal learning. First, the three modalities are projected to the same feature dimension $\mathcal{P}_m \in \mathbb{R}^{B \times N_m \times d_{hidden}}$, where B is the batch size. Then, by calculating the entropy H_m, mutual information $MI(\mathcal{P}_m, \mathcal{P}_L)$, and modality similarity $S_{\mathcal{P}_m \mathcal{P}_L}$, the weighted score $\mathcal{W}_m^{(i)}$ is obtained to measure the degree of modality missingness. Based on this information, a discriminative method is employed to determine if a modality is missing. If a modality is missing, conditional information C^i is generated using a cross-modal mechanism, and the MSSE is used to further enhance the feature expression. Through temporal and global context modeling, the generated conditional information is fused with the original missing modality features to obtain the enhanced modality feature $\mathcal{G}^i$. Otherwise, the current modality is output. Finally, the generated samples from the entire batch are concatenated with the original samples, resulting in the final multimodal representation $\hat{O}$, which is used for emotion prediction.

Dynamic Multi-feature Fusion Detector (DMFD)

Existing missingness detection methods face three key limitations: predefined rules (Wang et al. 2023b; Sun et al. 2024b), coarse granularity (Zhang et al. 2022), and heavy supervision (Rojas et al. 2025). To address these, we propose DMFD, an unsupervised, sample-level method that detects missingness by jointly analyzing intra-modal distribution properties and inter-modal correlations.

Multi-Metric Fusion Score Calculation. A weighted score calculation method based on multi-metric fusion is proposed to assess modality missingness. By combining

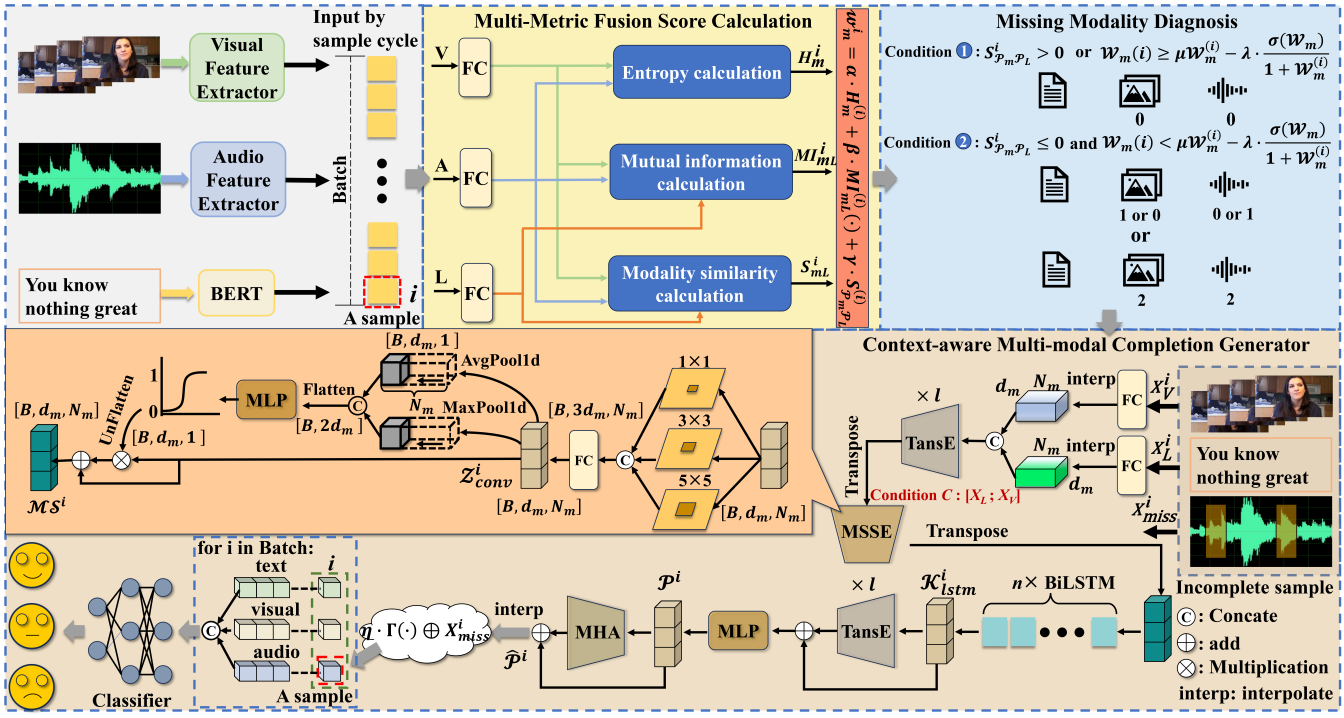

Figure 1: The SMCIR structure. It covers the entire process from multi-metric fusion score calculation (entropy, mutual information, and modality similarity), missing modality diagnosis, and cross-modal generation for missing modalities to sample-level modality fusion. An example of audio modality missing is shown.

multiple information metrics, the missingness and severity of each modality can be evaluated. Specifically, the text, visual, and audio features are first mapped to the same dimension, resulting in projected features $\mathcal{P}_m \in \mathbb{R}^{B\times N_m\times d_{hidden}}, m \in \{L, V, A\}$, where d_{hidden} is the mapped dimension. The weighted score is then calculated by evaluating modality relationships through three key metrics.

Entropy Calculation (H): Entropy serves as an effective measure for modality missingness. Higher entropy suggests more information, while lower entropy indicates potential missingness or sparsity. By calculating the entropy of each modality, we can identify missing modalities for compensation. For visual or audio modalities, we use a softmax-based entropy method. First, we compute the ℓ_2-norm of each feature vector to get the feature "size," then smooth this with a temperature parameter τ. The probability distribution p_k for each modality feature is computed as follows:

$$p_k = \frac{\exp\left(\|\mathcal{P}_{m,i,k}\|_2/\tau\right)}{\sum_{j=1}^{N_m}\exp\left(\|\mathcal{P}_{m,i,j}\|_2/\tau\right)}, \tag{1}$$

Where $\|\mathcal{P}_{m,i,k}\|_2$ denotes the ℓ_2-norm of the k-th modality feature of the i-th sample, and τ is the temperature coefficient. Then, the entropy H_i is calculated as the negative log expectation of the probability distribution.

$$H_i = -\sum_{k=1}^{N_m} p_k \log p_k. \tag{2}$$

Mutual Information Calculation (MI): In multimodal learning, mutual information measures the correlation and information sharing between modalities. To quantify this, we adopt a histogram discretization-based approximation method. Given a modality $\mathcal{P}_m$, we first discretize the features into fixed bins to obtain the joint and marginal probability distributions. Mutual information is then computed from these distributions. The formula is as follows:

$$MI(\mathcal{P}_m, \mathcal{P}_L) = \sum_{\mathfrak{k}_1=1}^{\mathfrak{K}}\sum_{\mathfrak{k}_2=1}^{\mathfrak{K}} \hat{p}(\mathfrak{k}_1,\mathfrak{k}_2)\log\frac{\hat{p}(\mathfrak{k}_1,\mathfrak{k}_2)}{\hat{p}(\mathfrak{k}_1)\hat{p}(\mathfrak{k}_2)} \tag{3}$$

Where $\mathfrak{K}$ represents the number of bins, $\hat{p}(\mathfrak{k}_1,\mathfrak{k}_2)$ is the empirical distribution of the joint features between modality $\mathcal{P}_m$ and text modality $\mathcal{P}_L$, and $\hat{p}(\mathfrak{k}_1)$ and $\hat{p}(\mathfrak{k}_2)$ are the marginal distributions, respectively.

Modality Similarity Calculation (S): Given two input modalities, $\mathcal{P}_m$ and $\mathcal{P}_L$, we first align their temporal dimensions to the target length N_L using adaptive average pooling. Then, for the i-th sample's pooled features $\tilde{\mathcal{P}}_m^{(i)}$ and $\tilde{\mathcal{P}}_L^{(i)}$, with shape $\mathbb{R}^{N_L\times d_{hidden}}$, we compute the cosine similarity at each time step. For each time step n, the cosine similarity $s_n^{(i)}$ is computed using the following formula:

$$s_n^{(i)} = \frac{\tilde{\mathcal{P}}_{m,n}^{(i)}\cdot\tilde{\mathcal{P}}_{L,n}^{(i)}}{\max\left(\|\tilde{\mathcal{P}}_{m,n}^{(i)}\|_2\cdot\|\tilde{\mathcal{P}}_{L,n}^{(i)}\|_2, \epsilon\right)} \tag{4}$$

Where $\tilde{\mathcal{P}}_{m,n}^{(i)}$ and $\tilde{\mathcal{P}}_{L,n}^{(i)}$ are the feature vectors of modality $\mathcal{P}_m$ and text modality $\mathcal{P}_L$ at time step n after pooling, and

$\|\tilde{\mathcal{P}}_{m,n}^{(i)}\|_2$ and $\|\tilde{\mathcal{P}}_{L,n}^{(i)}\|_2$ are their L_2-norms. To ensure numerical stability and avoid division by zero errors, we set $\epsilon = 10^{-8}$.

Finally, the modality similarity score is computed by averaging the cosine similarity across all time steps. For the i-th sample, the final modality correlation score $S_{\mathcal{P}_m\mathcal{P}_L}^{(i)}$ is computed as follows:

$$S_{\mathcal{P}_m\mathcal{P}_L}^{(i)} = \frac{1}{N_L}\sum_{n=1}^{N_L} s_n^{(i)} \tag{5}$$

To comprehensively consider the three metrics, including entropy, modality similarity, and mutual information, we adopt a weighted sum to compute the final missingness detection score for each sample. The weight coefficients α, β, and γ, which are fixed parameters, control the influence of each metric on the final score. The final missingness detection score $\mathcal{W}_m^{(i)}$ is calculated as follows:

$$\mathcal{W}_m^{(i)} = \alpha \cdot H_m^{(i)} + \beta \cdot MI(\mathcal{P}_m, \mathcal{P}_L)^{(i)} + \gamma \cdot S_{\mathcal{P}_m\mathcal{P}_L}^{(i)} \tag{6}$$

Where $H_m^{(i)}$ is the entropy value of sample i in modality $\mathcal{P}_m$, $MI(\mathcal{P}_m, \mathcal{P}_L)^{(i)}$ is the mutual information between modality $\mathcal{P}_m$ and text modality $\mathcal{P}_L$, and $S_{\mathcal{P}_m\mathcal{P}_L}^{(i)}$ is the modality correlation score between modality $\mathcal{P}_m$ and text modality $\mathcal{P}_L$.

Missing Modality Diagnosis. For modality $m \in \{V, A\}$, the modality missingness state $\hbar_m^{(i)}$ of the sample i is determined by the following decision rule:

$$\hbar_m^{(i)} = \begin{cases} 1, & \text{if } S_{\mathcal{P}_m\mathcal{P}_L}^{(i)} \le 0 \text{ and } \mathcal{W}_m^{(i)} < \mu_m - \lambda \cdot \frac{\sigma_m}{1+\mathcal{W}_m^{(i)}} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Where '0' and '1' represent the presence and absence of a modality, respectively. In this method, we focus on the scenario where the modality is missing. If $S_{\mathcal{P}_m\mathcal{P}_L}^{(i)} \le 0$ and $\mathcal{W}_m^{(i)}$ is below a certain threshold, defined by $\mu_m - \lambda \cdot \frac{\sigma_m}{1+\mathcal{W}_m^{(i)}}$, we classify the modality as missing ($\hbar_m^{(i)} = 1$). In all other cases, the modality is considered to exist ($\hbar_m^{(i)} = 0$). $\mathcal{W}_m^{(i)}$ is the feature quality score of modality m. $\mu_m = \frac{1}{B}\sum_{i=1}^{B} \mathcal{W}_m^{(i)}, \sigma_m = \sqrt{\frac{1}{B}\sum_{j=1}^{B}\left(\mathcal{W}_m^{(j)} - \mu_m\right)^2}$. λ is the hyperparameter that regulates the sensitivity of the threshold.

If both the visual and audio modalities are detected as missing (i.e., $\hbar_V^{(i)} = 1 \wedge \hbar_A^{(i)} = 1$), the state is updated to indicate a joint missing condition:

$$\hbar_V^{(i)} \leftarrow 2, \quad \hbar_A^{(i)} \leftarrow 2 \tag{8}$$

Context-aware Multi-modal Completion Generator (CMCG)

Next, we perform the completion or enhancement of the detected missing modalities. First, we identify the available modalities, with modality features represented as $X_m^i \in \mathbb{R}^{N_m \times d_m}$ and $X_L^i \in \mathbb{R}^{N_L \times d_L}$. When only the text modality is available, and both the audio and visual modalities are missing, the text features X_L^i do not require any transformation or fusion and are directly outputted. When both the text modality and other modalities (such as audio or visual) are available, the feature alignment function projects the input features to the same dimension through a linear transformation, followed by feature fusion.

$$E_m^i = W_o \cdot \text{TansE}\left(\left[W_L X_L^i \parallel \mathcal{I}(W_m X_m^i, N_L)\right]\right) \tag{9}$$

Where TansE represents the Transformer Encoder (Vaswani et al. 2017) with l layers, W_m and W_L are linear transformation matrices that map input features to the same dimension, and $\mathcal{I}(W_m X_m^i, N_L)$ aligns the features of modality m with the sequence length N_L of modality L. $\parallel$ denotes concatenation along the channel dimension, and W_o is the learned matrix used to output the aligned features. After transposition, E_m^i serves as the conditional information $C^i \in \mathbb{R}^{1\times d_L \times N_L}$, which is provided to the Multi-Scale Attention Enhancement Module (MSSE) to enhance the features of the missing modality. The structure of MSSE is inspired by multi-scale approaches (Chen et al. 2018) and attention mechanisms (Hu, Shen, and Sun 2018).

Different sizes of convolution kernels (k) are employed to model local and global contextual information from the input features. The outputs of the three convolution kernels with different sizes are concatenated and processed through a fully connected layer to produce $\mathcal{Z}_{\text{conv}}^i \in \mathbb{R}^{1\times d_L \times N_L}$. The corresponding output is defined as:

$$\Psi_k(C^i) = \text{Conv}_{k\times k}(C^i), \quad k \in \{1, 3, 5\} \tag{10}$$

$$\mathcal{Z}_{\text{conv}}^i = \text{FC}\left(\left[\Psi_1(C^i) \parallel \Psi_3(C^i) \parallel \Psi_5(C^i)\right]\right) \tag{11}$$

where Ψ_k represents the convolution operation and FC denotes the fully connected layer.

After multi-scale convolution, statistical features are extracted using max-pooling and average-pooling along the time dimension, and then concatenated along the channel dimension to extract global statistical information along the channel dimension, producing $\mathcal{Q}^i \in \mathbb{R}^{1\times 2d_L \times 1}$. To dynamically highlight key channel features, attention weights $\mathcal{A}^i$ are calculated using an MLP, defined as follows:

$$\mathcal{Q}^i = \left[\text{maxpool}_N(\mathcal{Z}_{\text{conv}}^i) \parallel \text{avgpool}_N(\mathcal{Z}_{\text{conv}}^i)\right] \tag{12}$$

$$\mathcal{A}^i = \sigma\left(W_2 \cdot \delta\left(\text{LN}\left(W_1 \cdot \text{vec}(\mathcal{Q}^i)\right)\right)\right) \tag{13}$$

where $\text{vec}(\cdot)$ denotes the flattening operation, δ represents the ReLU function, σ is the Sigmoid function, W_1 and W_2 are learnable parameters, and LN represents layer normalization. By combining the multi-scale convolution structure, dual pooling, and attention mechanism, the output of MSSE is given as:

$$\mathcal{MS}^i = \mathcal{Z}_{\text{conv}}^i \otimes \mathcal{A}^i \oplus \mathcal{Z}_{\text{conv}}^i \tag{14}$$

Where $\oplus$ denotes residual summation and $\otimes$ represents channel-wise multiplication. The final enhanced modality $\mathcal{G}^i$ is obtained through the following generation process. It is defined as follows:

$$\mathcal{G}^i = \eta \cdot \Gamma\left(\mathcal{MS}^i\right) \oplus X_{\text{miss}}^i \tag{15}$$

where η is the fusion weight and Γ denotes the core cross-modal modeling module. This generator fuses multiscale enhanced features with missing modality information for feature completion.

The Core Transformation Γ integrates temporal modeling and global context-aware capabilities. First, the features $\mathcal{MS}^i \in \mathbb{R}^{1\times N_L\times d_L}$ generated by MSSE are normalized and input into BiLSTM with *n* layers to capture temporal dependencies, producing output features $\mathcal{K}^i_{\text{lstm}}$. These features are then passed to TansE, where global feature correlations are modeled using self-attention, resulting in $\mathcal{K}^i_{\text{trans}}$. The features $\mathcal{K}^i_{\text{lstm}}$ and $\mathcal{K}^i_{\text{trans}}$ are fused using concatenation or residual addition, creating the joint representation $\mathcal{K}^i_{\text{fused}}$. This is linearly transformed to obtain intermediate features $\mathcal{P}^i$, which are input into the Multi-Head Attention Mechanism (MHA) for parallel subspace attention optimization, yielding the final representation $\hat{\mathcal{P}}^i$.

$$\mathcal{K}^i_{\text{lstm}} = \text{BiLSTM}\left(\text{LN}\left(\mathcal{MS}^i\right)\right) \tag{16}$$

$$\mathcal{K}^i_{\text{trans}} = \text{TansE}(\mathcal{K}^i_{\text{lstm}}) \tag{17}$$

$$\mathcal{K}^i_{\text{fused}} = \mathcal{K}^i_{\text{lstm}} \oplus \mathcal{K}^i_{\text{trans}} \tag{18}$$

$$\mathcal{P}^i = W_2 \cdot \delta(W_1 \cdot \mathcal{K}^i_{\text{fused}}) \tag{19}$$

$$\hat{\mathcal{P}}^i = \mathcal{I}(\text{LN}\left(\text{MHA}(\mathcal{P}^i) \oplus \mathcal{P}^i\right)) \tag{20}$$

The feature completion process handles both missing and non-missing modalities. Missing modalities are completed using the generator, while non-missing ones are preserved. The final output consists of the enhanced visual or audio modality, with the text modality unchanged. This process can be summarized as the following generation function:

$$F^{(i)}_{\text{out}} = \begin{cases} \mathcal{F}_{\mathcal{M}}\left(L^{(i)}, \mathcal{E}^{(i)}\right) \oplus X^{(i)}_m & \text{if } \hbar^{(i)}_m \in \{1,2\} \\ X^{(i)}_m & \text{otherwise} \end{cases} \tag{21}$$

$$\mathcal{E}^{(i)} = \begin{cases} A^{(i)} & \text{if } \mathcal{M} = V \\ V^{(i)} & \text{if } \mathcal{M} = A \end{cases} \tag{22}$$

Here, $\mathcal{F}_{\mathcal{M}}$ is the modality generation mapping function, $X^{(i)}_m$ represents the original modality features, which are directly preserved if the modality is not missing. Otherwise, the modality is completed by the generator. $L^{(i)}, V^{(i)}, A^{(i)}$ represent the modality features for the i-th sample, and $\mathcal{M} \in \{V, A\}$ is the identifier for the missing modality.

The features of the same modality in the batch are fused using a concatenation operation:

$$\hat{V}_{\text{out}} = \text{concat}\left(\{F^{(1)}_{\text{out}}[V], F^{(2)}_{\text{out}}[V], \ldots, F^{(B)}_{\text{out}}[V]\}\right) \tag{23}$$

$$\hat{A}_{\text{out}} = \text{concat}\left(\{F^{(1)}_{\text{out}}[A], F^{(2)}_{\text{out}}[A], \ldots, F^{(B)}_{\text{out}}[A]\}\right) \tag{24}$$

$$\hat{L}_{\text{out}} = \text{concat}\left(\{L_1, L_2, \ldots, L_B\}\right) \tag{25}$$

Finally, the three modalities are mapped to the same feature dimension using weight matrices $W_m, m \in \{L, V, A\}$ and aligned to the text modality's time steps N_L using interpolation operation $\mathcal{I}$. After processing through the fusion layer W_f, the final fused output $\hat{O}$ is obtained:

$$\hat{O} = \text{Fusion}(\hat{L}_{\text{out}}, \hat{V}_{\text{out}}, \hat{A}_{\text{out}}) = \begin{bmatrix} \hat{L}_{\text{out}} W_L \\ \mathcal{I}(\hat{V}_{\text{out}} W_V, N_L) \\ \mathcal{I}(\hat{A}_{\text{out}} W_A, N_L) \end{bmatrix}^{\top} W_f \tag{26}$$

Loss Function

The L1 loss is adopted as the task-specific loss function:

$$\mathcal{L}_{task} = \frac{1}{q}\sum_{i=1}^{q} |y_i - \hat{y}_i| \tag{27}$$

where q is the number of samples, y_i is the true label, and $\hat{y}_i$ is the predicted value.

For missing modality generation, we adopt an improved mean squared error to constrain generated features to resemble real modality features:

$$\mathcal{L}_{simse} = \frac{\left(\sum_{i=1}^{q}(r_i - \hat{r}_i)\right)^2}{q^2} \tag{28}$$

where r_i is the true modality feature, and $\hat{r}_i$ is the generated feature.

The final loss is a weighted sum of task loss and generation loss that:

$$\mathcal{L} = \mathcal{L}_{task} + \xi\mathcal{L}_{simse} \tag{29}$$

where ξ is a trade-off hyperparameter.

Experiment

Datasets and Evaluation Metrics

We evaluated the SMCIR model on two public benchmark multimodal sentiment analysis datasets:

MOSI: The MOSI dataset (Zadeh et al. 2016) consists of 2199 YouTube video segments with text, visual (facial expressions), and audio (speech) features, along with sentiment intensity labels in the range of [-3, 3].

MOSEI: The MOSEI dataset (Zadeh et al. 2018) contains over 23,000 YouTube video segments covering text, audio, and visual modalities. It provides sentiment intensity labels in the range of [-3, 3] and six emotion categories, making it suitable for both sentiment prediction and emotion classification tasks.

By following previous methods (Yang, Dong, and Qiang 2024) (Tao et al. 2025), the performance of the model on both datasets is evaluated using classification accuracy (Acc-2), weighted F1 score (F1), mean absolute error (MAE), and pearson correlation coefficient (Corr).

Feature Extraction

Language modality: We employ the pre-trained BERT model (Wolf et al. 2020) to encode raw text from the MOSI and MOSEI datasets, adding part-of-speech and word-level sentiment polarity embeddings for each sentence. **Visual Features:** The OpenFace (Baltrušaitis, Robinson, and Morency 2016) is employed to extract facial features from visual frames. Frames with open eyes are selected, and those with closed eyes are discarded to avoid facial expression analysis errors. **Acoustic Features:** The COVAREP (Degottex et al. 2014) is adopted to extract acoustic features, including Mel-frequency cepstral coefficients (MFCCs), pitch, and voicing features related to speech emotion and pitch.

Methods	CMU-MOSI				CMU-MOSEI			
	MAE	Corr	ACC-2	F1	MAE	Corr	ACC-2	F1
FDMER (Yang et al. 2022)	0.845	0.732	-/84.20	-/83.90	0.568	0.736	-/83.90	-/83.80
TETFN (Wang et al. 2023a)	0.717	0.800	**84.05**/86.10	83.83/86.07	0.551	0.748	84.25/85.18	84.18/85.27
EMT (Sun et al. 2024a)	0.705	0.798	83.30/85.00	83.20/85.00	0.527	0.774	83.40/86.00	83.70/86.00
MTMD (Lin and Hu 2024)	0.705	0.799	84.00/86.00	**83.90**/86.00	0.531	0.767	**84.80**/86.10	84.90/85.90
CLGSI (Yang, Dong, and Qiang 2024)	0.703	0.790	83.97/**86.43**	83.63/**86.25**	0.532	0.763	84.01/86.32	84.21/86.18
MFMB-Net (Tao et al. 2025)	0.709	0.798	82.70/85.70	83.20/86.00	0.532	0.758	84.70/85.10	**85.00**/85.10
DLF (Wang et al. 2025)	0.731	0.781	-/85.06	-/85.04	0.536	0.764	-/85.42	-/85.27
RGM-LoRA (Liu, Fu, and Wang 2025)	0.698	0.801	83.10/85.70	83.40/86.10	0.532	0.769	82.90/85.30	83.70/85.80
SMCIR	**0.696**	**0.804**	83.15/84.82	83.12/84.80	**0.523**	**0.777**	83.34/**86.35**	83.76/**86.35**

Table 1: Comparison with the full-modal methods based on the **BERT** language model on the CMU-MOSI and CMU-MOSEI datasets. F1 and Acc-2 are reported as negative/non-negative (left) and negative/positive (right). The best results are marked in bold, while the second best with underline.

Datasets	Methods	Testing Conditions				
		{L}	{L, V}	{L, A}	Avg	{L,V,A}
CMU-MOSI	SMIL	78.26	79.15	79.82	79.08	82.85
	GCNet	80.42	82.78	83.23	82.14	83.78
	MPLMM	79.52	80.12	80.48	80.03	82.96
	SMCIR	**83.18**	**84.12**	**83.40**	**83.57**	**84.36**
CMU-MOSEI	SMIL	77.46	78.36	77.89	77.90	81.56
	GCNet	81.35	83.43	82.96	82.58	83.42
	MPLMM	78.69	79.67	79.45	79.27	82.67
	SMCIR	**84.39**	**85.35**	**84.88**	**84.87**	**85.78**

Table 2: Comparison of modality generation ability. For example, "L" indicates that only the text modality is available when visual and audio modalities are missing. "Avg" represents the average performance across the three possible conditions. The evaluation metric is the F1 score.

Datasets	DMFD	MSSE	CMCG	MAE	Corr
CMU-MOSI	✓		✓	0.704	0.802
	✓	✓		0.705	0.800
		✓	✓	0.723	0.795
	✓	✓	✓	**0.696**	**0.804**
CMU-MOSEI	✓		✓	0.537	0.771
	✓	✓		0.530	0.771
		✓	✓	0.534	0.767
	✓	✓	✓	**0.523**	**0.777**

Table 3: Ablation study of the proposed DMFD, MSSE, and CMCG.

Implementation Details

On the MOSI and MOSEI datasets, SMCIR is optimized using the Adam optimizer with a batch size of 64. The learning rates are set to 2×10^{-5} and 1×10^{-5}, and the Adam epsilon values are 3×10^{-8} and 1×10^{-8}, respectively. The coefficients α, β, γ, and ξ are set to 0.5, 0.3, 10, and 0.2, while η is set to 0.6 and 0.4 for the two datasets. Both the l layers of *TransE* and the n layers of BiLSTM are set to 2 and 4, respectively. Non-reconstructive methods are trained using the same setup with publicly available code, and all experiments are conducted on an RTX 3090 GPU.

Comparison to State-of-the-art Methods

As shown in Table 1, under ideal conditions with complete modality input, the SMCIR outperforms full-modal methods, particularly on CMU-MOSI, setting new records across multiple metrics, demonstrating strong expressive and modeling capabilities. Existing methods such as TETFN, MTMD, and FDMER assume complete modality information and cannot handle missing or degraded modalities. In contrast, the SMCIR introduces a modality state perception mechanism, enabling dynamic identification and selective enhancement, improving adaptability and robustness in real-world scenarios. While methods such as EMT-DLFR and RGM-LoRA explore modality collaboration, they rely on static models, making sample-level modality perception challenging. SMCIR's "state perception + selective enhancement" strategy avoids redundant modeling, improving fusion quality and emotion recognition.

To evaluate SMCIR's generative ability with missing modalities, we compared it with three non-reconstructive methods (SMIL (Ma et al. 2021b), GCNet (Lian et al. 2023), and MPLMM (Guo, Jin, and Zhao 2024)) on the two datasets in Table 2. On CMU-MOSI, SMCIR outperforms all methods with an average score of 83.57%, excelling in missing modality detection and enhancement. On CMU-MOSEI, it leads with an average accuracy of 84.87%, surpassing other methods. Even with complete input, it achieves 85.78%, confirming its stability and generalizability. The regression results in three missing modality scenarios show SMCIR's superior performance in Figure 2, especially under the LV condition, where its predictions closely match the ideal, demonstrating better emotion recognition and generation. These results highlight SMCIR's advantage in handling missing modalities and emotional intensity.

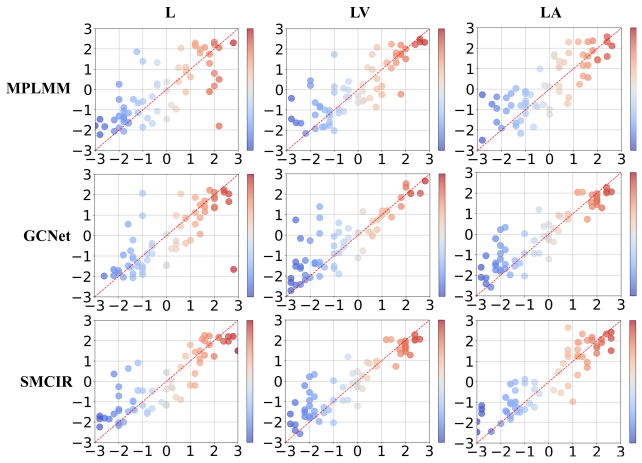

Figure 2: Visualization of regression performance of different methods under three missing conditions on the MOSI dataset. The varying depth of color represents the magnitude of emotional intensity (from -3 to 3).

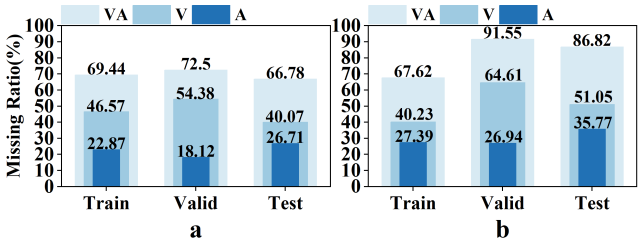

Figure 3: Modality missing rate detection: Figures a and b display the modality missing rate detection results for the MOSI and MOSEI datasets, respectively. For example, "A" in the legend indicates the missing rate of the audio modality in the total samples of both the visual and audio modalities.

Ablation Study

Structural contribution of SMCIR. To validate the contributions of DMFD, MSSE, and CMCG, we conduct an ablation study on both datasets (Table 3). The results show: (i) Using DMFD alone performs well in missing modality scenarios but is inferior to using all modules, indicating that combining DMFD with others enhances robustness and performance. (ii) Using MSSE alone slightly reduces performance, especially MAE on CMU-MOSI, suggesting that MSSE improves feature expression but redundancy and inconsistency remain without other modules. (iii) Removing CMCG causes a significant drop, particularly on CMU-MOSI, showing that CMCG is essential for modality completion and context-guided feature reconstruction. (iv) Using all modules together confirms that the combined approach substantially improves robustness and accuracy across various missing modality scenarios.

Structural contribution of DMFD. Table 4 shows the results of different indicator combinations in the DMFD model. The single indicator *H* outperforms both *M* and *S* in classification and regression tasks. Among two-indicator combinations, *HM* improves regression but slightly reduces classification performance, while *MS* offers no significant gain. The three-indicator combination (*HMS*) provides the best performance, highlighting the complementary benefits of combining diverse indicators.

Methods	CMU-MOSEI			
	MAE	Corr	ACC-2	F1
H	0.529	0.773	**83.51**/86.19	**83.95**/86.16
M	0.533	0.770	83.14/85.39	83.48/85.36
S	0.532	0.769	83.27/85.69	83.55/85.64
HM	0.527	0.772	83.19/85.39	83.42/85.28
MS	0.528	0.771	83.01/85.20	83.36/85.07
HMS	**0.523**	**0.777**	83.34/**86.35**	83.76/**86.35**

Table 4: A study on the effectiveness of different combinations of indicators in the proposed DMFD.

Datasets	Available	MAE	Corr	ACC-2	F1
CMU-MOSI	{V⋆, A⋆}	**0.696**	**0.804**	83.15/**84.82**	83.12/**84.80**
	{V, A⋆}	0.709	0.792	**83.24**/84.43	**83.23**/84.54
	{V⋆, A}	0.726	0.788	83.09/84.46	83.04/84.48
CMU-MOSEI	{V⋆, A⋆}	**0.523**	**0.777**	**83.34/86.35**	**83.76/86.35**
	{V, A⋆}	0.529	0.771	82.93/85.69	83.37/85.64
	{V⋆, A}	0.539	0.763	81.36/84.84	81.92/84.80

Table 5: Ablation experiment of DMFD.The text modality is always available. ⋆ indicates the modality where DMFD is applied.

Ablation results of DMFD. The ablation results in Table 5 confirm the effectiveness of DMFD. Applying it to both visual and audio modalities yields the best performance on both datasets, demonstrating that the dual-modality collaborative enhancement effectively handles fuzzy features. Individually, the detection module benefits the audio modality more, especially on CMU-MOSEI, likely due to richer discriminative features. In contrast, the visual modality often identifies more ambiguous samples (Figure 3), leading to outputs that rely on cross-modal generation and perform worse than real samples.

Conclusion

This paper proposes SMCIR, a sample-level modality diagnosis and cross-modal enhancement framework, which addresses modality missing and information imbalance in multimodal data through the DMFD and CMCG. Experimental results show that SMCIR significantly improves performance on the CMU-MOSI and CMU-MOSEI datasets. The framework dynamically enhances each sample, improving the recognition and handling of ambiguous modalities. It demonstrates good scalability, offering a novel solution for multimodal tasks. Future work could optimize module interaction and sample-level processing efficiency, which remains a key challenge. More efficient modality generation schemes are needed to further enhance performance on large-scale datasets.

Acknowledgments

This work was supported in part by the National Science and Technology Innovation 2030 Major Project under (No. 2022ZD0209103); in part by the National Natural Science Foundation of China(No. 62476281, 62306324, 62376279, U24A20333); in part by the National Natural Science Foundation of China Joint Found (No. U24A20323); in part by the Science and Technology Innovation Program of Hunan Province (No. 2024RC3128); and in part by the National University of Defense Technology Research Foundation (No. ZK24-30).

References

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.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848.

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.

Guo, Z.; Jin, T.; and Zhao, Z. 2024. Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition. arXiv:2407.05374.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, C.; Chen, J.; Huang, Q.; Wang, S.; Tu, Y.; and Huang, X. 2025. AtCAF: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114: 102725.

Huang, J.; Zhou, J.; Tang, Z.; Lin, J.; and Chen, C. Y.-C. 2024. TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowledge-Based Systems*, 285: 111346.

Kang, M.; Zhu, R.; Chen, D.; Liu, X.; and Yu, W. 2024. CM-GAN: A Cross-Modal Generative Adversarial Network for Imputing Completely Missing Data in Digital Industry. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3): 2917–2926.

Ke, G.; He, S.; Wang, X.; Wang, B.; Chao, G.; Zhang, Y.; Xie, Y.; and Su, H. 2025. Knowledge Bridger: Towards Training-Free Missing Modality Completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 25864–25873.

Le, H. Q.; Thwal, C. M.; Qiao, Y.; Tun, Y. L.; Nguyen, M. N.; Huh, E.-N.; and Hong, C. S. 2025. Cross-modal prototype based multimodal federated learning under severely missing modality. *Information Fusion*, 122: 103219.

Li, J.; Cai, S.; Li, L.; Sun, R.; Yuan, G.; and Zhu, R. 2024a. MIT-FRNet: Modality-invariant temporal representation learning-based feature reconstruction network for missing modalities. *Expert Systems with Applications*, 249: 123655.

Li, M.; Yang, D.; Liu, Y.; Wang, S.; Chen, J.; Wang, S.; Wei, J.; Jiang, Y.; Xu, Q.; Hou, X.; Sun, M.; Qian, Z.; Kou, D.; and Zhang, L. 2024b. Toward Robust Incomplete Multimodal Sentiment Analysis via Hierarchical Representation Learning. arXiv:2411.02793.

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

Li, Y.; Zheng, C.; Zuo, R.; and Lu, W. 2024d. Semantic Reconstruction Guided Missing Cross-modal Hashing. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Lian, Z.; Chen, L.; Sun, L.; Liu, B.; and Tao, J. 2023. GCNet: Graph Completion Network for Incomplete Multimodal Learning in Conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8419–8432.

Lin, R.; and Hu, H. 2024. Multi-Task Momentum Distillation for Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*, 15(2): 549–565.

Liu, F.; Fu, Z.; and Wang, Y. 2025. Reward-Based Gradient Modulation for Multimodal Emotion Recognition With LoRA. *IEEE Transactions on Computational Social Systems*, 1–11.

Liu, R.; Zuo, H.; Lian, Z.; Schuller, B. W.; and Li, H. 2024. Contrastive Learning Based Modality-Invariant Feature Acquisition for Robust Multimodal Emotion Recognition With Missing Modalities. *IEEE Transactions on Affective Computing*, 15(4): 1856–1873.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021a. SMIL: Multimodal Learning with Severely Missing Modality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3): 2302–2310.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021b. SMIL: Multimodal Learning with Severely Missing Modality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3): 2302–2310.

Rojas, K.; Zhu, Y.; Zhu, S.; Ye, F. X. F.; and Tao, M. 2025. Diffuse Everything: Multimodal Diffusion Models on Arbitrary State Spaces. arXiv:2506.07903.

Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2024a. Efficient Multimodal Transformer With Dual-Level Feature Restoration for Robust Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*, 15(1): 309–325.

Sun, Y.; Liu, Z.; Sheng, Q. Z.; Chu, D.; Yu, J.; and Sun, H. 2024b. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 110: 102454.

Tao, C.; Li, J.; Zang, T.; and Gao, P. 2025. A Multi-Focus-Driven Multi-Branch Network for Robust Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2): 1547–1555.

Tran, L.; Liu, X.; Zhou, J.; and Jin, R. 2017. Missing Modalities Imputation via Cascaded Residual Autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, D.; Guo, X.; Tian, Y.; Liu, J.; He, L.; and Luo, X. 2023a. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136: 109259.

Wang, H.; Chen, Y.; Ma, C.; Avery, J.; Hull, L.; and Carneiro, G. 2023b. Multi-Modal Learning With Missing Modality via Shared-Specific Feature Modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15878–15887.

Wang, H.; Ma, C.; Zhang, J.; Zhang, Y.; Avery, J.; Hull, L.; and Carneiro, G. 2023c. Learnable Cross-modal Knowledge Distillation for Multi-modal Learning with Missing Modality. volume 14223 LNCS, 216 – 226.

Wang, P.; Zhou, Q.; Wu, Y.; Chen, T.; and Hu, J. 2025. DLF: Disentangled-Language-Focused Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20): 21180–21188.

Wang, Y.; Li, Y.; and Cui, Z. 2023. Incomplete Multimodality-Diffused Emotion Recognition. In *Advances in Neural Information Processing Systems*, volume 36, 17117–17128. Curran Associates, Inc.

Wang, Y.; Yao, X.; Zhu, P.; Li, W.; Cao, M.; and Hu, Q. 2024. Integrated Heterogeneous Graph Attention Network for Incomplete Multi-modal Clustering. *INTERNATIONAL JOURNAL OF COMPUTER VISION*, 1–20.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Wu, Q.; Lu, T.; Ding, Z.; Shi, L.; Xu, Y.; and Xu, X. 2024. Automatic Completion Method of Power Missing Data in Distribution Network Based on GCN-VAE Model. In *2024 China International Conference on Electricity Distribution (CICED)*, 1117–1121.

Wu, R.; Wang, H.; Chen, H.-T.; and Carneiro, G. 2024. Deep Multimodal Learning with Missing Modality: A Survey. *arXiv e-prints*, arXiv:2409.07825.

Yan, A.; Wang, W.; Ren, Y.; and Geng, H. 2021. A Clustering Algorithm for Multi-Modal Heterogeneous Big Data With Abnormal Data. *Frontiers in Neurorobotics*, Volume 15 - 2021.

Yang, D.; Huang, S.; Kuang, H.; Du, Y.; and Zhang, L. 2022. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 1642–1651. ISBN 9781450392037.

Yang, D.; Kuang, H.; Yang, K.; Li, M.; and Zhang, L. 2024a. Towards Asynchronous Multimodal Signal Interaction and Fusion via Tailored Transformers. *IEEE Signal Processing Letters*, 31: 1550–1554.

Yang, D.; Yang, K.; Kuang, H.; Chen, Z.; Wang, Y.; and Zhang, L. 2024b. Towards Context-Aware Emotion Recognition Debiasing From a Causal Demystification Perspective via De-Confounded Training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10663–10680.

Yang, D.; Yang, K.; Li, M.; Wang, S.; Wang, S.; and Zhang, L. 2024c. Robust Emotion Recognition in Context Debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12447–12457.

Yang, Y.; Dong, X.; and Qiang, Y. 2024. CLGSI: A Multimodal Sentiment Analysis Framework based on Contrastive Learning Guided by Sentiment Intensity. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2099–2110.

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. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.

Zhang, C.; Chu, X.; Ma, L.; Zhu, Y.; Wang, Y.; Wang, J.; and Zhao, J. 2022. M3Care: Learning with Missing Modalities in Multimodal Healthcare Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 2418–2428. ISBN 9781450393850.

Zhang, Y.; Zhang, Y.; Gao, M.; Wang, X.; Dai, B.; and Shen, W. 2025. Multimodal behavior recognition for dairy cow digital twin construction under incomplete modalities: A modality mapping completion network approach. *Artificial Intelligence in Agriculture*, 15(3): 459–469.

Zhu, A.; Hu, M.; Wang, X.; Yang, J.; Tang, Y.; and Ren, F. 2024. KEBR: Knowledge Enhanced Self-Supervised Balanced Representation for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 5732–5741. ISBN 9798400706868.

Zhuang, Y.; Zhang, Y.; Hu, Z.; Zhang, X.; Deng, J.; and Ren, F. 2024. GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 1800–1809. ISBN 9798400706868.