MRCF: RETHINKING MODALITY RELIABILITY IN MULTIMODAL LEARNING WITH INCOMPLETE OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Sentiment Analysis (MSA) integrates text, audio, and vision to infer human affect, yet real-world multimodal observations are often incomplete. Existing methods for incomplete-observation MSA mainly follow two broad directions. Reconstruction-based methods recover missing information from observed modalities, while joint-representation methods learn directly from incomplete inputs. Although effective, these methods usually treat modality reliability only implicitly within representation learning or fusion design rather than modeling it explicitly. We argue that modality reliability is a central variable in incomplete-observation settings. Failure to model it explicitly gives rise to two related issues. The first is reliability mismatch, in which the affective evidence retained by each modality varies across samples and missing rates. The second is reliability propagation bias, in which messages from degraded modalities may adversely affect cross-modal interaction and predictive performance. To address these issues, we propose MRCF, a Modality Reliability-Calibrated Framework for MSA with incomplete observations. MRCF contains a Reliability-Aware Branch that estimates sample-specific modality reliability from intramodal quality cues and cross-modal semantic consistency, a Reliability-Guided Interaction Branch that uses the estimated scores to modulate cross-modal information flow, and a Reliability-Calibrated Fusion Module that integrates reliability and semantic cues for final prediction. Experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS show that MRCF achieves strong performance under standard incomplete-observation protocols. Further analyses provide evidence that explicit reliability modeling helps mitigate reliability mismatch and reliability propagation bias during interaction and fusion.

1 INTRODUCTION

Multimodal Sentiment Analysis (MSA) aims to infer human sentiment by integrating complementary cues from text, audio, and vision (Meng et al., 2026a). Its effectiveness relies on the fact that different modalities provide heterogeneous yet mutually informative evidence for affect understanding (Meng et al., 2026b). In real-world applications, multimodal observations are often incomplete due to sensor failures, occlusion, noise, privacy constraints, and transcription errors (Zhang et al., 2024; Zhu et al., 2025). Under such conditions, models must infer sentiment reliably from partial evidence (Li et al., 2025), making incomplete-observation MSA increasingly important in practice.

Existing studies mainly address this problem through two broad directions: reconstruction-based learning and joint-representation learning. These directions emphasize missing-information recovery and robust joint representations, respectively, and can overlap in individual methods. Reconstruction-based methods attempt to recover missing information from observed modalities before performing downstream prediction (Yuan et al., 2024). Representative methods include TFR-Net (Yuan et al., 2021), which employs attention-based extractors and a reconstruction module to recover missing semantics in unaligned multimodal sequences, and DAR (Yang et al., 2025), which decouples modality-common and modality-specific semantics and reconstructs them before fusion. Joint-representation methods learn robust multimodal representations directly from incomplete inputs via distillation, dominant-modality preservation, proxy modeling, or fusion redesign (Sun et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2024b). CorrKD (Li et al., 2024) uses correlation-decoupled distillation for uncertain missing patterns, LNLN (Zhang et al., 2024) preserves the dominant language modality, and P-RMF (Zhu et al., 2025) models proxy modalities and uncertainty in a Gaussian latent space. Although designed for complete observations, ALMT (Zhang et al., 2023) offers related insight through language-guided suppression of disruptive auxiliary cues. See Appendix A for details.

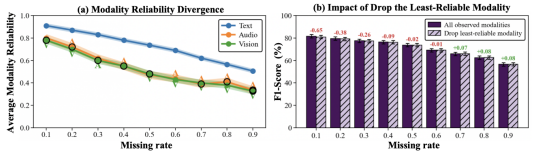

Figure 1: Modality reliability analysis under incomplete observations. (a) Reliability degradation varies across modalities with increasing missing rates, causing the least-reliable modality to change across stages. (b) Modality removal is measured by $\Delta\mathrm{F1} = \mathrm{F1}_{\mathrm{drop}} - \mathrm{F1}_{\mathrm{all}}$; negative and positive values indicate degradation and negative transfer, respectively.

Despite these advances, existing methods do not explicitly and systematically model modality reliability under incomplete observations. Instead, their focus on missing-information recovery or representation fusion leaves reliability implicit rather than treating it as a sample-specific variable. Fig. 1 provides two diagnostic analyses of this issue. In Fig. 1(a), we use the agreement between a modality-specific prediction and the corresponding full-observation prediction at each missing rate as a proxy for modality reliability. The resulting curves show that this reliability proxy is neither uniform across modalities nor static across missing rates. As the missing rate increases, the proxy scores of different modalities decrease at different rates. In particular, while text remains the highest-scoring modality overall, audio and vision alternate as the lowest-scoring modality at different missing stages, revealing a clear reliability mismatch across modalities. This indicates that the affective evidence retained under incomplete observations varies dynamically across modalities and corruption levels, whereas current methods do not generally expose this variation as an explicit sample-specific quantity. Figure 1(b) further examines the predictive effect of removing the lowest-scoring modality at each missing rate. The comparison shows a systematic transition as missingness increases. At low missing rates, dropping the lowest-scoring modality leads to a noticeable performance decrease, suggesting that even a relatively low-scoring modality may retain weak but useful complementary cues. As the missing rate increases further, the performance gap gradually shrinks and eventually becomes slightly positive, indicating that evidence from the lowest-scoring modality becomes increasingly likely to disrupt interaction and fusion rather than provide useful support. This pattern is consistent with reliability propagation bias: errors introduced by degraded evidence may spread through cross-modal interaction and adversely affect the joint representation. This interpretation is also consistent with several recent findings. LNLN (Zhang et al., 2024) identifies temporal information corruption and modality laziness under random missingness, P-RMF (Zhu et al., 2025) highlights unequal modality contributions, cross-modal discrepancies, and uncertainty, and CMAD (Zhuang et al., 2025) reveals that uncontrolled interaction causes representation inconsistency and optimization instability. These findings motivate explicitly modeling sample-specific modality reliability under incomplete observations.

Motivated by these, we propose the Modality Reliability-Calibrated Framework (MRCF) for MSA with incomplete observations. MRCF contains two complementary branches and a reliability-calibrated fusion module. First, a Reliability-Aware Branch explicitly estimates sample-specific modality reliability by jointly modeling intramodal quality cues and cross-modal semantic consistency and is designed to address reliability mismatch. Second, a Reliability-Guided Interaction Branch uses the estimated reliability to regulate cross-modal information flow, assigning relatively greater influence to messages from higher-scoring modalities and reducing the influence of lower-scoring ones, thereby targeting reliability propagation bias during interaction. Finally, a Reliability-Calibrated Fusion Module integrates semantic and reliability cues for final prediction and uses complete observations as training-time anchors for proxy reliability supervision and fusion calibration. In contrast to methods that focus mainly on processing incomplete inputs, our framework emphasizes estimating modality reliability, using it during interaction, and calibrating the final fusion decision. The contributions are: **1.** We characterize dynamic modality reliability in incomplete-observation MSA and propose the Modality Reliability-Calibrated Framework (MRCF) to address reliability mismatch and propagation bias. **2.** We design a Reliability-Aware Branch (RAB) to estimate sample-specific modality reliability from intra- and cross-modal cues, and a Reliability-Guided Interaction Branch (RGIB) to suppress unreliable cross-modal messages.

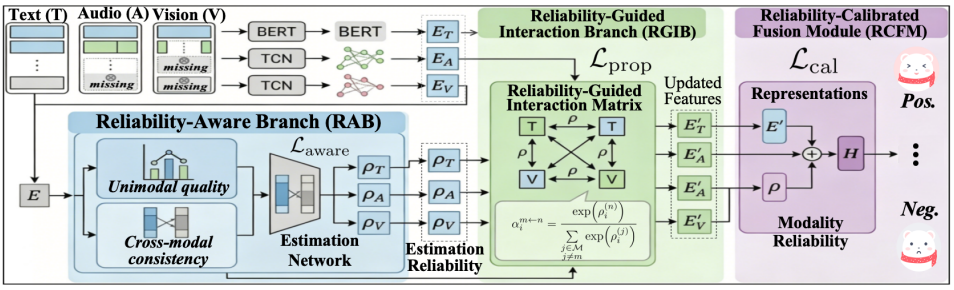

Figure 2: MRCF estimates modality reliability for cross-modal interaction and calibrated fusion under incomplete observations.

3. We introduce a Reliability-Calibrated Fusion Module (RCFM) that jointly models semantic and reliability information and uses a complete-observation semantic anchor to calibrate fusion, yielding stable representations and improved prediction.

2 Methodology

2.1 Overview

We study MSA with incomplete observations. Given an input sample, one or more modalities may be partially missing, making the remaining affective evidence both incomplete and unevenly reliable across modalities. To reduce the influence of degraded evidence on cross-modal interaction, we propose the Modality Reliability-Calibrated Framework (MRCF). As shown in Fig. 2, MRCF contains three tightly coupled components. First, a Reliability-Aware Branch estimates sample-specific modality reliability from intramodal quality cues and cross-modal semantic consistency. Second, a Reliability-Guided Interaction Branch uses the estimated reliability to modulate cross-modal information flow, assigning relatively greater influence to messages from higher-scoring modalities. Third, a Reliability-Calibrated Fusion Module integrates semantic and reliability cues for final prediction and aligns incomplete-observation fusion with a complete-observation semantic anchor during training. A key property of MRCF is that modality reliability is treated as an explicitly estimated quantity rather than an implicit by-product of feature learning. Accordingly, the framework is optimized for two coupled goals: estimating modality reliability under incomplete observations and using it to reduce reliability propagation bias during interaction and fusion.

2.2 Feature Extraction

Let $\mathcal{M} = \{t, a, v\}$ denote the modality set, where t, a, and v represent text, audio, and vision, respectively. Given a training set $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$, the complete multimodal input of the i-th sample is written as $\mathbf{X}_i = \{X_i^{(m)}\}_{m\in\mathcal{M}}$, where $X_i^{(m)} \in \mathbb{R}^{L_i^{(m)}\times d_m}$ is the modality-specific sequence, $L_i^{(m)}$ is the sequence length, and d_m is the feature dimension. The sentiment label is $y_i \in \mathbb{R}$. Following the standard incomplete-observation protocol, we construct an incomplete counterpart $\widetilde{\mathbf{X}}_i = \{\widetilde{X}_i^{(m)}\}_{m\in\mathcal{M}}$ by applying $\widetilde{X}_i^{(m)} = \mathcal{C}\left(X_i^{(m)}, B_i^{(m)}\right)$, where $B_i^{(m)} \in \{0,1\}^{L_i^{(m)}}$ is a binary observation mask, $B_{ij}^{(m)} = 1$ denotes an observed temporal position, and $\mathcal{C}(\cdot)$ is the corruption operator. We define the observed ratio of modality m as $\gamma_i^{(m)} = \frac{1}{L_i^{(m)}}\sum_{j=1}^{L_i^{(m)}} B_{ij}^{(m)}$. This input construction is standard and not our primary contribution. Unlike existing incomplete-observation MSA methods that directly reconstruct, distill, or fuse masked inputs, we explicitly model modality reliability as a sample-specific quantity. Since the affective evidence retained by each modality varies across samples and missing rates, the model should first assess what information remains and how reliable it is before cross-modal interaction. For each modality $m \in \mathcal{M}$, a modality-specific encoder E_m maps the incomplete input to a hidden sequence $\mathbf{H}_i^{(m)} = E_m\left(\widetilde{X}_i^{(m)}\right)$, with $\mathbf{H}_i^{(m)} \in \mathbb{R}^{L_i^{(m)}\times d}$, where d is the shared hidden dimension. A pooling layer then produces a compact modality rep-

resentation $\mathbf{h}_i^{(m)} = \mathrm{Pool}\Big(\mathbf{H}_i^{(m)}\Big) \in \mathbb{R}^d$. During training, we additionally encode the complete observations using the same encoders: $\mathbf{H}_{i,c}^{(m)} = E_m\Big(X_i^{(m)}\Big)$ and $\mathbf{h}_{i,c}^{(m)} = \mathrm{Pool}\Big(\mathbf{H}_{i,c}^{(m)}\Big)$. These complete-observation features are used only as training-time anchors for reliability supervision and fusion calibration, and are not required during inference.

2.3 Reliability-Aware Branch

The first component of MRCF estimates how trustworthy each modality is for the current sample under incomplete observations. This branch is motivated by the fact that modality reliability should not be treated as a static prior, because it depends jointly on what remains observable within the modality and whether the retained semantics agree with those of other modalities. To this end, we use two complementary cues. **Intramodal quality cue.** We first characterize the internal quality of the observed modality itself. The pooled modality representation is combined with its observed ratio to obtain an intramodal quality feature $\mathbf{q}_i^{(m)} = \phi_q^{(m)}\Big([\mathbf{h}_i^{(m)}; \gamma_i^{(m)}]\Big) \in \mathbb{R}^{d_q}$, where $\phi_q^{(m)}(\cdot)$ is a modality-specific projection network and $[\cdot;\cdot]$ denotes concatenation. **Cross-modal semantic consistency cue.** We then evaluate whether modality m remains semantically consistent with the other modalities. If it still preserves useful affective evidence, its representation should be more compatible with the representations derived from the others. We therefore define the semantic consistency score of modality m as $s_i^{(m)}$, where $W_s^{(m)} \in \mathbb{R}^{d\times d}$ is a learnable projection matrix and $\cos(\cdot,\cdot)$ is cosine similarity.

$$s_i^{(m)} = \frac{1}{|\mathcal{M}|-1} \sum_{n\in\mathcal{M}, n\neq m} \cos\Big(W_s^{(m)}\mathbf{h}_i^{(m)}, W_s^{(n)}\mathbf{h}_i^{(n)}\Big), \tag{1}$$

Reliability estimation. Modality reliability is estimated by combining the two cues through a nonlinear mapping, $\rho_i^{(m)} = \sigma\Big(\mathbf{w}_r^\top \tanh\Big(W_r[\mathbf{q}_i^{(m)}; s_i^{(m)}] + \mathbf{b}_r\Big)\Big)$. Here $W_r \in \mathbb{R}^{d_r\times(d_q+1)}$ and $\mathbf{w}_r, \mathbf{b}_r \in \mathbb{R}^{d_r}$ are learnable parameters, and the sigmoid gives $\rho_i^{(m)} \in (0,1)$. A larger $\rho_i^{(m)}$ means that modality m is estimated to provide more trustworthy affective evidence for sample i. **Proxy reliability supervision.** Since complete observations are available during training, we construct a training-time proxy target from the agreement between the incomplete unimodal prediction and the complete-observation prediction. For modality m, we first obtain a unimodal prediction $\hat{y}_i^{(m)} = g_m\Big(\mathbf{h}_i^{(m)}\Big)$. The complete-observation prediction $\hat{y}_i^c$ is obtained from the complete-observation fusion anchor defined in Sec. 2.5. We then define $\bar{\rho}_i^{(m)} = \exp\Big(-\Big|\hat{y}_i^{(m)} - \mathrm{sg}(\hat{y}_i^c)\Big|\Big)$, where $\mathrm{sg}(\cdot)$ stops gradients through the complete-observation prediction. This target is a model-derived training proxy, not a ground-truth reliability label; it reflects how closely an incomplete unimodal prediction agrees with the complete-observation prediction. The reliability regression loss is, $\mathcal{L}_{\mathrm{rel}} = \frac{1}{N|\mathcal{M}|}\sum_{i=1}^N \sum_{m\in\mathcal{M}} \mathrm{SL1}\Big(\rho_i^{(m)}, \bar{\rho}_i^{(m)}\Big)$, where $\mathrm{SL1}(\cdot,\cdot)$ denotes the smooth-L_1 loss. To encourage the ordering induced by the proxy targets, we use the pairwise ranking term

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{N}\sum_{i=1}^N \frac{1}{\max\{1, |\mathcal{P}_i|\}} \sum_{(m,n)\in\mathcal{P}_i} \log\Big(1 + \exp\Big(\rho_i^{(n)} - \rho_i^{(m)}\Big)\Big), \tag{2}$$

where $\mathcal{P}_i = \Big\{(m,n) \mid \bar{\rho}_i^{(m)} > \bar{\rho}_i^{(n)}\Big\}$; the inner sum is zero when $\mathcal{P}_i$ is empty. The overall objective of the Reliability-Aware Branch is $\mathcal{L}_{\mathrm{aware}} = \mathcal{L}_{\mathrm{rel}} + \lambda_r\mathcal{L}_{\mathrm{rank}}$ where λ_r is a balancing coefficient.

2.4 Reliability-Guided Interaction Branch

Estimating reliability alone is not sufficient. The crucial question is how to use reliability once it has been estimated. This motivates our Reliability-Guided Interaction Branch.

The need for this design follows from the second phenomenon discussed in the introduction. When representations from low-reliability modalities enter cross-modal interaction without explicit reweighting, their distorted evidence is no longer confined to their original streams. Instead, its errors may propagate into other modalities and affect the joint representation. Therefore, reliability

should not be used only to assign a higher or lower fusion weight at the end. It should also affect the earlier stage at which cross-modal messages are generated and routed.

For each ordered modality pair (m, n) with $m \neq n$, we define a cross-modal message from source modality n to target modality m. We first compute $Q_i^{(m)} = \mathbf{H}_i^{(m)} W_Q^{(m)}, K_i^{(n)} = \mathbf{H}_i^{(n)} W_K^{(n)}, V_i^{(n)} = \mathbf{H}_i^{(n)} W_V^{(n)}$, where $W_Q^{(m)}, W_K^{(n)}, W_V^{(n)} \in \mathbb{R}^{d \times d}$ are learnable projection matrices. The pairwise attention matrix is then $A_i^{m \leftarrow n} = \mathrm{Softmax}\left(\frac{Q_i^{(m)} (K_i^{(n)})^\top}{\sqrt{d}}\right)$ and the corresponding message is $\mathbf{U}_i^{m \leftarrow n} = A_i^{m \leftarrow n} V_i^{(n)}$, where $\mathrm{Softmax}(\cdot)$ is applied row-wise, so that $A_i^{m \leftarrow n} \in \mathbb{R}^{L_i^{(m)} \times L_i^{(n)}}$ and $\mathbf{U}_i^{m \leftarrow n} \in \mathbb{R}^{L_i^{(m)} \times d}$.

A conventional interaction module would directly aggregate these pairwise messages. In contrast, MRCF explicitly regulates them according to modality reliability. We define the source-side routing weight as $\alpha_i^{m \leftarrow n} = \frac{\exp\left(\rho_i^{(n)}\right)}{\sum_{j \in \mathcal{M}, j \neq m} \exp\left(\rho_i^{(j)}\right)}$. The updated representation of modality m is then computed by $\widetilde{\mathbf{H}}_i^{(m)} = \mathbf{H}_i^{(m)} + (1 - \rho_i^{(m)}) \sum_{\substack{n \in \mathcal{M} \\ n \neq m}} \alpha_i^{m \leftarrow n} \mathbf{U}_i^{m \leftarrow n}$. This design reflects the intended interaction logic. If target modality m has a high reliability score, then $1 - \rho_i^{(m)}$ is small and its update is reduced. If target modality m has a lower score, the update is larger, with relatively more weight assigned to higher-scoring source modalities. Thus, reliability affects both the direction and magnitude of cross-modal information flow.

The proposed interaction mechanism is intended to preserve high-scoring modalities rather than overwrite them unnecessarily. It also turns cross-modal interaction into a reliability-aware process rather than uniform semantic mixing. To reduce the energy of outgoing messages from low-scoring source modalities, we use the following regularizer:

$$\mathcal{L}_{\mathrm{prop}} = \frac{1}{N|\mathcal{M}|(|\mathcal{M}|-1)} \sum_{i=1}^{N} \sum_{\substack{m,n \in \mathcal{M} \\ m \neq n}} \left(1 - \bar{\rho}_i^{(n)}\right) \|\alpha_i^{m \leftarrow n} \mathbf{U}_i^{m \leftarrow n}\|_F^2 \tag{3}$$

This term penalizes the squared Frobenius norm of routed messages more strongly when the source has a low proxy-reliability score; it is therefore a reliability-weighted message-energy regularizer.

Table 1: Results on MOSI and MOSEI under IMM. MRCF results are averaged over $r \in \{0.0, 0.1, \ldots, 0.9\}$. a/b denotes the negative vs. non-negative and negative vs. positive settings for both Acc-2 and F1. Boldface highlights MRCF.

	CMU-MOSI						CMU-MOSEI					
Model	Acc-2	F1	Acc-5	Acc-7	MAE (↓)	Corr	Acc-2	F1	Acc-5	Acc-7	MAE (↓)	Corr
MISA	70.33/71.49	70.00/71.28	33.08	29.85	1.085	0.524	75.82/71.27	68.73/63.85	39.39	40.84	0.780	0.503
Self-MM	69.26/70.51	67.54/66.60	34.67	29.55	1.070	0.512	77.42/73.89	72.31/68.92	45.38	44.70	0.695	0.498
MMIM	67.06/69.14	64.04/66.65	33.77	31.30	1.077	0.507	75.89/73.32	70.32/68.72	41.74	40.75	0.739	0.489
CENET	67.73/71.46	64.85/68.41	33.62	30.38	1.080	0.504	77.34/74.67	74.08/70.68	47.83	47.18	0.685	0.535
ALMT	68.39/70.40	71.80/72.57	33.42	30.30	1.083	0.498	77.54/76.64	78.03/77.14	41.64	40.92	0.674	0.481
LNLN	70.94/72.55	71.25/72.73	38.27	34.26	1.046	0.527	78.19/76.30	79.95/77.77	46.17	45.42	0.692	0.530
P-RMF	71.53/72.81	71.69/72.93	38.50	34.19	1.038	0.525	78.83/78.14	80.39/79.33	45.87	44.63	0.658	0.589
DAR	71.60/73.18	71.51/73.15	38.65	34.47	1.069	0.520	77.48/78.14	77.44/77.51	48.02	47.01	0.665	0.583
MIG-HCL	72.62	72.17	-	34.4	0.993	0.508	77.73	76.72	-	48.32	0.646	0.557
MRCF	**73.17/74.26**	**73.31/74.39**	**39.87**	**35.79**	**0.986**	**0.552**	**79.38/80.07**	**80.27/80.39**	**48.87**	**48.66**	**0.642**	**0.606**

2.5 Reliability-Calibrated Fusion Module

After reliability-guided interaction, the model still needs to produce a final joint representation for sentiment prediction. A standard fusion strategy would pool the updated modality features and weight them through a learned attention mechanism. However, this strategy does not explicitly condition its modality weights on the estimated reliability scores. Fusion under incomplete observations

should satisfy two requirements simultaneously. It should preserve modality complementarity, because a modality with lower reliability may still retain partial but useful affective cues. It should also reduce the influence of low-reliability evidence on the final decision. A semantic-only attention mechanism does not explicitly address the second requirement. This motivates the Reliability-Calibrated Fusion Module. We first pool each updated modality sequence $\widetilde{\mathbf{h}}_i^{(m)} = \mathrm{Pool}\Big(\widetilde{\mathbf{H}}_i^{(m)}\Big)$. The fusion weight of modality m is then defined by jointly considering its semantic representation and estimated reliability, $\beta_i^{(m)} = \frac{\exp\Big(\mathbf{w}_f^\top [\widetilde{\mathbf{h}}_i^{(m)}; \rho_i^{(m)}]\Big)}{\sum_{n\in\mathcal{M}} \exp\Big(\mathbf{w}_f^\top [\widetilde{\mathbf{h}}_i^{(n)}; \rho_i^{(n)}]\Big)}$, where $\mathbf{w}_f \in \mathbb{R}^{d+1}$ is a learnable parameter vector. Unlike a semantic-only fusion logit, the proposed logit also receives the estimated reliability score. The final contribution of a modality can therefore depend jointly on its semantic evidence and its estimated reliability under the current incomplete observation. In this way, reliability can influence the final representation rather than only an auxiliary loss. The final fused representation is $\mathbf{z}_i = \sum_{m\in\mathcal{M}} \beta_i^{(m)} \widetilde{\mathbf{h}}_i^{(m)}$, and the sentiment prediction is $\hat{y}_i = g_f(\mathbf{z}_i)$, where $g_f(\cdot)$ is the prediction head. To further stabilize fusion under incomplete observations, we introduce a complete-observation fusion anchor during training. Using the complete-observation features $\{\mathbf{h}_{i,c}^{(m)}\}_{m\in\mathcal{M}}$, we obtain $\mathbf{z}_i^c = \sum_{m\in\mathcal{M}} \beta_{i,c}^{(m)} \mathbf{h}_{i,c}^{(m)}$, where $\beta_{i,c}^{(m)}$ is computed by the same fusion head. The corresponding complete-observation prediction is $\hat{y}_i^c = g_f(\mathbf{z}_i^c)$. We then calibrate the incomplete-observation fused representation toward this complete-observation semantic anchor: $\mathcal{L}_{\mathrm{cal}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{z}_i - \mathrm{sg}(\mathbf{z}_i^c)\|_2^2$. This calibration serves two purposes. It is intended to discourage the fusion module from over-pruning low-reliability modalities and discarding useful complementary evidence. It also discourages excessive deviation between paired incomplete- and complete-observation fused representations, thereby encouraging representation stability.

2.6 Training Objective

The main sentiment regression objective is $\mathcal{L}_{\mathrm{task}} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|$. The final objective of MRCF is $\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_1 \mathcal{L}_{\mathrm{aware}} + \lambda_2 \mathcal{L}_{\mathrm{prop}} + \lambda_3 \mathcal{L}_{\mathrm{cal}}$, where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters. During inference, MRCF operates on the available observations and does not require the complete-observation branch. The complete-observation branch and proxy reliability targets are used only during training.

3 Experiments

3.1 Experimental Settings

Benchmarks. We evaluate MRCF on standard MSA benchmarks, namely MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020). **Evaluation Metrics.** We evaluate MOSI and MOSEI using MAE, Corr, F1, Acc-2, Acc-5, and Acc-7, with Acc-2 and F1 reported under both negative–non-negative and negative–positive partitions, and CH-SIMS using MAE, Corr, F1, Acc-2, Acc-3, and Acc-5. **Implementation Details.** All experiments are conducted in PyTorch on an NVIDIA A100 GPU. We train for 100 epochs using Adam with a learning rate of 10^{-4} and a batch size of 32. BERT (Devlin et al., 2019) is used as the language backbone, and the audio and visual feature extraction pipelines follow established baselines (Li et al., 2025; Zhu et al., 2025). MRCF results are averaged over three random seeds. Additional protocol details are provided in Appendix B. **Baseline references.** MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), CENET (Wang et al., 2023a), TFR-Net (Yuan et al., 2021), ALMT (Zhang et al., 2023), LNLN (Zhang et al., 2024), P-RMF (Zhu et al., 2025), DAR (Yang et al., 2025), MIG-HCL (Gong et al., 2026), DMD (Li et al., 2023), NIAT (Yuan et al., 2024), GCNet (Lian et al., 2023), IMDer (Wang et al., 2023b), EMT (Sun et al., 2024a), GMD (Wang et al., 2024), CorrKD (Li et al., 2024), and ROSA (Li et al., 2025).

3.2 Training Protocol

We follow the two protocols used in prior incomplete-observation MSA (Li et al., 2025; Zhu et al., 2025). Under intra-modal missingness (IMM), we erase a fraction $r \in \{0.0, 0.1, \ldots, 0.9\}$ of temporal positions. Missing-aware methods, including MRCF, P-RMF, LNLN, and ROSA, are trained and evaluated at the same target missing rate; methods requiring complete training inputs, such as

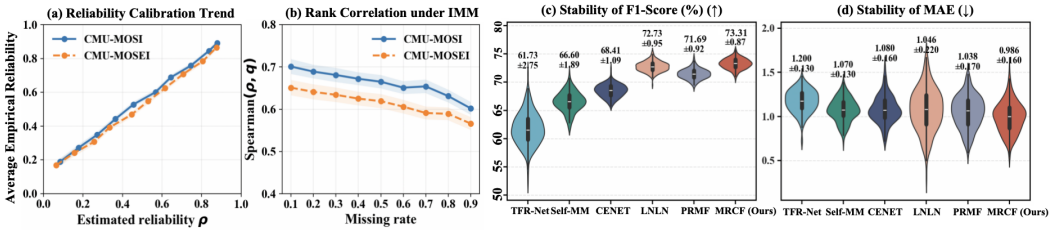

Figure 3: Modality reliability and prediction stability analysis. (a) Reliability trends. (b) Spearman correlations across missing rates, with higher values indicating stronger diagnostic consistency. (c,d) Violin distributions of F1 and MAE on MOSI over ten missing rates $r \in \{0.0, 0.1, \ldots, 0.9\}$ and three seeds, box plots show median and IQR, where smaller IQR indicates higher stability.

MISA and Self-MM, are trained on complete observations. All methods are evaluated with the same test corruption masks and operator at each r, and the main evaluation protocol reports mean performance across all ten missing rates. Under fixed-modality missingness (FMM), all methods are trained on complete observations and evaluated under seven modality-availability conditions, including the complete-input condition. MRCF's paired complete-observation branch is used only during training and is absent at inference. Further corruption and input-handling details are provided in Appendix B.

3.3 Results under Intra-Modal Missingness

We evaluate robustness to random temporal erasure on CMU-MOSI, CMU-MOSEI, and CH-SIMS. As shown in Tables 1 and 2, MRCF achieves strong overall results on CMU-MOSI and CMU-MOSEI, while reaching 75.16% Acc-2 and 79.71% F1 on CH-SIMS, exceeding the reported P-RMF scores. It also obtains Corr scores of 0.552 and 0.606 on CMU-MOSI and CMU-MOSEI, respectively, demonstrating stronger consistency with continuous ground-truth sentiment under temporal corruption. Per-rate results are reported in Appendix B. Random sequence erasure disrupts local temporal continuity and induces reliability mismatch across modalities. Conventional fusion mechanisms may integrate corrupted streams without explicitly accounting for their unequal reliability, leading to reliability propagation bias. MRCF is designed to mitigate this issue. The Reliability-Aware Branch evaluates the internal quality of corrupted sequences and their cross-modal consistency. Guided by these estimates, the Reliability-Guided Interaction Branch reweights source messages and scales target-side updates according to estimated reliability. The resulting interaction is intended to reduce the influence of degraded streams at high missing rates.

Table 2: Robustness on SIMS under IMM. Results are averaged over $r \in \{0.0, \ldots, 0.9\}$.

Model	Acc-2	F1	Acc-3	MAE (↓)	Corr
MISA	72.71	66.30	56.87	0.539	0.348
Self-MM	72.81	68.43	56.75	0.508	0.376
MMIM	69.86	66.21	52.76	0.544	0.339
CENET	68.13	57.90	53.17	0.589	0.107
P-RMF	73.64	74.65	54.75	0.500	0.414
LNLN	72.73	–	57.14	0.514	0.397
MRCF	**75.16**	**79.71**	**57.97**	**0.481**	**0.435**

3.4 Results under Fixed-Modality Missingness

We evaluate robustness under fixed-modality missingness, where one or more modalities are unavailable during inference. As shown in Table 3, MRCF achieves the best average F1 score on both CMU-MOSI and CMU-MOSEI and performs strongly with audio or audio–vision inputs, although its vision-only results are lower than those of P-RMF. For instance, using only audio and vision modalities, MRCF reaches 84.47% F1 on CMU-MOSI, exceeding ROSA by 0.66 percentage points. These results support the effectiveness of MRCF under the evaluated incomplete-modality conditions.

3.5 Ablation Studies

Table 3: Results on CMU-MOSI and CMU-MOSEI under fixed-modality missingness (FMM). Here l, v, and a denote the available language, visual, and acoustic modalities. Average includes all seven conditions. Boldface highlights MRCF.

Model	Available Modalities (F1, %)							
	$\{l,v,a\}$	$\{l,v\}$	$\{l,a\}$	$\{v,a\}$	$\{l\}$	$\{v\}$	$\{a\}$	**Average**
CMU-MOSI								
Self-MM	84.64	74.97	69.81	47.12	67.80	38.52	40.95	60.54
DMD	84.50	68.45	70.51	50.47	68.97	42.26	43.33	61.21
NIAT	83.16	81.75	82.03	43.58	80.72	42.79	42.10	65.16
GCNet	83.20	83.58	84.73	70.02	80.91	58.70	65.07	75.17
IMDer	84.71	81.67	82.53	57.30	81.40	56.71	55.29	71.37
EMT	82.43	83.30	82.64	55.90	80.37	53.63	51.50	69.97
GMD	85.70	84.39	83.53	54.60	81.74	52.88	51.24	70.58
CorrKD	83.94	82.41	83.56	73.74	81.20	60.72	66.52	76.01
P-RMF	84.37	81.94	82.10	73.11	81.36	70.32	71.44	77.81
ROSA	86.30	85.39	85.13	83.81	83.64	56.79	81.91	80.42
MRCF (Ours)	**86.79**	**86.08**	**85.85**	**84.47**	**84.34**	**61.67**	**82.35**	**81.65**
CMU-MOSEI								
Self-MM	83.69	74.62	75.91	49.52	71.53	37.61	43.57	62.35
DMD	84.78	72.45	74.78	52.70	70.26	39.84	46.18	63.00
NIAT	86.35	85.31	85.10	48.57	83.29	45.76	44.43	68.40
GCNet	82.35	81.15	81.96	69.21	80.52	61.83	66.54	74.79
IMDer	84.71	82.39	83.41	53.30	82.77	51.83	50.47	69.84
EMT	85.17	84.45	83.81	52.39	83.04	50.79	51.28	70.13
GMD	87.07	85.52	83.10	52.67	81.45	49.96	50.39	70.02
CorrKD	82.16	81.28	81.74	71.92	80.76	62.30	66.09	75.18
P-RMF	85.48	85.17	84.61	76.88	81.91	73.19	75.91	80.45
ROSA	89.56	87.32	86.64	84.19	84.42	54.38	79.04	80.79
MRCF (Ours)	**89.92**	**87.68**	**87.15**	**85.72**	**84.80**	**56.77**	**79.54**	**81.65**

Table 4 presents the ablation studies on CMU-MOSI, CMU-MOSEI, and CH-SIMS under intra-modal missingness. The reported F1 and MAE values are averaged over missing rates from 0.1 to 0.9. **Importance of Core Components.** Removing both the Reliability-Aware Branch and the Reliability-Guided Interaction Branch causes the largest component-level degradation (e.g., CMU-MOSI F1 decreases from 73.31% to 69.90%), suggesting that their contributions are complementary. Removing only the interaction branch causes a larger decrease than removing the Reliability-Aware Branch alone. This indicates that estimating reliability is most useful when the estimates also regulate cross-modal message passing. Omitting the Reliability-Calibrated Fusion Module also reduces performance, supporting the importance of reliability integration at the final decision stage. **Effectiveness of Reliability Estimation.** Within the estimation strategy, removing cross-modal consistency causes a larger decline than removing the intramodal quality cue. This result suggests that semantic compatibility across modalities is an important component of the estimated reliability under incomplete observations. Removing the ranking loss also reduces performance. **Interaction and Fusion Variants.** MRCF outperforms

Table 4: Ablation studies under intra-modal missingness. w/o (R,G): w/o (RAB+RGIB). Metrics: F1 (%) and MAE (↓).

Variant	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	F1	**MAE↓**	**F1**	**MAE↓**	**F1**	**MAE↓**
MRCF	**73.31**	**0.986**	**80.27**	**0.642**	**79.71**	**0.481**
(1) Core Components						
w/o RAB	72.09	1.017	79.09	0.666	78.25	0.504
w/o RGIB	71.86	1.024	78.95	0.672	77.96	0.510
w/o (R,G)	69.90	1.063	77.42	0.694	76.35	0.535
w/o RCFM	72.36	1.009	79.36	0.656	78.63	0.494
(2) Reliability Estimation						
w/o Intra-Q	72.46	1.001	79.56	0.654	78.86	0.492
w/o Cross-C.	71.98	1.015	79.02	0.668	78.15	0.506
w/o $\mathcal{L}_{\text{rank}}$	72.76	0.997	79.74	0.650	79.02	0.489
(3) Interaction/Fusion Variants						
Uniform Inter.	71.22	1.033	78.53	0.676	77.54	0.514
Late Rel.	71.59	1.026	78.74	0.670	77.81	0.509
Addition	70.46	1.048	77.88	0.689	76.95	0.522
Concat+MLP	70.85	1.041	78.10	0.684	77.13	0.517
(4) Training Objectives						
w/o $\mathcal{L}_{\text{rel}}$	71.95	1.015	78.96	0.668	78.01	0.507
w/o $\mathcal{L}_{\text{prop}}$	72.28	1.010	79.31	0.660	78.46	0.499
w/o $\mathcal{L}_{\text{cal}}$	72.62	0.999	79.58	0.651	78.87	0.492

Uniform Interaction, indicating that reliability-aware message passing is useful relative to modality-agnostic interaction. Late-Reliability fusion also performs worse, supporting the use of reliability during interaction rather than only at final fusion. Addition and Concat+MLP yield lower results than the full model, further supporting the coupled interaction-fusion design. **Evaluation of Training Objectives.** Removing $\mathcal{L}_{\text{rel}}$ causes the largest decrease among the loss-term ablations. Omitting $\mathcal{L}_{\text{prop}}$ or $\mathcal{L}_{\text{cal}}$ also consistently reduces performance, supporting their respective roles in message-energy regularization and complete-observation calibration.

3.6 Further Analysis

Alignment with the Reliability Proxy. We examine whether the estimated score is aligned with the intended prediction-agreement criterion. For sample i, modality m, and missing rate r, we define the diagnostic $q_i^{(m)}(r) = \exp\left(-\left|\hat{y}_i^{(m,r)} - \hat{y}_i^{c}\right|\right)$, where $\hat{y}_i^{(m,r)}$ is the modality-m prediction at missing rate r, and $\hat{y}_i^{c}$ is the complete-observation prediction defined in Sec. 2.5. A larger $q_i^{(m)}(r)$ indicates closer prediction agreement. As shown in Fig. 3(a)(b), the estimated reliability $\rho_i^{(m)}$ is positively rank-correlated with $q_i^{(m)}(r)$ across CMU-MOSI and CMU-MOSEI. Because $q_i^{(m)}(r)$ follows the same agreement principle as the training proxy, this analysis measures alignment with the intended criterion rather than independently validating actual modality trustworthiness. This result is also consistent with the ablation in Table 4, where removing cross-modal consistency has a larger effect than removing the intramodal cue.

Behavior of Reliability-Guided Interaction. To examine how estimated reliability affects cross-modal interaction, we define the effective message strength as $S_i^{m\leftarrow n} = (1-\rho_i^{(m)})\alpha_i^{m\leftarrow n}\|\mathbf{U}_i^{m\leftarrow n}\|_F$ and compute the ratio of interaction from higher- to lower-scoring modalities to interaction in the reverse direction. As shown in Fig. 4(a), MRCF consistently yields higher directional interaction ratios than Uniform Interaction (near 1.0), increasing from 1.17 to 1.43 on MOSI as missingness grows. Fig. 4(b) further shows stronger routing from higher- to lower-scoring modalities under severe missingness, supporting reliability-aware rather than uniform semantic interaction. **Statistical Stability Analysis.** We evaluate CMU-MOSI across missing rates $r \in \{0.0, 0.1, \ldots, 0.9\}$ with three random seeds (Fig. 3(c)(d)). MRCF achieves a mean score of 73.31 and MAE of 0.986, while exhibiting tighter distributions than P-RMF. This demonstrates stable performance under random modality missingness. **Branch-weight Visualizations, Hyperparameter Sensitivity and Efficiency Comparisons are in Appendix** C, D, and E.

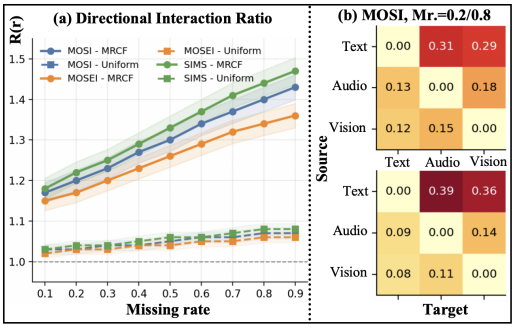

Figure 4: Reliability-guided cross-modal interaction. (a) Interaction ratios across missing rates on CMU-MOSI, CMU-MOSEI, and CH-SIMS, relative to the baseline (1.0). (b) CMU-MOSI routing at missing rates 0.2 and 0.8, showing stronger routing from higher- to lower-scoring modalities.

4 Conclusion

This paper investigates dynamically varying modality reliability and the resulting reliability mismatch and propagation bias in Multimodal Sentiment Analysis with incomplete observations. We propose MRCF, which explicitly models sample-specific modality reliability using intramodal quality and cross-modal consistency cues. The estimated reliability scores guide cross-modal message passing to suppress degraded evidence and are further incorporated into a reliability-calibrated fusion mechanism that aligns incomplete-observation representations with complete-observation anchors. Experiments across diverse multimodal benchmarks demonstrate that MRCF maintains robust performance under incomplete observations, with additional analyses validating its stability and efficiency. These results highlight reliability modeling as an effective principle for robust multimodal learning.

AI USE STATEMENT

In this work, we used generative AI tools for Paper polishing and grammar proofreading. We have not used generative AI tools for Method design and Experiment. The grammar of the text polished and proofread using LLM has been verified by the authors for accuracy. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies multimodal sentiment analysis using the existing CMU-MOSI, CMU-MOSEI, and CH-SIMS benchmarks. Predictions may reflect dataset biases and should not be treated as definitive assessments of individuals. Any downstream use should respect privacy, applicable dataset terms, and the limitations of automated sentiment inference.

REFERENCES

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 (Volume 1: Long Papers)*, pp. 2236–2246. Association for Computational Linguistics, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Peizhu Gong, Jin Liu, Xiliang Zhang, Xingye Li, Lai Wei, and Huihua He. Towards robust sentiment analysis with multimodal interaction graph and hybrid contrastive learning. *Pattern Recognition*, 169:111870, 2026.

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 (EMNLP)*, pp. 9180–9192, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pp. 1122–1131, 2020.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 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)*, pp. 12458–12468. IEEE, 2024.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6631–6640. IEEE, 2023.

Ziming Li, Yaxin Liu, Chuanpeng Yang, Yan Zhou, and Songlin Hu. ROSA: A robust self-adaptive model for multimodal emotion recognition with uncertain missing modalities. *IEEE Transactions on Multimedia*, 27:6766–6779, 2025.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. GCNet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8419–8432, 2023.

Chunlei Meng, Guanhong Huang, Rong Fu, Runmin Jian, Zhongxue Gan, and Chun Ouyang. CLCR: Cross-level semantic collaborative representation for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1606–1615, 2026a.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8791–8800, 2026b.

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, 2024a.

Teng Sun, Yinwei Wei, Juntong Ni, Zixin Liu, Xuemeng Song, Yaowei Wang, and Liqiang Nie. Muti-modal emotion recognition via hierarchical knowledge distillation. *IEEE Transactions on Multimedia*, 26:9036–9046, 2024b.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921, 2023a.

Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. Gradient-guided modality decoupling for missing-modality robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15483–15491, 2024.

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17117–17128, 2023b.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7617–7630. Association for Computational Linguistics, 2023.

Mingzheng Yang, Kai Zhang, Yuyang Ye, Yanghai Zhang, Runlong Yu, and Min Hou. Decoupling and reconstructing: A multimodal sentiment analysis framework towards robustness. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6803–6811, 2025.

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 (ACL)*, pp. 3718–3727. Association for Computational Linguistics, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 10790–10797, 2021.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4400–4407, 2021.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539, 2024.

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.

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 (EMNLP)*, pp. 756–767, 2023.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 55943–55974, 2024.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, pp. 22123–22138. Association for Computational Linguistics, 2025.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. CMAD: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4626–4636, 2025.

A Related Work

Research on Multimodal Sentiment Analysis (MSA) under complete observations has mainly advanced along the lines of representation decomposition and cross-modal fusion. Early methods such as MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), and CENET (Wang et al., 2023a) improve multimodal affect modeling by learning modality-invariant and modality-specific representations, exploiting self-supervised objectives, maximizing mutual information, or enhancing cross-modal interactions. ALMT (Zhang et al., 2023) further shows that language-guided adaptive modulation can suppress disruptive auxiliary cues and improve sentiment prediction. Although effective, these methods are developed for fully observed inputs and largely assume that multimodal evidence is available and comparably trustworthy during interaction and fusion.

For incomplete-observation MSA, one important line of work treats missing or corrupted inputs as a reconstruction problem. TFR-Net (Yuan et al., 2021) reconstructs missing features for non-aligned multimodal sequences, while DAR (Yang et al., 2025) decouples modality-common and modality-specific semantics and reconstructs them before fusion. Related robustness-oriented designs, such as NIAT (Yuan et al., 2024) and EMT (Sun et al., 2024a), also improve performance under degraded observations through corruption-aware or restoration-based learning. These methods substantially alleviate feature corruption, but their primary focus remains on recovering missing information, which tends to absorb modality reliability into reconstruction quality rather than modeling it explicitly during multimodal reasoning.

Another line of work emphasizes robust multimodal representation learning through distillation, dominant-modality preservation, and proxy-modality modeling. CorrKD (Li et al., 2024) transfers robust knowledge under uncertain missing modalities through distillation. LNLN (Zhang et al., 2024) emphasizes the dominant language modality to preserve reliable affective cues under incomplete data. P-RMF (Zhu et al., 2025) introduces proxy-modality learning in a latent Gaussian space and explicitly considers intrinsic modality uncertainty. Related studies such as CMAD (Zhuang et al., 2025), GMD (Wang et al., 2024), and ROSA (Li et al., 2025) further indicate that unequal modality contributions and unstable multimodal interactions are central issues under missing conditions. However, these methods mainly encode robustness through distillation, proxy construction, decoupling, or uncertainty-aware representation learning. In contrast, our method treats modality reliability as an explicit sample-level variable, estimates it jointly from intramodal quality and cross-modal semantic consistency, and uses it to regulate both cross-modal interaction and final fusion under incomplete observations.

B Experimental Protocol and Additional Results

B.1 Training and Evaluation Protocol Details

For intra-modal missingness (IMM), a fraction $r \in \{0.0, 0.1, \ldots, 0.9\}$ of temporal positions is randomly erased while sequence order is preserved. Following P-RMF (Zhu et al., 2025), missing acoustic and visual frames are zero-padded, and missing language tokens are replaced with `[UNK]`. At each target rate, methods supporting incomplete-input training, including MRCF, P-RMF, and LNLN, are trained and evaluated at that rate. Methods requiring complete training inputs, including MISA and Self-MM, are trained on uncorrupted observations. All methods are evaluated with the same test corruption masks and operator at the target rate. The main IMM protocol uses the ten-rate

Table 5: Detailed robustness evaluation results of MRCF on CMU-MOSI under different Intra-Modal Missing (IMM) rates. Avg. is computed over $r \in \{0, \ldots, 0.9\}$.

Missing Rate r	CMU-MOSI					
	Acc-2 (%)	F1 (%)	Acc-5 (%)	Acc-7 (%)	MAE (↓)	Corr
0.0	83.48 / 84.88	83.55 / 85.12	49.43	45.02	0.706	0.792
0.1	82.35 / 83.52	82.37 / 83.79	48.30	43.03	0.773	0.744
0.2	79.32 / 80.62	79.30 / 82.03	45.70	41.48	0.819	0.685
0.3	77.17 / 78.04	77.18 / 80.49	43.82	40.51	0.881	0.642
0.4	75.31 / 76.84	75.63 / 79.09	41.96	37.09	0.953	0.609
0.5	73.01 / 74.56	73.36 / 74.87	39.36	35.37	1.022	0.552
0.6	69.26 / 70.44	69.51 / 70.34	34.87	31.20	1.085	0.465
0.7	67.25 / 68.02	67.37 / 66.55	34.74	29.94	1.160	0.419
0.8	63.35 / 64.06	63.44 / 62.78	31.71	28.27	1.199	0.356
0.9	61.20 / 61.62	61.39 / 58.84	28.81	25.99	1.262	0.256
Avg.	**73.17 / 74.26**	**73.31 / 74.39**	**39.87**	**35.79**	**0.986**	**0.552**

grid $r \in \{0.0, 0.1, \ldots, 0.9\}$. The per-rate tables in this appendix report the available results for the missing rates $r \in \{0, 0.1, \ldots, 0.9\}$; each Avg. entry is computed only over these ten displayed rates.

During MRCF training, each corrupted input is paired with its complete counterpart only to construct the agreement-based reliability proxy and the fusion-calibration anchor. The complete-observation branch is not used at inference. Under fixed-modality missingness (FMM), all methods are trained on complete observations and evaluated under the seven modality-availability conditions reported in the main paper. For implementations requiring fixed-shape inputs, unavailable acoustic and visual streams are represented by zero tensors and unavailable language inputs by `[UNK]`. No method uses a complete test view when evaluated under missingness.

B.2 Additional Results

To complement the averaged IMM results reported in the main paper, we further present the available per-rate performance of MRCF at the missing rates listed in the following tables, including the complete-input condition. Detailed results on CMU-MOSI, CMU-MOSEI, and CH-SIMS are reported in Tables 5, 6, and 7, respectively. These results provide a more detailed view of robustness beyond the averaged scores and show how MRCF behaves as intra-modal corruption becomes progressively more severe.

Across all three benchmarks, the overall trend is consistent: performance gradually degrades as the missing rate increases, while MAE increases and correlation decreases accordingly. However, the degradation remains moderate in the low-to-moderate missingness regime. This behavior is consistent with the design of MRCF. The Reliability-Aware Branch (RAB) first estimates sample-level modality reliability from intramodal quality cues and cross-modal semantic consistency, allowing the model to distinguish partially preserved evidence from heavily corrupted evidence before deep multimodal interaction. As a result, when the missing rate is still moderate, MRCF can retain a substantial portion of useful affective information and avoid early performance deterioration.

As the missing rate increases further, the performance drop becomes more evident, especially when $r \geq 0.7$, where local temporal continuity and semantic completeness are both severely damaged. Even in this regime, however, MRCF maintains a controlled degradation pattern. This robustness stems from the Reliability-Guided Interaction Branch (RGIB), which explicitly regulates cross-modal message passing according to the estimated reliability. Instead of uniformly mixing corrupted streams, RGIB favors corrective guidance from more reliable modalities and suppresses misleading outgoing messages from less reliable ones. Consequently, the model is less vulnerable to reliability propagation bias under severe corruption.

Another notable observation is that the degradation of MRCF remains smooth across increasing missing rates. This pattern is consistent with the role of the Reliability-Calibrated Fusion Module (RCFM). By jointly modeling semantic representations and modality reliability and calibrating incomplete-observation fusion toward a complete-observation semantic anchor during training, RCFM is intended to reduce excessive changes in the final representation. Therefore, MRCF does not simply downweight uncertain modalities through a fixed heuristic; it uses estimated reliability during interaction and final fusion.

Overall, the detailed per-rate results describe how MRCF performance varies across the evaluated missing rates. Performance generally declines as missingness increases.

Table 6: Detailed robustness evaluation results of MRCF on CMU-MOSEI under different Intra-Modal Missing (IMM) rates. Avg. is computed over $r \in \{0, \ldots, 0.9\}$.

Missing Rate r	**CMU-MOSEI**					
	Acc-2 (%)	F1 (%)	Acc-5 (%)	Acc-7 (%)	MAE (↓)	Corr
0.0	83.81 / 86.06	83.42 / 85.90	53.64	51.77	0.535	0.773
0.1	83.06 / 85.08	82.72 / 84.94	53.33	51.50	0.548	0.756
0.2	82.60 / 84.30	82.39 / 84.08	51.55	50.82	0.566	0.733
0.3	81.31 / 82.84	81.24 / 82.71	50.30	49.40	0.598	0.696
0.4	80.27 / 81.78	80.45 / 81.73	49.57	49.31	0.616	0.669
0.5	79.23 / 80.67	79.64 / 80.62	48.06	48.20	0.648	0.619
0.6	78.11 / 78.40	78.90 / 78.85	46.60	46.97	0.683	0.566
0.7	76.62 / 77.17	78.08 / 77.29	46.21	46.89	0.710	0.504
0.8	75.29 / 73.53	77.94 / 74.58	45.12	46.07	0.739	0.427
0.9	73.50 / 70.87	77.92 / 73.20	44.32	45.67	0.777	0.317
Avg.	**79.38 / 80.07**	**80.27 / 80.39**	**48.87**	**48.66**	**0.642**	**0.606**

Table 7: Detailed robustness evaluation results of MRCF on CH-SIMS under different intra-modal missingness (IMM) rates. Avg. is computed over $r \in \{0, \ldots, 0.9\}$.

Missing Rate r	**CH-SIMS**					
	Acc-2 (%)	F1 (%)	Acc-3 (%)	Acc-5 (%)	MAE (↓)	Corr
0.0	79.05	83.85	63.02	39.20	0.431	0.562
0.1	78.13	83.24	62.11	38.05	0.442	0.544
0.2	77.30	82.02	62.07	37.47	0.446	0.528
0.3	76.91	80.65	59.84	37.38	0.459	0.501
0.4	76.75	81.26	58.93	37.11	0.463	0.492
0.5	75.23	80.00	58.15	35.45	0.475	0.426
0.6	73.91	78.68	56.54	34.65	0.493	0.407
0.7	73.55	77.78	56.19	33.05	0.507	0.395
0.8	71.67	76.12	53.74	31.96	0.541	0.297
0.9	69.10	73.50	49.11	30.08	0.553	0.198
Avg.	**75.16**	**79.71**	**57.97**	**35.44**	**0.481**	**0.435**

C Visualization of Reliability Weights and Branch Importance

To further examine how the three core branches of MRCF contribute under increasing missingness, we visualize branch-level performance-gain weights and module importance on CMU-MOSI and CMU-MOSEI across $r \in \{0.1, \ldots, 0.9\}$. As shown in Fig. 5, this analysis is branch-oriented and

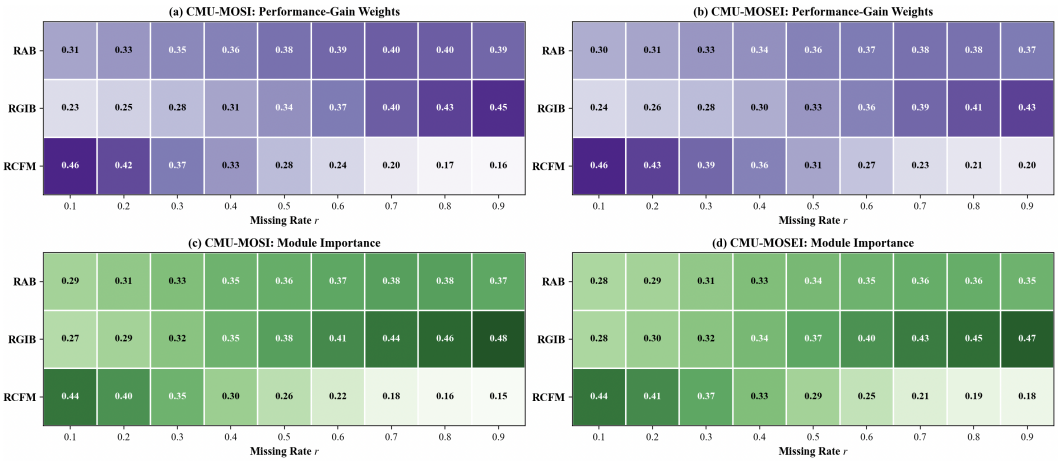

Figure 5: Visualization of branch-level performance-gain weights and module importance on CMU-MOSI and CMU-MOSEI across missing rates $r \in \{0.1, \ldots, 0.9\}$. Each heatmap is organized by branch (RAB, RGIB, and RCFM) and missing rate. The top row shows the normalized performance-gain weights obtained by sequentially introducing each branch into the MRCF pipeline, while the bottom row shows the normalized module importance measured by the relative performance drop after removing each branch from the full model. As missingness increases, the contributions of RAB and RGIB become progressively more pronounced, especially in high-missingness regimes where reliability mismatch and propagation bias are more severe. In contrast, RCFM remains consistently useful across all missing rates by stabilizing the final reliability-calibrated decision. These patterns support that the three branches are all necessary but play distinct roles in the full MRCF pipeline.

directly evaluates the roles of the Reliability-Aware Branch (RAB), the Reliability-Guided Interaction Branch (RGIB), and the Reliability-Calibrated Fusion Module (RCFM). Specifically, for each missing rate, we first compute the normalized performance gain obtained by sequentially introducing each branch into the pipeline and use it as the branch weight. We then compute the normalized performance drop caused by removing each branch from the full model and use it as the branch importance.

Several consistent patterns can be observed. At low missing rates, RCFM accounts for a relatively larger share, indicating that when multimodal evidence is still largely preserved, reliability-calibrated fusion plays a central role in organizing complementary information for prediction. As missingness increases, the contributions of RAB and RGIB generally increase. This suggests that under more severe corruption, accurate reliability estimation becomes increasingly necessary, while reliability-guided cross-modal correction becomes more important for preventing unreliable streams from contaminating healthier modalities. In particular, RGIB shows the highest importance in the high-missingness regime, which is consistent with the observation that propagation bias becomes more harmful when the retained evidence is sparse and unevenly trustworthy. Although RCFM becomes relatively less dominant at very high missing rates, its contribution remains consistently non-negligible across all stages, indicating that reliable final fusion is still necessary after reliability-aware interaction.

These observations are consistent with the main ablation results. They further show that the three branches of MRCF are functionally distinct rather than redundant. RAB identifies which modality remains trustworthy, RGIB determines how reliability regulates cross-modal correction, and RCFM stabilizes the final decision by integrating semantic and reliability cues. Therefore, the effectiveness of MRCF does not rely on a single dominant component, but on the coordinated operation of three tightly coupled branches with different roles.

D Sensitivity Analysis

We further examine whether the performance gains of MRCF depend on careful coefficient selection. Since the core mechanism of MRCF is realized through the reliability-aware objective, the

Table 8: Sensitivity analysis of λ_1 under IMM. Results are averaged over $r \in \{0, \ldots, 0.9\}$.

λ_1	MOSI F1 (%)	MOSI MAE ↓	MOSEI F1 (%)	MOSEI MAE ↓
0.2	72.58	1.006	79.66	0.653
0.5	73.16	0.987	80.16	0.647
1.0	**73.31**	**0.986**	**80.27**	**0.642**
1.5	73.25	0.991	80.25	0.646
2.0	72.86	0.998	79.89	0.651

Table 9: Sensitivity analysis of λ_2 under IMM. Results are averaged over $r \in \{0, \ldots, 0.9\}$.

λ_2	MOSI F1 (%)	MOSI MAE ↓	MOSEI F1 (%)	MOSEI MAE ↓
0.005	72.84	0.997	79.96	0.648
0.01	73.03	0.992	80.09	0.647
0.02	73.18	0.988	80.22	0.643
0.05	**73.31**	**0.986**	**80.27**	**0.642**
0.07	73.28	0.987	80.26	**0.642**
0.10	73.04	0.992	80.06	0.646

propagation-suppression term, the fusion-calibration term, and the ranking regularizer within reliability estimation, we conduct a sensitivity analysis by varying λ_1, λ_2, λ_3, and λ_r one at a time. Following the main paper, all experiments are performed under the IMM protocol on CMU-MOSI and CMU-MOSEI, and we report the average F1 score and MAE over missing rates $r \in \{0.1, \ldots, 0.9\}$. Unless otherwise specified, the default setting is fixed to $\lambda_1 = 1.0$, $\lambda_2 = 0.05$, $\lambda_3 = 0.10$, and $\lambda_r = 0.50$.

Impact of the reliability-aware weight λ_1. The coefficient λ_1 controls the overall strength of $\mathcal{L}_{\text{aware}}$, which is responsible for learning sample-level modality reliability before it is used by reliability-guided interaction and reliability-calibrated fusion. As shown in Table 8, performance improves when λ_1 increases from a small value to a moderate range, and the best overall trade-off is obtained around $\lambda_1 = 1.0$. This behavior is consistent with the role of the Reliability-Aware Branch. When λ_1 is too small, reliability learning is under-constrained, and the subsequent interaction and fusion stages cannot fully exploit reliable guidance. When λ_1 becomes too large, optimization is more strongly affected by the auxiliary reliability objective, which can slightly weaken task-oriented representation learning. The results indicate that MRCF benefits from sufficiently strong, but not excessive, reliability supervision.

Impact of the propagation-suppression weight λ_2. The coefficient λ_2 controls $\mathcal{L}_{\text{prop}}$, which penalizes high-energy outgoing messages from unreliable source modalities. Table 9 shows that the best results are achieved around $\lambda_2 = 0.05$, with a stable plateau in the range $[0.05, 0.07]$. This pattern is consistent with the design of the Reliability-Guided Interaction Branch. If λ_2 is too small, unreliable modalities are not sufficiently suppressed and can still contaminate healthier streams during cross-modal interaction. If λ_2 is too large, cross-modal communication becomes overly restricted, and weak but useful complementary cues may also be suppressed. Therefore, the optimal setting lies in a moderate interval where misleading evidence is reduced without harming beneficial interaction.

Impact of the fusion-calibration weight λ_3. The coefficient λ_3 determines how strongly the incomplete-observation fused representation is aligned with the complete-observation semantic anchor. As reported in Table 10, MRCF reaches its best performance around $\lambda_3 = 0.10$ and remains competitive in the interval $[0.10, 0.15]$. This trend is also consistent with the role of the Reliability-Calibrated Fusion Module. A small λ_3 leads to insufficient semantic anchoring, which makes the fused representation more vulnerable to instability under corrupted observations. A large λ_3, in contrast, over-constrains the model toward the complete-observation geometry and reduces the flexibility needed to adapt to genuinely incomplete inputs. The results suggest that a moderate calibration strength is sufficient to stabilize fusion while preserving adaptability.

Table 10: Sensitivity analysis of λ_3 under IMM. Results are averaged over $r \in \{0, \ldots, 0.9\}$.

λ_3	MOSI F1 (%)	MOSI MAE ↓	MOSEI F1 (%)	MOSEI MAE ↓
0.02	72.89	0.995	80.01	0.648
0.05	73.19	0.990	80.20	0.646
0.10	**73.31**	**0.986**	**80.27**	**0.642**
0.15	73.27	0.987	80.26	0.643
0.20	73.08	0.991	80.12	0.649

Table 11: Sensitivity analysis of λ_r under IMM. Results are averaged over $r \in \{0, \ldots, 0.9\}$.

λ_r	MOSI F1 (%)	MOSI MAE ↓	MOSEI F1 (%)	MOSEI MAE ↓
0.10	72.97	0.992	80.07	0.646
0.25	73.21	0.988	80.22	0.644
0.50	**73.31**	**0.986**	**80.27**	**0.642**
0.75	73.29	0.988	80.25	0.643
1.00	73.12	0.993	80.13	0.647

Impact of the ranking coefficient λ_r. The coefficient λ_r balances the pairwise ranking term inside $\mathcal{L}_{\text{aware}}$. As shown in Table 11, the best results are obtained around $\lambda_r = 0.50$, while the performance remains nearly unchanged in the range $[0.50, 0.75]$. This behavior indicates that the ranking term mainly plays a stabilizing role. With a very small λ_r, the model relies primarily on pointwise reliability regression and may produce less discriminative relative reliability ordering across modalities. When λ_r becomes too large, optimization places more emphasis on pairwise ordering and slightly weakens the absolute calibration of reliability values. The observed plateau around the default setting suggests that MRCF does not rely on finely tuned ranking strength to achieve robust performance.

Overall observation. Across the tested coefficient values, performance varies modestly around the default settings. The results characterize the evaluated ranges and do not establish a ranking of coefficient sensitivity or component importance.

Table 12: Efficiency analysis on CMU-MOSI. We report parameter counts and average training time per epoch. Baseline references: MISA (Hazarika et al., 2020), ConFEDE (Yang et al., 2023), P-RMF (Zhu et al., 2025), and LNLN (Zhang et al., 2024).

Model	**Params.**	**Time/Epoch**	F1(%)	MAE(↓)
MISA	112.45M	14.31 s	70.00	1.085
ConFEDE	246.98M	40.12 s	70.56	1.068
P-RMF	117.82M	18.62 s	71.69	1.038
LNLN	116.75M	24.58 s	71.25	1.046
MRCF	118.11M	20.06 s	**73.31**	**0.986**

E Efficiency Analysis

Table 12 evaluates the computational efficiency of MRCF on CMU-MOSI. With 118.11M parameters and a 20.06-second training time per epoch, MRCF is comparable to P-RMF and LNLN and smaller than ConFEDE in this comparison. It also achieves the highest reported F1 (73.31%) and the lowest MAE (0.986). The method avoids explicit feature reconstruction and distribution sampling; reliability estimation uses linear projections and cosine similarities, while interaction uses scalar reliability-dependent weights. The complete-observation anchors and proxy targets are used only during training, so the complete-observation branch is absent at inference. Because the table

does not report inference latency, memory, or floating-point operations, its efficiency conclusions are limited to parameter count and per-epoch training time.