This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Conflict-Aware Adaptive Cross-Reconstruction for Multimodal Sentiment Analysis

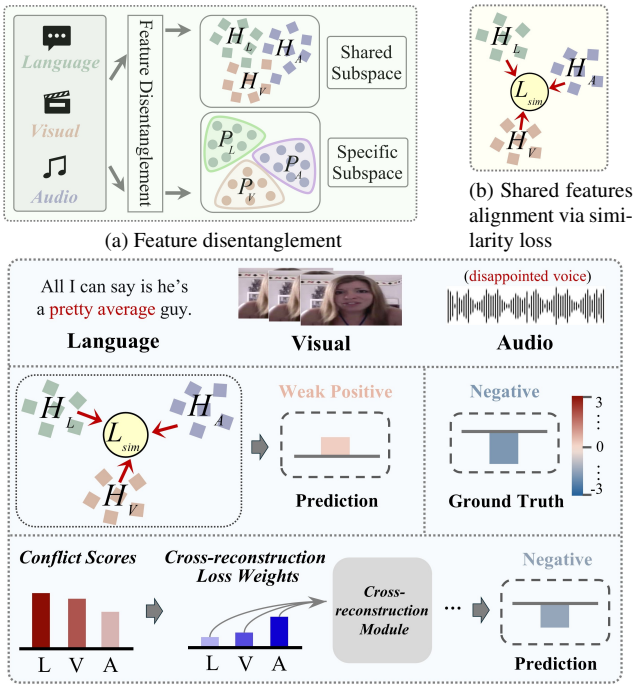
Yan Wang Fuyuan Cao* Xingwang Zhao
School of Computer and Information Technology, Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, China
{wangyan17, cfy, zhaoxw}@sxu.edu.cn

Abstract

*Disentanglement-based methods for learning shared representations are widely used in multimodal sentiment analysis. However, existing methods often overlook potential emotional conflicts across modalities within the same sample. They typically adopt intra-modal reconstruction and rely on similarity losses to align shared representations, which may distort the shared semantics under such conflicts. To address these, we propose a **C**onflict-aware **A**daptive **C**ross-**R**econstruction approach (**CACR**). First, we formally define emotional conflict and design a conflict-aware weighting strategy. This strategy calculates conflict scores for each modality and maps them to the corresponding cross-reconstruction weights. Second, we construct a cross-reconstruction module. It reconstructs each target modality using its own specific features and the shared features of other modalities, achieving implicit alignment of shared representations. Coupled with the aforementioned weights, this module effectively suppresses conflicting modalities and mitigates semantic ambiguity. Furthermore, we develop a fine-grained sentiment refinement module that captures fine-grained cues from visual- and audio-specific features to supplement textual semantics. Extensive experiments on three datasets show that CACR outperforms existing state-of-the-art methods, demonstrating its effectiveness in handling emotional conflict.*

1. Introduction

Multimodal Sentiment Analysis (MSA) aims to characterize human emotions [20, 21] from heterogeneous sequences such as language, visual, and audio modalities, showing great potential in applications including mental health assessment [3] and human-computer interaction [6]. Although numerous methods have made considerable progress in addressing the inherent heterogeneity across modalities, the potential emotional conflicts within the same sample remain an unresolved bottleneck limiting the accuracy of sentiment prediction. How to learn consistent shared representations while suppressing conflicting information remains a core challenge in multimodal learning.



(a) Feature disentanglement

(b) Shared features alignment via similarity loss

(c) Similarity-based method vs. our CACR

Figure 1. Motivation of the proposed CACR approach. (a) Feature disentanglement. (b) Similarity-based alignment of shared representations (adapted from DLF [25]). (c) Example of a sample with emotional conflict, where similarity-based alignment (middle) distorts the shared semantic, while our conflict-aware weighting (bottom) down-weights the conflicting modality and yields a correct prediction.

*Corresponding author.

Existing approaches can be broadly categorized into two groups: one focuses on building robust multimodal representations [10, 22, 28, 32], while the other is dedicated to designing cross-modal fusion mechanisms [15, 27, 29]. Within the representation-learning group, disentanglement-based methods [13, 25, 28] decompose features into shared and specific subspaces to reduce information redundancy (Figure 1a). However, these methods implicitly assume that different modalities of the same sample maintain consistent emotional polarity, thereby overlooking potential emotional conflicts across modalities within the same sample. Under this assumption, they typically reconstruct each modality using only its own shared and specific features, enforcing alignment in the shared subspace via similarity losses (Figure 1b). Figure 1c provides an example: in the shared subspace, the language modality is in clear conflict with the visual and audio modalities in sentiment polarity. Conventional similarity-based alignment pulls these contradictory features together, ultimately leading to an incorrect emotion prediction (middle of Figure 1c).

In this work, we propose a Conflict-aware Adaptive Cross-Reconstruction approach (CACR) to suppress conflicting modalities and mitigate semantic ambiguity of the shared representations (bottom of Figure 1c). We first disentangle the features of each modality into shared and specific subspaces. Second, we formally define emotional conflict within the shared subspace and design a conflict-aware weighting strategy (CAW) that calculates sample-level conflict scores and maps them to the corresponding cross-reconstruction loss weights for each modality. Based on this, we construct a cross-reconstruction module. Guided by the aforementioned weights, CRM reconstructs the target modality by integrating its own specific features with the shared features of other modalities. This adaptively weighted reconstruction process not only achieves implicit alignment of the shared representations but also effectively suppresses the influence of conflicting modalities. Furthermore, considering that the language modality typically dominates sentiment expression while visual and audio signals provide complementary cues, we develop a fine-grained sentiment refinement (FSR) module to enrich textual semantics with these emotional cues. Finally, we fuse the shared representations derived from the cross-reconstructed features with the enhanced specific features for sentiment prediction. The main contributions can be summarized as follows:

- We formally define emotional conflict and propose a conflict-aware weighting strategy to mitigate semantic distortions of the shared representations
- We design a cross-reconstruction module to implicitly align shared representations, coupled with the conflict-aware weighting strategy to suppress the influence of conflicting modalities on the reconstruction process.
- We introduce a fine-grained sentiment refinement module that captures emotional cues from visual- and audio-specific features to supplement textual semantics, thereby enhancing the representation of fine-grained sentiments.

2. Related Work

2.1. Multimodal Sentiment Analysis

Unlike unimodal sentiment analysis [1, 19] that relies solely on single-modality data, multimodal sentiment analysis (MSA) integrates heterogeneous modalities, including the language, visual, and audio modalities, to capture richer human sentiment. Current MSA methods can be roughly categorized into representation learning approaches [7, 13, 25] and fusion strategy methods [15, 29, 34]. Representation learning methods aim to produce unified multimodal representations by extracting high-level semantics. For example, Self-MM [32] enhances cross-modal mutual information via self-supervised contrastive and masked learning. DEVA [27] textualizes visual and audio inputs and integrates them into the language branch, thereby augmenting language representations. In contrast, fusion methods concentrate on effectively integrating multimodal information. Tensor Fusion Network (TFN)[34] is a representative model that progressively fuses multimodal information through Cartesian product operations. Subsequently, MMIM [9] improves multimodal fusion through hierarchical mutual information maximization, and ALMT [36] treats language as the dominant modality to capture information from auxiliary modalities and produce fused representations. However, many of these approaches adopt uniform processing across all modalities, which may introduce redundancy. This has motivated the development of disentanglement-based representation learning methods to address these issues.

2.2. Disentangled Representation Learning

Disentangled representation learning aims to separate multimodal features into shared and specific subspaces to reduce information redundancy. For instance, FDMER [28] combines metric learning with adversarial learning to construct modality-invariant and modality-private subspaces, thereby enhancing multimodal representations. DMD [13] disentangles feature spaces and employs a dynamic graph distillation mechanism, achieving more flexible cross-modal knowledge transfer. DLF [25] transfers visual and audio information into the language representations, effectively improving overall performance. However, these methods typically overlook potential emotional conflicts across modalities within the same sample, which limits the accuracy of sentiment prediction. Our proposed CACR incorporates a dynamic conflict-aware weighting strategy into the cross-modal reconstruction process. This design achieves implicit alignment of shared representations while

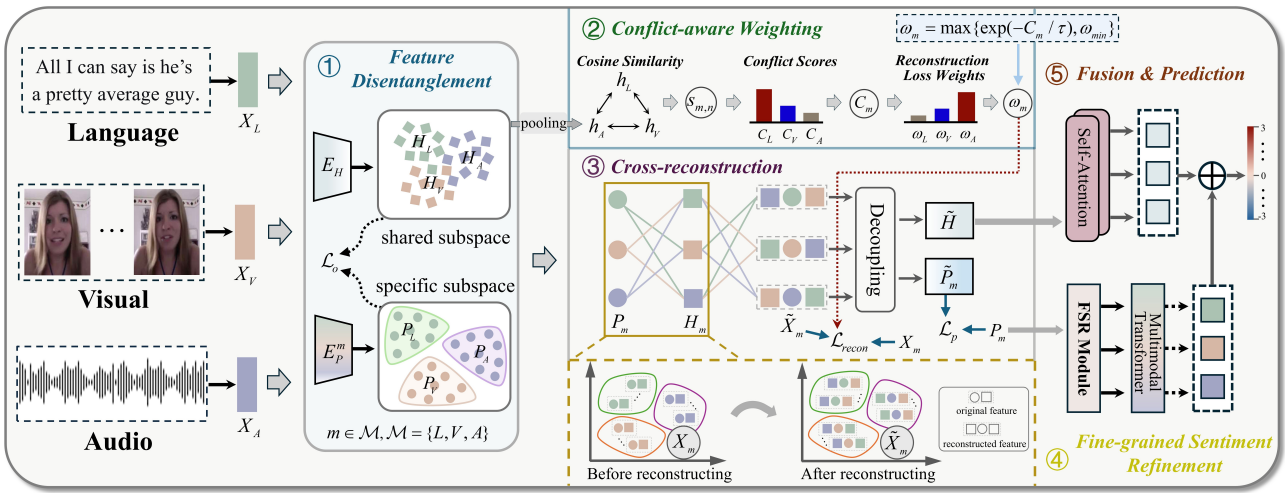


Figure 2. **Architecture illustration of CACR.** ① After unimodal feature extraction, unimodal features are decomposed into shared and specific subspaces. ② The Conflict-aware Weighting strategy calculates emotional conflict scores for each modality and maps them to cross-reconstruction loss weights. ③ The Cross-reconstruction module then reconstructs each target modality from its own specific features together with the shared features of the other modalities. This process achieves implicit alignment of shared representations while utilizing the aforementioned weights to adaptively suppress conflicting modalities. ④ The Fine-grained Sentiment Refinement module extracts fine-grained emotional cues from visual- and audio-specific features to supplement and refine the textual semantics. ⑤ The refined shared and modality-specific representations are fused to produce the final sentiment prediction.

adaptively suppressing the influence of conflicting modalities, thereby alleviating semantic misalignment.

3. The Proposed Method

Overview. To address the issues of existing methods that overlook emotional conflicts and rely on similarity constraints, we propose the Conflict-aware Adaptive Cross-Reconstruction (CACR) approach. The overall architecture of CACR is illustrated in Figure 2. First, CACR decomposes multimodal features into modality-shared and modality-specific subspaces to minimize information redundancy (Section 3.1). Subsequently, we formally define emotional conflict within the shared subspace and design a Conflict-Aware Weighting (CAW) strategy that calculates sample-level conflict scores and maps them to the corresponding cross-reconstruction loss weights (Section 3.2). Then, a Cross-Reconstruction Module (CRM) is employed to reconstruct each modality by combining its own specific features with the shared features of other modalities, achieving implicit alignment without relying on similarity-based constraints. Coupled with the CAW strategy, this module effectively suppresses the influence of conflicting modalities (Section 3.3). Further, the Fine-grained Sentiment Refinement (FSR) module extracts emotional cues from visual- and audio-specific features to supplement the language branch (Section 3.4). Finally, the shared representations obtained via CRM are fused with the refined specific features to perform hierarchical prediction (Section 3.5).

3.1. Feature Disentanglement

Given a multimodal dataset with modalities $\mathcal{M} = \{L, V, A\}$, we first extract unimodal features via dedicated encoders and unify their dimensions using a 1D convolutional layer. For the i-th sample, the sequence representation of each modality m is denoted as $X_m^{(i)} \in \mathbb{R}^{T_m \times d}$, where $m \in \mathcal{M}$, T_m is the sequence length, and d is the unified dimension. For simplicity, we omit the sample index (i) in subsequent formulations—except when calculating the conflict score—and denote the representation as X_m.

Then, we employ one modality-shared encoder and three modality-specific encoders to decompose the multimodal representations into modality-shared and -specific features:

$$H_m = E_H\left(X_m\right), P_m = E_P^m\left(X_m\right), \tag{1}$$

where E_H is the shared encoder to capture cross-modal consistency semantics, and E_P^m is the specific encoder for the m-th modality to preserve specific information. The resulting H_m and P_m are used in subsequent modules.

To further refine the decoupling between the two subspaces, we introduce a soft orthogonality loss $\mathcal{L}_o$ to reduce the redundant information between the shared and specific features. It is defined as

$$\mathcal{L}_o = \sum_{m \in \mathcal{M}} \left|\cos(H_m, P_m)\right|, \tag{2}$$

where $|\cos(\cdot,\cdot)|$ denotes the non-negative value of the cosine similarity between H_m and P_m. Minimizing this term

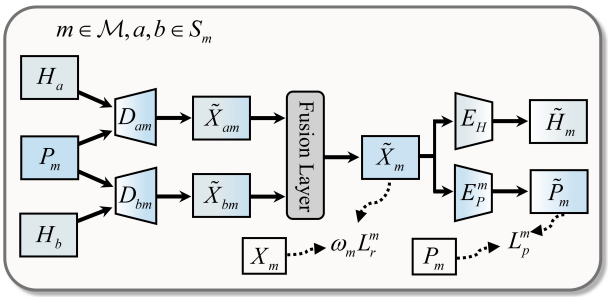


Figure 3. The architecture of CRM.

effectively enhances the orthogonality of the two subspaces.

3.2. Conflict-aware Weighting Strategy

To avoid treating inherent modality heterogeneity as conflict, we formally define emotional conflict and explicitly measure it within the shared feature space rather than raw or specific feature spaces. To this end, we compute sample-level conflict scores and design a Conflict-Aware Weighting (CAW) strategy. This strategy maps these scores to corresponding cross-reconstruction loss weights, thereby effectively mitigating the negative impact of conflicting modalities throughout the optimization process.

Definition (Emotional Conflict). Within the shared subspace, different modalities of the same sample express contradictory emotional semantics.

For a target modality $m \in \mathcal{M}$, we define its complementary set as $S_m = \mathcal{M} \setminus \{m\}$. For the i-th sample, temporal pooling is applied to the shared features $H_m^{(i)}$ to obtain a global vector $h_m^{(i)}$. The conflict score between the target modality m and any modality $n \in S_m$ is computed based on their cosine similarity $s_{m,n}^{(i)}$:

$$s_{m,n}^{(i)} = \cos(h_m^{(i)}, h_n^{(i)}), d_{m,n}^{(i)} = 1 - s_{m,n}^{(i)}. \tag{3}$$

Accordingly, the overall conflict score for the target modality m is computed by averaging its scores across all complementary modalities:

$$C_m^{(i)} = \frac{1}{|S_m|} \sum_n d_{m,n}^{(i)}, \tag{4}$$

where $C_m^{(i)}$ denotes the divergence between the target modality m and the other modalities n. Based on $C_m^{(i)}$, the weight of the corresponding cross-reconstruction term is computed via an exponential decay function:

$$\hat{\omega}_m^{(i)} = \exp(\frac{-C_m^{(i)}}{\tau}), \omega_m^{(i)} = \max\{\hat{\omega}_m^{(i)}, \omega_{min}\}, \tag{5}$$

where $\tau > 0$ controls the decay rate and ω_{min} is the lower bound, with their sensitivity analysis in Section 4.5.

These computed weights $\omega_m^{(i)}$ are then used to adaptively modulate the cross-reconstruction loss in Section 3.3.

3.3. Cross-reconstruction with Adaptive Weights

As illustrated in Figure 3, giving a sample, a cross-reconstruction module (CRM) is employed for each target modality $m \in \mathcal{M}$.

First, the specific features P_m of the target modality m are concatenated with the shared features H_n from the other modalities $n \in S_m$ and passed through a cross-reconstruction decoder D to generate cross-modal interaction features:

$$\tilde{X}_{nm} = D_{nm}\big([H_n \oplus P_m]\big). \tag{6}$$

These features are then fused to form the reconstructed feature of the target modality m:

$$\tilde{X}_m = F\big(\tilde{X}_{nm}\big), \tag{7}$$

where D_{nm} denotes a 1D convolutional decoder for cross-reconstruction, $\oplus$ denotes concatenation, and F is a linear fusion layer.

To effectively suppress the influence of conflicting modalities during the reconstruction process, we incorporate the weights derived in Section 3.2 into the reconstruction objective. Let $\ell_m^{(i)} = \|\tilde{X}_m^{(i)} - X_m^{(i)}\|^2$ denote the per-sample cross-reconstruction loss. The final weighted cross-reconstruction loss $\mathcal{L}_m$ for the target modality m and the overall reconstruction loss $\mathcal{L}_{recon}$ are formulated as

$$\mathcal{L}_m = \frac{\sum_{i=1}^N \omega_m^{(i)} \ell_m^{(i)}}{\sum_{i=1}^N \omega_m^{(i)}}, \mathcal{L}_{recon} = \sum_m \mathcal{L}_m, \tag{8}$$

where N is the number of samples. This mechanism enables the model to dynamically handle varying degrees of emotional conflict. Specifically, **for low-conflict modalities** ($C_m \to 0$)**:** $\omega_m \to 1.0$, which preserves full reconstruction constraints and encourages the shared features of other modalities n to align with the target modality m. **For high-conflict modalities:** ω_m decays exponentially, effectively reducing the contribution of conflicting modalities to the reconstruction loss, thus explicitly suppressing their negative impact.

Subsequently, the same disentanglement operation described in Section 3.1 is applied to $\tilde{X}_m$ to extract the shared and specific features of the reconstructed representation:

$$\tilde{H}_m = E_H\left(\tilde{X}_m\right), \tilde{P}_m = E_P^m\left(\tilde{X}_m\right), \tag{9}$$

where $\tilde{H}_m$ and $\tilde{P}_m$ denote the shared and specific features estimated from the cross-reconstructed features, respectively. We further concatenate the shared features of all modalities to form $\tilde{H} = [\tilde{H}_L; \tilde{H}_V; \tilde{H}_A]$. Through this adaptively weighted cross-reconstruction, shared semantics are implicitly aligned across modalities without relying on similarity constraints.

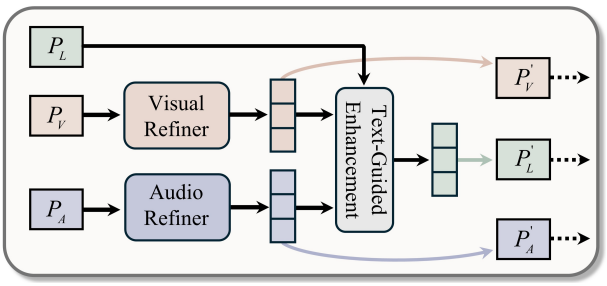


Figure 4. The details of the FSR.

Finally, to preserve the distinctiveness of each modality, we introduce a specific loss $\mathcal{L}_p$, defined as

$$\mathcal{L}_p = \sum_m \| \tilde{P}_m - P_m \|^2 . \tag{10}$$

This loss term ensures that specific features do not degrade during the reconstruction process.

3.4. Fine-grained Sentiment Refinement

Unlike methods that adopt mutual cross-modal attention [25], we consider that the language modality typically dominates sentiment semantics, while visual and audio modalities provide complementary fine-grained cues. Therefore, we propose a Fine-grained Sentiment Refinement (FSR) module that captures fine-grained emotional cues (e.g., transient frowning) from visual- and audio-specific features to supplement textual semantics.

The detailed architecture of FSR is illustrated in Figure 4. Specifically, FSR first extracts fine-grained features P'_V and P'_A from P_V and P_A. Subsequently, using the language feature P_L as a query, a text-guided enhancement layer selectively aggregates relevant emotional cues from P'_V and P'_A. These cues are then fused with P_L to output the enhanced feature P'_L. This process enriches the language representation with sentiment cues, thereby enhancing its capability to express fine-grained emotional semantics.

3.5. Fusion and Prediction

Multimodal Fusion. To capture global contextual interactions among the refined features, CACR employs a multimodal Transformer layer [23] to enhance both intra- and inter-modal dependencies. Taking P'_L, P'_V, and P'_A as inputs, this layer leverages self-attention and cross-attention mechanisms to comprehensively model the complex intra- and inter-modal interactions, yielding the enhanced modality-specific features TP'_m. Simultaneously, the shared features $\tilde{H}_m$ derived from the reconstructed features are processed by a self-attention Transformer to further refine global semantic information, yielding the enhanced shared feature $T\tilde{H}$. Finally, TP'_m and $T\tilde{H}$ are fused to produce the final representation M:

$$M = Cat\left(TP'_L, TP'_V, TP'_A, T\tilde{H}\right), \tag{11}$$

where Cat denotes the concatenation operation.

Hierarchical Predictions. After multimodal fusion, a prediction head is used to obtain the final output. The fusion prediction loss $\mathcal{L}_f$ is defined as

$$\mathcal{L}_f = \frac{1}{N} \sum_{i=1}^{N} | \hat{y}_i - y_i |, \tag{12}$$

where N denotes the number of samples, $\hat{y}_i$ is the prediction, and y_i is the ground truth of sample i.

Following the hierarchical prediction design of DLF [25], we introduce auxiliary classifiers on both the shared and specific branches. Consequently, the model is jointly optimized via the shared loss $\mathcal{L}_{T\tilde{H}}$, the modality-specific losses $\mathcal{L}_{TP'_m}$, and the prediction loss $\mathcal{L}_f$. Notably, unlike DLF [25], the shared features in our framework are derived from cross-reconstruction estimates rather than the initial disentangled features, yielding stronger cross-modal consistency. The overall MSA loss is expressed as

$$\mathcal{L}_{MSA} = \alpha \mathcal{L}_f + \beta \mathcal{L}_{T\tilde{H}} + \sum_{m \in \mathcal{M}} \mathcal{L}_{TP'_m}, \tag{13}$$

where α and β are hyper-parameters, whose sensitivity is analyzed in Section 4.5.

Overall, the optimization objective of the CACR framework is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{MSA} + \mathcal{L}_{recon} + \mathcal{L}_o + \mathcal{L}_p. \tag{14}$$

4. Experiments

Datasets. To ensure fair comparison with prior methods, we evaluate CACR on three widely used multimodal sentiment analysis benchmarks: CMU-MOSI [33], CMU-MOSEI [35], and CH-SIMS [31]. All three datasets contain language, visual, and audio modalities, with each sample annotated with sentiment intensity scores.

CMU-MOSI consists of 2,199 video clips, split into 1,284 training, 229 validation, and 686 test samples. The sentiment labels range from -3 (highly negative) to +3 (highly positive). **CMU-MOSEI** is an extended version of CMU-MOSI, containing 22,856 samples divided into 16,326 for training, 1,871 for validation, and 4,659 for testing, with the same sentiment annotation range of -3 to +3. **CH-SIMS** is a Chinese multimodal sentiment dataset comprising 2,281 samples, split into 1,368 training, 456 validation, and 457 test samples, annotated with sentiment scores from -1 (negative) to +1 (positive).

Evaluation Metrics. Following prior studies [13, 16, 38],

Table 1. Performance comparison on the MOSI dataset. The results are reported as average over 5 runs.

Methods	$ACC_7 \uparrow$	$ACC_2 \uparrow$	F1 $\uparrow$	MAE $\downarrow$
TFN [34]	34.90	80.08	80.07	0.901
LMF [14]	33.20	82.50	82.40	0.917
MFM [24]	35.83	78.87	78.90	0.927
MulT [23]	40.00	83.00	82.00	0.871
MISA [10]	41.37	83.54	83.58	0.777
MAG-BERT [17]	43.62	84.43	84.61	0.727
MFSA [30]	41.10	83.30	83.70	0.856
Self-MM [32]	43.61	82.69	82.97	0.760
FDMER [28]	44.03	83.48	83.51	0.761
DMD [13]	45.77	84.15	84.19	0.728
CGGM [8]	45.04	85.52	85.51	0.761
EMOE [7]	<u>47.80</u>	85.40	85.30	**0.710**
DLF [25]	45.34	<u>85.67</u>	<u>85.68</u>	0.733
CACR (Ours)	**48.69**	**87.04**	**86.90**	**0.710**

Table 2. Performance comparison on the MOSEI dataset. The results are reported as average over 5 runs.

Methods	$ACC_7 \uparrow$	$ACC_2 \uparrow$	F1 $\uparrow$	MAE $\downarrow$
TFN [34]	51.60	81.89	81.74	0.572
LMF [14]	51.59	83.48	83.36	0.575
MFM [24]	51.34	82.85	82.85	0.575
MulT [23]	52.84	84.63	84.52	0.565
MISA [10]	52.05	84.67	84.66	0.557
MAG-BERT [17]	52.67	84.82	84.71	0.543
MFSA [30]	53.20	83.80	83.60	0.574
Self-MM [32]	52.15	82.50	82.53	0.579
FDMER [28]	52.33	83.56	83.29	0.572
DMD [13]	52.61	84.53	84.49	0.543
CGGM [8]	53.06	84.81	84.86	0.544
EMOE [7]	<u>54.10</u>	<u>85.30</u>	<u>85.30</u>	<u>0.536</u>
DLF [25]	52.20	85.00	85.04	0.541
CACR (Ours)	**54.41**	**85.37**	**85.36**	**0.532**

Table 3. Performance comparison on the CH-SIMS dataset. The results are reported as average over 5 runs.

Methods	$ACC_2 \uparrow$	F1 $\uparrow$	Corr $\uparrow$	MAE $\downarrow$
TFN [34]	78.38	78.62	0.591	0.432
LMF [14]	77.77	77.88	0.575	0.441
MulT [23]	78.56	79.66	0.564	0.453
MISA [10]	76.54	76.59	0.563	0.447
MAG-BERT [17]	74.44	71.75	0.399	0.492
Self-MM [32]	77.31	77.91	0.543	0.430
DMD [13]	78.34	<u>78.56</u>	0.563	<u>0.426</u>
DLF [25]	<u>78.56</u>	78.53	<u>0.564</u>	0.439
CACR (Ours)	**81.62**	**80.99**	**0.596**	**0.423**

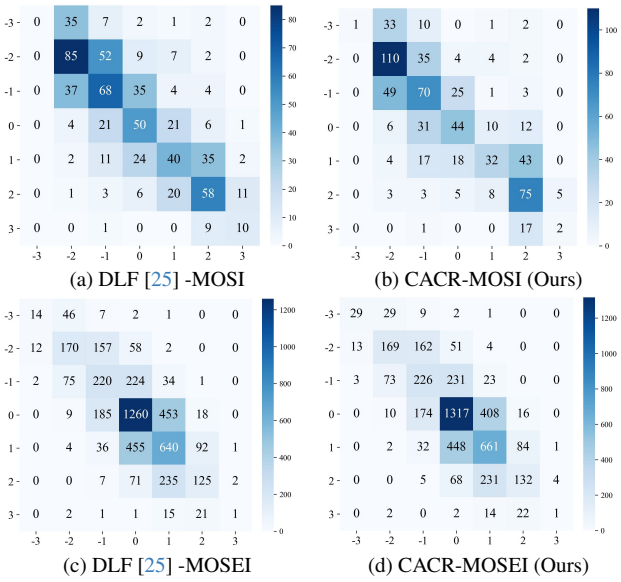
we adopt consistent evaluation metrics. For the MOSI and MOSEI datasets, we comprehensively evaluate the performance of CACR using 7-class accuracy (ACC_7), Binary accuracy (ACC_2), F1-score (F1), and Mean Absolute Error (MAE). For CH-SIMS dataset, we report ACC_2, F1 score, the Pearson correlation between model predictions and human annotations (Corr), and MAE as the main metrics.

Implementation Details. In our CACR framework, we adopt the same feature extraction settings as prior works [13, 26]. Specifically, for the language modality, we adopt the BERT-base-uncased pre-trained model [5] to extract language features, yielding 768-dimensional word embeddings. For the visual modality, we use the Facet toolkit [2] to encode videos frame by frame and extract 35-dimensional facial action unit features [12]. For the audio modality, we employ the COVAREP toolkit [4] to obtain 74-dimensional audio features. Our CACR is implemented in PyTorch, and all experiments are executed on an NVIDIA A800 GPU with 80GB of memory.

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

We compare our CACR approach with state-of-the-art methods under the same dataset settings: TFN [34], LMF [14], MFM [24], MulT [23], MISA [10], MAG-BERT [17], MFSA [30], Self-MM [32], FDMER [28], DMD [13], CGGM [8], EMOE [7], and DLF [25]. As shown in Tables 1 to 3, CACR consistently outperforms existing methods on all reported metrics. On MOSI and MOSEI, CACR achieves the best results on ACC_7, ACC_2, F1 score. Similarly, on CH-SIMS, it obtains higher ACC_2, F1, and Corr than prior approaches, clearly demonstrating the superiority of CACR. Specifically, compared to existing disentanglement-based methods (MISA [10], FDMER [28], DMD [13], and DLF [25]), they typically assume inter-modal emotional consistency without considering potential emotional conflicts. In contrast, building upon the feature disentanglement paradigm, CACR implicitly aligns shared representations via CRM, effectively mitigating the semantic ambiguity induced by similarity constraints. To prevent modality interference, our CAW strategy quantifies sample-level conflict scores and adaptively suppresses the influence of conflicting modalities. Additionally, the FSR module extracts fine-grained sentiment cues from visual- and audio-specific features to explicitly supplement the textual semantics. These components enables CACR to achieve superior performance on both classification and regression metrics across multiple datasets.

Confusion Matrix Analysis. To further evaluate the discriminative capability of CACR, we compare its confusion matrices with those of the representative disentanglement-based method DLF [25] on the MOSI and MOSEI datasets.



(a) DLF [25] -MOSI (b) CACR-MOSI (Ours)

(c) DLF [25] -MOSEI (d) CACR-MOSEI (Ours)

Figure 5. Comparison of confusion matrices between CACR and DLF [25] on MOSI and MOSEI datasets, showing the predicted sentiment classes against ground-truth labels.

As illustrated in Figure 5, CACR achieves a higher accurate prediction rate and substantially fewer misclassification errors compared to DLF across all sentiment levels. In particular, DLF shows pronounced confusion in classes -2 and 2, whereas CACR correctly classifies more samples in these categories, resulting in clearer boundaries between adjacent sentiment intensities. These results demonstrate that the CAW strategy and CRM module effectively alleviate inter-class confusion, enabling the FSR module to capture fine-grained sentiment cues for more robust sentiment discrimination.

4.2. Ablation Studies

To investigate the impact of different core components and loss regularization terms on the overall performance of CACR, we conduct comprehensive ablation studies on the MOSI dataset, using ACC_7, ACC_2, F1, and MAE as evaluation metrics, as reported in Table 4.

Core Components. Notably, removing CRM leads to a drastic drop in accuracy and a substantial increase in MAE, verifying that cross-modal semantic alignment is the basis of our model. Similarly, removing FSR decreases overall performance, confirming the necessity of capturing fine-grained emotional cues and enhancing the textual representation of subtle emotions. When CAW is either removed (by fixing the cross-reconstruction loss weights to 1) or replaced with a standard adaptive weighting mechanism (denoted as "gating"), the model performance notably degrades. This indicates that the performance improvements stem specifically from CAW rather than generic adaptive weighting. This demonstrates that by explicitly modeling sample-level conflicts and adaptively down-weighting the negative impact of conflicting modalities during reconstruction, the model effectively mitigates modality interference.

Table 4. Results of ablation studies on MOSI dataset.

Method	$ACC_7 \uparrow$	$ACC_2 \uparrow$	F1 $\uparrow$	MAE $\downarrow$
Different Critical Components				
w/o CRM	40.23	84.15	84.15	0.791
w/o CAW	44.31	83.84	83.83	0.752
gating	44.98	84.91	84.85	0.747
w/o FSR	46.79	86.43	86.37	0.751
Different Loss Regularization				
w/o $\mathcal{L}_o$	40.96	85.06	85.13	0.779
w/o $\mathcal{L}_p$	42.53	84.76	84.71	0.788
CACR (Ours)	**48.69**	**87.04**	**86.90**	**0.710**

Table 5. Comparison results of CACR and its variant CACR-S.

Method	MOSI		MOSEI	
	$ACC_2 \uparrow$	F1 $\uparrow$	$ACC_2 \uparrow$	F1 $\uparrow$
CACR	87.04	86.90	85.37	85.36
CACR-S	**87.17**	**87.13**	**86.21**	**86.12**

Loss Regularization Terms. We further perform ablation studies by removing different loss terms to analyze the effect of each regularization component. When the soft orthogonality loss $\mathcal{L}_o$ is removed, the model can no longer effectively separate shared and specific features, leading to a sharp decline in overall classification and regression metrics. This indicates that the orthogonality regularization is crucial for stable and reliable feature disentanglement. Furthermore, removing the specific loss $\mathcal{L}_p$ also leads to consistent performance drops, showing that this loss term effectively prevents the modality-specific features from drifting during cross-reconstruction process.

4.3. Compatibility Analysis

To further investigate whether the generalization of our framework is limited by specific feature encoders, we equip CACR with stronger encoders. Specifically, we employ the pre-trained DeBERTa-large [11], MA-Net [37], and wav2vec-large [18] models to extract features for the language, visual, and audio modalities, respectively. We denote this variant as CACR-S.

As reported in Table 5, CACR-S yields further performance gains. CACR operates at the feature level and does not depend on whether the initial features are handcrafted or extracted by end-to-end models, thus exhibiting excellent compatibility across different backbones.

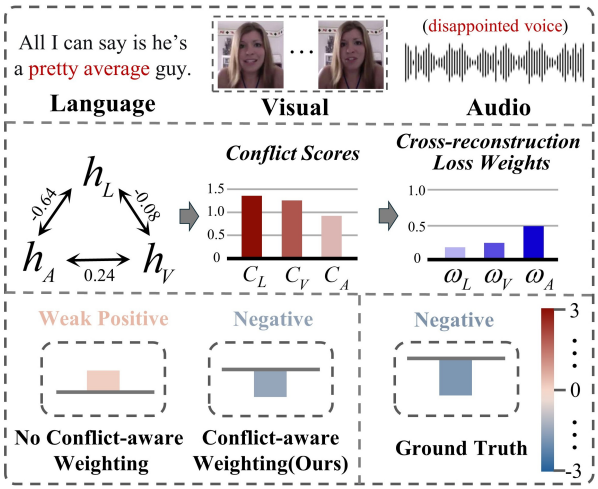


Figure 6. Case study of samples with emotional conflicts on the MOSEI dataset and its binary prediction results.

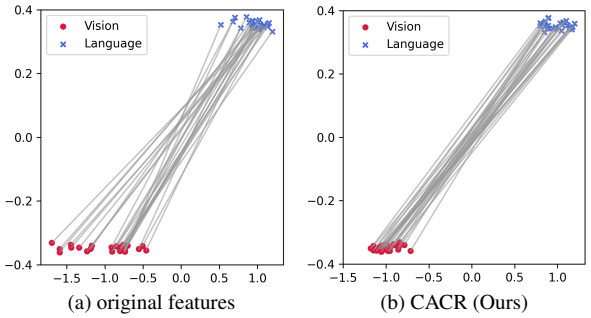
4.4. Qualitative Analysis

Case Study. Figure 6 visualizes a MOSEI sample with emotional conflicts to demonstrate the effectiveness of our conflict-aware weighting (CAW) strategy. In this sample, language and visual cues exhibit a neutral-to-positive tendency, which contradicts the clearly disappointed audio tone. As shown in the middle of Figure 6, CACR quantifies this conflict via cosine similarity and maps conflict scores into adaptive cross-reconstruction weights, among which the audio modality receives the highest weight. Without CAW, the model fails to resolve this conflict and yields an incorrect weakly positive prediction. Conversely, CACR successfully suppresses the misleading language and visual modalities while reinforcing the reliable audio modality, outputting a negative prediction aligned with the ground truth. This proves that suppressing conflicting modalities enhances overall prediction accuracy.

Analysis of the Modality Gap. Figure 7 presents a feature visualization case study on the language and visual modalities, illustrating the effectiveness of CACR in mitigating the semantic gap between modalities. In the original feature space (Figure 7a), language and visual features exhibit clear distributional discrepancies, which hinder cross-modal semantic alignment. By employing our CRM module (Figure 7b), the feature distributions of the two modalities are pulled closer together, significantly reducing the inter-modal distance. These results indicate that our CRM effectively alleviates the semantic gap between the language and visual modalities.

4.5. Hyper-parameter Sensitivity Analysis

To evaluate the robustness of CACR, we conduct a sensitivity analysis on key hyper-parameters ω_{min}, τ, α, and β on the MOSI dataset. As shown in Figure 8, excluding extreme settings, the ACC_7 and MAE scores remain relatively stable as these parameters vary across different values, demonstrating the framework's robustness to hyper-parameter variations. Consequently, we uniformly adopt the optimal settings of $\omega_{min} = 0.1$, $\tau = 0.5$, $\alpha = 1$, and $\beta = 2$ as our default configuration for all experiments.



(a) original features (b) CACR (Ours)

Figure 7. Visualization of the modality gap between vision and language on MOSEI dataset.

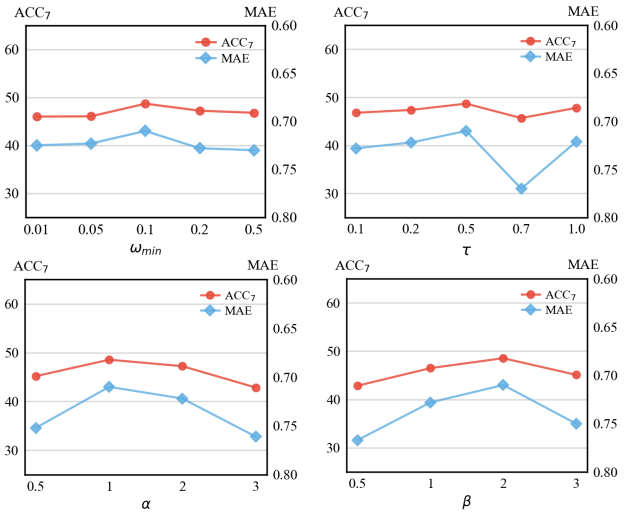


Figure 8. Hyper-parameter sensitivity analysis on MOSI dataset.

5. Conclusion and Future Discussion

In this work, we propose CACR, a conflict-aware cross-reconstruction framework designed to explicitly model sample-level emotional conflicts. By adaptively down-weighting conflicting modalities and capturing fine-grained cues, our approach effectively mitigates inter-modal interference and achieves superior performance on three multimodal benchmarks. In future work, we will extend CACR to broader multimodal tasks and further investigate its robustness in real-world scenarios with incomplete or missing modalities, aiming to maintain reliable predictions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U24A20323, 62376145, and 62406184), the Science and Technology Innovation Talent Team of Shanxi Province (202204051002016), and the Taiyuan City "Double Hundred Research Action" of the first batch project about "Leading the Charge with Open Competition" (2024TYJB0127).

References

[1] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205, 2007. 2

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision*, pages 1–10, 2016. 6

[3] Ringki Das and Thoudam Doren Singh. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38, 2023. 1

[4] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *ICASSP*, pages 960–964, 2014. 6

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, 2019. 6

[6] Kaouther Ezzameli and Hela Mahersia. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 99:101847, 2023. 1

[7] Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *CVPR*, pages 14314–14324, 2025. 2, 6

[8] Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In *NeurIPS*, pages 133328–133344, 2024. 6

[9] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, 2021. 2

[10] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM MM*, pages 1122–1131, 2020. 2, 6

[11] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*, 2021. 7

[12] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *CVPR*, pages 10924–10933, 2019. 6

[13] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *CVPR*, pages 6631–6640, 2023. 2, 5, 6

[14] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2247–2256, 2018. 6

[15] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *CVPR*, pages 2554–2562, 2021. 2

[16] Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. M-sema: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 204–213, 2022. 5

[17] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, 2020. 6

[18] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. Wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019. 7

[19] Amir Shirian and Tanaya Guha. Compact graph architecture for speech emotion recognition. In *ICASSP*, pages 6284–6288, 2021. 2

[20] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *ACM MM*, pages 3722–3729, 2022. 1

[21] Hao Sun, Ziwei Niu, Hongyi Wang, Xinyao Yu, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. Multimodal sentiment analysis with mutual information-based disentangled representation learning. *IEEE Transactions on Affective Computing*, 16 (3):1606–1617, 2025. 1

[22] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325, 2023. 2

[23] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019. 5, 6

[24] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019. 6

[25] Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *AAAI*, pages 21180–21188, 2025. 1, 2, 5, 6, 7

[26] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *ICCV*, pages 22025–22034, 2023. 6

[27] Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *AAAI*, pages 1601–1609, 2025. 2

[28] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *ACM MM*, pages 1642–1651, 2022. 2, 6

[29] Dingkang Yang, Shuai Huang, Yang Liu, and Lihua Zhang. Contextual and cross-modal interaction for multi-modal speech emotion recognition. *IEEE Sign. Process. Letters*, 29:2093–2097, 2022. 2

[30] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences. In *ACM MM*, pages 1708–1717, 2022. 6

[31] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, 2020. 5

[32] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*, pages 10790–10797, 2021. 2, 6

[33] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 5

[34] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017. 2, 6

[35] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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*, pages 2236–2246, 2018. 5

[36] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767, 2023. 2

[37] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE TIP*, 30:6544–6556, 2021. 7

[38] Ying Zhou, Xuefeng Liang, Han Chen, Yin Zhao, Xin Chen, and Lida Yu. Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, 114: 102663, 2025. 5