HIGH UNCERTAINTY IS NOT ALWAYS HARMFUL: CONSISTENCY-GUIDED CROSS-MODAL UNCERTAINTY RECTIFICATION FOR ROBUST MULTIMODAL AFFECTIVE COMPUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Affective Computing (MAC) has attracted increasing attention in recent years, aiming to predict human affective states by exploiting complementary signals across different modalities. Despite substantial progress, existing methods remain vulnerable to noisy inputs, missing modalities, and the intrinsic uncertainty of multimodal data, limiting their reliability and robustness. In this paper, we propose a Consistency-guided Cross-modal Uncertainty Rectification (C^2UR) framework. Unlike existing approaches that primarily adjust the fusion contribution of each modality according to its uncertainty, C^2UR leverages cross-modal consistency to rectify unreliable yet task-informative modalities, thereby enabling robust prediction. Specifically, we first probabilistically model modality-specific representations and jointly estimate their uncertainty and task utility. We then adaptively rectify the distribution of each target modality by aggregating reliable evidence from the remaining modalities. Finally, we introduce a dual-branch evidence fusion mechanism that integrates the original modality information with the rectified cross-modal consensus, suppressing unreliable information while preserving task-relevant cues. Extensive experiments across four MAC tasks and diverse degradation scenarios demonstrate the effectiveness and robustness of C^2UR.

1 INTRODUCTION

As an important foundation of human-centric computing, Multimodal Affective Computing (MAC) aims to achieve a comprehensive understanding of human emotions by integrating complementary information across modalities and has attracted increasing research attention in recent years (Kim & Park, 2023; Mai et al., 2025; Li et al., 2024). Existing methods primarily focus on multimodal fusion, ranging from early fusion (Yu et al., 2021) at the input or feature level to late fusion (Cao et al., 2024) of modality-specific predictions at the decision level. While these methods effectively facilitate cross-modal interactions, they typically rely on relatively complete and reliable inputs. In real-world scenarios, missing modalities (Lian et al., 2023), sensor noise (Mai et al., 2023), and inherent uncertainty (Gao et al., 2024b; Hu et al., 2025) in multimodal data can substantially reduce modality reliability, limiting the robustness and effectiveness of MAC systems.

Recently, numerous studies have sought to address multimodal scenarios involving noisy or missing modalities. For example, some reconstruct missing modalities from the remaining observations (Guo et al., 2024), while others employ hyper-modality representation learning (Zhang et al., 2024a), diffusion models (Wang et al., 2023b), or attention mechanisms (Zhuang et al., 2026) for cross-modal knowledge transfer. However, these methods typically lack explicit modeling of input reliability and thus remain vulnerable to noise artifacts and unreliable information (Zhu et al., 2026). Reliable multimodal learning has therefore received increasing attention, focusing on explicitly quantifying modality uncertainty and dynamically adjusting each modality's contribution to fusion (Zhang et al., 2023b; Gao et al., 2024a; Zhu et al., 2025; Chen et al., 2026). This paradigm improves model trustworthiness and interpretability, supporting reliable decision-making in safety-critical applications such as autonomous driving (Tang et al., 2022).

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.

Motivation. However, modality uncertainty reflects representation reliability rather than task utility. As shown in Figure 1, highly uncertain modalities can still make positive task contributions. Notably, for audio and visual modalities, whose affective cues are often more ambiguous, the high-uncertainty group even shows a higher proportion of positive contributions than the low-uncertainty group. This suggests that simply down-weighting uncertain modalities may suppress valuable complementary information, motivating us to distinguish modality reliability from task utility.

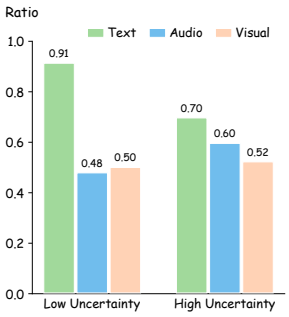


Figure 1: Proportion of samples with positive modality contributions in the lowest and highest uncertainty groups (bottom / top 30%) on CMU-MOSI.

Based on these observations, we propose C^{2}UR, a framework that distinguishes modality utility from representation uncertainty and rectifies uncertain representations through cross-modal consensus. The key idea is to preserve valuable information in uncertain yet task-relevant modalities using reliable cross-modal evidence. Specifically, C^{2}UR comprises three modules: ❶ *Modality Utility–Uncertainty Modeling.* We model each modality representation as a Gaussian distribution in a shared latent space and quantify uncertainty through posterior variance. Meanwhile, we derive conditional utility supervision from the task-loss change induced by modality removal and train a lightweight predictor to estimate each modality's task contribution, thereby modeling utility and uncertainty separately. ❷ *Consistency-guided Uncertainty Rectification.* We construct cross-modal consensus by aggregating probabilistic evidence from auxiliary modalities via a reliability-weighted Product-of-Experts (Hinton, 2002). The target modality's utility and uncertainty, together with auxiliary prediction consistency, adaptively determine the degree of posterior rectification. ❸ *Dual-branch Evidence Fusion.* We combine utility-weighted original evidence with reliability-aware rectified evidence, integrating shared cross-modal information while preserving modality-specific information for final affective prediction. Our main contributions are:

- We propose C^{2}UR, a framework that explicitly decouples modality task utility from representation uncertainty, preserving valuable task information in uncertain yet informative modalities.
- We introduce consistency-guided posterior rectification to adaptively refine uncertain representations using reliable cross-modal consensus, followed by dual-branch fusion of original and rectified evidence to learn robust multimodal representations.
- Extensive experiments on four MAC tasks demonstrate that C^{2}UR achieves state-of-the-art performance. We further evaluate C^{2}UR under continuous combinations of noise and modality missingness to reflect real-world degradation, demonstrating greater robustness than leading baselines.

2 Related Work

2.1 Robust Multimodal Affective Computing

Existing MAC research primarily focuses on modality fusion and representation learning. The former models cross-modal interactions to integrate complementary information. For example, tensor fusion (Zadeh et al., 2017) explicitly captures cross-modal interactions, while Transformer-based architectures (Tsai et al., 2019; Zhang et al., 2023a) use cross-modal attention to align asynchronous sequences. The latter aims to learn more affect-discriminative representations, such as by reducing redundancy through disentanglement (Hazarika et al., 2020; Yang et al., 2022; Qian et al., 2026) or incorporating external affective knowledge (Zhu et al., 2024). However, most existing methods assume complete and noise-free modalities, limiting their applicability in real-world scenarios.

Robust multimodal learning has therefore received increasing attention. For noisy inputs, existing methods often employ information-theoretic mechanisms, such as the information bottleneck (Tishby et al., 2000), to suppress irrelevant information (Mai et al., 2023; Xiao et al., 2024; Luo et al., 2025). For missing modalities, they typically reconstruct missing information from the observed modalities (Zhao et al., 2021; Li et al., 2024; Zhuang et al., 2025b). For example, Guo et al. (2024) use generative and missing-aware prompts to recover missing modality features and adapt pretrained

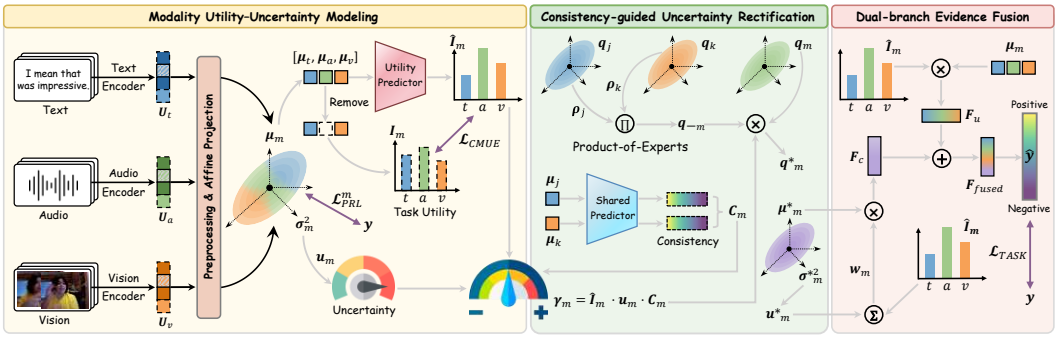


Figure 2: Illustration of the proposed C^{2}UR framework. It consists of three modules: (A) Modality Utility–Uncertainty Modeling Module, (B) Consistency-guided Uncertainty Rectification Module, and (C) Dual-branch Evidence Fusion Module.

multimodal models, while Wang et al. (2023a) use class-specific flow-based generators to recover missing modalities. However, these methods do not explicitly model representation reliability, making it difficult to distinguish task-relevant semantics from uninformative noise and potentially weakening robust affective representation learning.

2.2 Uncertainty-aware Learning

Uncertainty modeling is an important component of reliable AI systems, particularly in safety-critical applications (Hu et al., 2025). It is commonly categorized into epistemic uncertainty, arising from model limitations, and aleatoric uncertainty, stemming from inherent data noise (Kendall & Gal, 2017). Representative approaches include Bayesian neural networks (Kendall & Gal, 2017), Monte Carlo sampling (Zhang, 2021), ensemble learning (Lakshminarayanan et al., 2017), and variational probabilistic representation learning (Shi & Jain, 2019). Recent studies have incorporated uncertainty modeling into multimodal learning to improve robustness (Zhang et al., 2023b; Gao et al., 2024b; Chen et al., 2026). For example, Gao et al. (2024a) employ an uncertainty-enhanced mixture of experts to mitigate noise; Zhuang et al. (2025a) introduce uncertainty-aware fusion to adaptively balance original and hyper-modality enhanced representations; and Zhu et al. (2026) propose a quality-aware mixture-of-experts framework based on a continuous reliability spectrum to handle heterogeneous noise and missing modalities across samples. However, these methods do not explicitly distinguish representation reliability from task utility and may therefore suppress uncertain yet informative modalities. In contrast, we decouple these two factors and use reliable cross-modal consensus to rectify uncertain representations while preserving their task contributions.

3 Methodology

The overall architecture of the proposed C^{2}UR framework is illustrated in Figure 2. It comprises three modules. The Modality Utility–Uncertainty Modeling (MUM) module learns probabilistic modality representations and separately estimates their task utility and representation uncertainty (Section 3.2). The Consistency-guided Uncertainty Rectification (CUR) module constructs reliable cross-modal consensus and adaptively rectifies the target-modality posterior based on modality utility, uncertainty, and prediction consistency (Section 3.3). Finally, the Dual-branch Evidence Fusion (DEF) module combines utility-weighted original modality evidence with reliability-aware rectified common evidence for robust prediction (Section 3.4).

3.1 Problem Statement

MAC aims to infer a speaker's affective state from video segments comprising textual (t), acoustic (a), and visual (v) modalities. Given a training dataset $\mathcal{D} = \{(\boldsymbol{U}_t^i, \boldsymbol{U}_a^i, \boldsymbol{U}_v^i, y^i)\}_{i=1}^N$, where N denotes the number of training samples and y^i denotes the affective label of the i-th sample, our goal is to learn a parameterized multimodal model $\mathcal{M}_\phi$ to predict the affective state from multimodal inputs:

$\hat{y}^i = \mathcal{M}_\phi(\boldsymbol{U}_t^i, \boldsymbol{U}_a^i, \boldsymbol{U}_v^i)$, where $\hat{y}^i$ represents either a continuous affect intensity score for regression tasks or a discrete affective category for classification tasks, depending on the task. For each modality $m \in \{t, a, v\}$, the input feature sequence is denoted by $\boldsymbol{U}_m \in \mathbb{R}^{l_m \times d_m}$, where l_m and d_m denote the sequence length and feature dimension of modality m, respectively.

For each modality m, we employ a modality-specific encoder to extract features from the input sequence $\boldsymbol{U}_m$, yielding a unimodal representation $\boldsymbol{X}_m \in \mathbb{R}^d$, where d denotes the unified representation dimension. Details of the modality-specific encoders are provided in Appendix A.1.

3.2 Modality Utility–Uncertainty Modeling

To distinguish modality task contribution from representation reliability, we propose the MUM module, consisting of two key components: Probabilistic Representation Learning (PRL) and Conditional Modality Utility Estimation (CMUE). We describe each component in detail below.

Probabilistic Representation Learning. To model the uncertainty associated with different modalities and establish a unified parameter space for subsequent cross-modal probabilistic evidence aggregation, we map each modality-specific representation to a probabilistic estimate of a shared task-relevant latent variable through shared Gaussian parameterization networks. Specifically, for each modality $m \in \{t, a, v\}$, the mean and variance are parameterized as

$$\boldsymbol{\mu}_m = f_\mu(\boldsymbol{X}_m), \qquad \boldsymbol{\sigma}_m^2 = Softplus(f_\sigma(\boldsymbol{X}_m)) + \epsilon, \tag{1}$$

where $f_\mu(\cdot)$ and $f_\sigma(\cdot)$ denote the mean and variance parameterization networks shared across the three modalities, respectively, and $\epsilon > 0$ is a small constant introduced for numerical stability. We then model the modality-specific latent variable $\boldsymbol{Z}_m$ as a conditional multivariate Gaussian distribution:

$$q(\boldsymbol{Z}_m \mid \boldsymbol{X}_m) = \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \qquad \boldsymbol{\Sigma}_m = diag(\boldsymbol{\sigma}_m^2). \tag{2}$$

Here, $\boldsymbol{\mu}_m$ represents the modality-specific central estimate of the shared task-relevant information, whereas $\boldsymbol{\Sigma}_m$ captures the uncertainty associated with this estimate. Based on the average posterior variance, we further quantify the uncertainty of modality m as

$$u_m = 1 - \exp\left(-\frac{1}{d}\,\mathrm{Tr}(\boldsymbol{\Sigma}_m)\right), \tag{3}$$

where $u_m \in [0, 1)$. A larger u_m indicates a less reliable probabilistic representation for the corresponding modality, potentially due to noisy or missing inputs. Thus, u_m serves as a dynamic quality indicator for guiding the subsequent fusion process.

However, probabilistic modeling alone does not ensure that the learned latent representations preserve task-discriminative information. To encourage each modality to retain task-relevant semantics, we therefore impose unimodal task supervision using a prediction head $g_u(\cdot)$ shared across modalities:

$$\mathcal{L}_{PRL}^m = \mathbb{E}_{q(\boldsymbol{Z}_m|\boldsymbol{X}_m)}\left[\mathcal{L}_{TASK}\left(g_u(\boldsymbol{Z}_m), y\right)\right] + \beta D_{KL}\left(q(\boldsymbol{Z}_m \mid \boldsymbol{X}_m) \,\|\, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\right), \tag{4}$$

where β controls the strength of the KL regularization. During training, $\boldsymbol{Z}_m$ is sampled from $q(\boldsymbol{Z}_m \mid \boldsymbol{X}_m)$ using the reparameterization trick (Kingma & Welling, 2013). Further details of the Gaussian parameterization, reparameterization procedure, and closed-form KL divergence are provided in the Appendix A.2.

Conditional Modality Utility Estimation. Modality uncertainty characterizes the reliability of a probabilistic representation, but does not directly reflect the modality's contribution to task prediction. In particular, a highly uncertain modality may still contain important modality-specific or complementary information. We therefore estimate the conditional task utility of each modality from the change in task loss caused by its removal, using the mean representations $\boldsymbol{\mu}_m$. Let $\boldsymbol{\mu} = \{\boldsymbol{\mu}_t, \boldsymbol{\mu}_a, \boldsymbol{\mu}_v\}$, and let $\mathcal{F}_\theta(\cdot)$ and $g(\cdot)$ denote the multimodal fusion network and task prediction head, respectively. We use simple concatenation for multimodal fusion and project the concatenated features to a d-dimensional representation through an MLP before prediction. Given the complete multimodal input, we compute

$$\boldsymbol{F} = \mathcal{F}_\theta(\boldsymbol{\mu}), \qquad \widehat{y} = g(\boldsymbol{F}), \qquad \ell = \mathcal{L}_{TASK}(\widehat{y}, y). \tag{5}$$

For each modality $m \in \{t, a, v\}$, we replace $\boldsymbol{\mu}_m$ with a zero vector and recompute the prediction using the same networks:

$$\boldsymbol{F}_{-m} = \mathcal{F}_\theta(\boldsymbol{\mu}_{-m}), \qquad \widehat{y}_{-m} = g(\boldsymbol{F}_{-m}), \qquad \ell_{-m} = \mathcal{L}_{TASK}(\widehat{y}_{-m}, y), \tag{6}$$

where $\boldsymbol{\mu}_{-m}$ denotes the multimodal mean representations with modality m removed. We define the resulting positive increase in task loss as the conditional task utility of modality m: $\mathcal{V}_m = [\ell_{-m} - \ell]_+$, $[x]_+ = \max(x, 0)$. A larger $\mathcal{V}_m$ indicates that removing modality m causes a greater performance degradation, implying a larger task contribution conditional on the remaining modalities. We normalize these utilities using a temperature-controlled Softmax: $I_m = \exp(\mathcal{V}_m/\tau_I) \Big/ \left(\sum_{j\in\{t,a,v\}} \exp(\mathcal{V}_j/\tau_I)\right)$, where $\tau_I > 0$ is the temperature parameter.

Since $\boldsymbol{I} = [I_t, I_a, I_v]$ depends on the ground-truth label, it is used only during training as supervision for modality utility estimation. To avoid repeated modality removal at inference time, we introduce a lightweight predictor $G_I(\cdot)$ that directly estimates the relative utilities of the three modalities from their mean representations: $\widehat{\boldsymbol{I}} = Softmax(G_I([\boldsymbol{\mu}_t; \boldsymbol{\mu}_a; \boldsymbol{\mu}_v]))$. The predictor is trained to match the removal-based utility distribution:

$$\mathcal{L}_{CMUE} = \frac{1}{B}\sum_{i=1}^{B} D_{KL}\left(SG(\boldsymbol{I}^i) \parallel \widehat{\boldsymbol{I}}^i\right) = \frac{1}{B}\sum_{i=1}^{B}\sum_{m\in\{t,a,v\}} SG(I_m^i)\log\frac{SG(I_m^i)}{\widehat{I}_m^i}, \tag{7}$$

where B is the batch size and $SG(\cdot)$ denotes the stop-gradient operator, ensuring that the removal-based utility distribution serves only as the supervision target.

This design yields the predicted modality utility $\widehat{I}_m$ and uncertainty score u_m, which capture task contribution and representation reliability, respectively. Since a high-utility modality can still be highly uncertain, directly suppressing it may discard valuable information. We therefore selectively rectify the uncertainty of such modalities using reliable cross-modal evidence, as described next.

3.3 Consistency-guided Uncertainty Rectification

Although a modality with both high task utility and high uncertainty can contribute substantially to the task, existing uncertainty-aware learning tends to suppress its contribution due to low reliability, potentially discarding valuable task-relevant information. We therefore rectify its uncertainty rather than directly downweight it. Since auxiliary modalities may suffer from noise or missing information, we assess the cross-modal evidence for rectification from two complementary perspectives: auxiliary modality reliability and task-prediction consistency.

Reliable Cross-modal Consensus. For a target modality m, we first define the reliability of each auxiliary modality based on its uncertainty: $\rho_j = 1 - u_j$, $j \neq m$. Since $u_j \in [0, 1)$, we have $\rho_j \in (0, 1]$, such that reliable modalities contribute more strongly, while highly uncertain modalities are adaptively downweighted.

However, reliability alone does not guarantee consistent task evidence. Forcing inconsistent cross-modal evidence into the target modality may introduce misleading information and lead to negative transfer. For target modality m, let j and k denote the two auxiliary modalities. We obtain their task predictions using the shared prediction head and measure their consistency in the task output space:

$$\widetilde{y}_j = g_u(\boldsymbol{\mu}_j), \qquad \widetilde{y}_k = g_u(\boldsymbol{\mu}_k), \qquad C_m = \exp\left(-\frac{d_{TASK}\left(\widetilde{y}_j, \widetilde{y}_k\right)}{\tau_C}\right), \tag{8}$$

where $\tau_C > 0$ is a temperature parameter and $d_{TASK}(\cdot, \cdot)$ denotes a task-specific prediction distance. We use the squared Euclidean distance for regression and the Jensen–Shannon divergence (Lin, 1991) between predicted probability distributions for classification. Thus, $C_m \in (0, 1]$, where a larger value indicates stronger agreement between the auxiliary modalities.

Since all modality-specific distributions estimate a common task-relevant latent variable, we construct a cross-modal consensus distribution for target modality m using a reliability-weighted Product-of-Experts (Hinton, 2002). Let $\mathcal{A}_m = \{t, a, v\} \setminus \{m\}$ denote the set of auxiliary modalities. The consensus distribution is defined as

$$q_{-m}(\boldsymbol{z}) = \frac{1}{\mathcal{Z}_{-m}} \prod_{r\in\mathcal{A}_m} \left[q(\boldsymbol{z} \mid \boldsymbol{X}_r)\right]^{\rho_r}, \qquad \rho_r = 1 - u_r, \tag{9}$$

where $\mathcal{Z}_{-m}$ is the normalization constant. This formulation aggregates the probabilistic evidence from all auxiliary modalities according to their reliability, allowing reliable modalities to exert greater

influence on the consensus distribution while adaptively suppressing uncertain ones. Meanwhile, C_m captures the predictive agreement among the auxiliary modalities, providing a complementary signal for determining whether the resulting consensus is sufficiently reliable for uncertainty rectification.

Adaptive Posterior Rectification. Given the cross-modal consensus distribution, we determine the degree of posterior rectification based on the task utility and uncertainty of the target modality, as well as the predictive consistency of the auxiliary modalities. For target modality m, we define the rectification coefficient as $\gamma_m = \widehat{I}_m \cdot u_m \cdot C_m$, where $\widehat{I}_m$, u_m, and C_m denote the predicted task utility, representation uncertainty, and auxiliary-modal consistency, respectively. Thus, cross-modal evidence contributes strongly to rectification only when the target modality is both important and uncertain, while the auxiliary modalities provide consistent evidence.

We treat the cross-modal consensus distribution as additional probabilistic evidence and rectify the target posterior using a weighted Product-of-Experts:

$$q_m^*(\boldsymbol{z}) = \frac{1}{\mathcal{Z}_m^*} q(\boldsymbol{z} \mid \boldsymbol{X}_m) \left[q_{-m}(\boldsymbol{z})\right]^{\gamma_m}, \tag{10}$$

where $\mathcal{Z}_m^*$ is the normalization constant. When γ_m is small, the rectified posterior largely preserves the target modality's own probabilistic evidence; as γ_m increases, reliable cross-modal evidence exerts a stronger influence on the posterior.

Since all modality-specific probabilistic representations are Gaussian, this update admits a closed-form solution. Let $q_m^*(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu}_m^*, \boldsymbol{\Sigma}_m^*)$. The rectified posterior covariance satisfies $\boldsymbol{\Sigma}_m^* \preceq \boldsymbol{\Sigma}_m$, where $\preceq$ denotes the positive semidefinite ordering. Therefore, incorporating cross-modal evidence cannot increase the target modality's posterior uncertainty. Since our uncertainty measure increases monotonically with the posterior covariance, we further obtain $u_m^* \leq u_m$. The closed-form solution of the weighted Product-of-Experts and its proof are provided in the Appendix A.3.

3.4 Dual-branch Evidence Fusion

Cross-modal consensus primarily captures probabilistic evidence jointly supported by multiple modalities. Replacing the original modality representations with their rectified counterparts may therefore bias modality-specific or complementary task information toward the consensus. We thus retain both original and rectified modality evidence through two branches.

The first branch weights the original modality means by their estimated utilities and fuses them through a Transformer-based network:

$$\boldsymbol{F}_u = \mathcal{F}_u\left(\widehat{I}_t\boldsymbol{\mu}_t, \widehat{I}_a\boldsymbol{\mu}_a, \widehat{I}_v\boldsymbol{\mu}_v\right), \tag{11}$$

where $\mathcal{F}_u$ applies a two-layer Transformer encoder to the three weighted modality tokens, concatenates the outputs, and projects them through a two-layer MLP. This branch retains the original task-relevant evidence from each modality according to its conditional utility.

The second branch aggregates the rectified modality means using weights that combine task utility and rectified reliability:

$$\boldsymbol{F}_c = \sum_{m\in\{t,a,v\}} w_m \boldsymbol{\mu}_m^*, \qquad w_m = \frac{\widehat{I}_m(1-u_m^*)}{\sum_{j\in\{t,a,v\}} \widehat{I}_j(1-u_j^*)}, \tag{12}$$

where $\boldsymbol{\mu}_m^*$ and u_m^* denote the posterior mean and uncertainty after cross-modal rectification. These weights favor modalities with higher task utility and greater rectified reliability. Finally, we concatenate the two branch representations for affective prediction:

$$\boldsymbol{F}_{fused} = \mathcal{F}_f\left([\boldsymbol{F}_u; \boldsymbol{F}_c]\right), \qquad \widehat{y} = g(\boldsymbol{F}_{fused}), \tag{13}$$

where $\mathcal{F}_f$ consists of a linear layer followed by ReLU and dropout, and g is a linear prediction head.

3.5 Training Objective

The overall training objective of C^{2}UR is defined as

$$\mathcal{L}_{All} = \mathcal{L}_{TASK}(\widehat{y}, y) + \lambda_1\mathcal{L}_{PRL} + \lambda_2\mathcal{L}_{CMUE}, \tag{14}$$

where $\mathcal{L}_{PRL} = \frac{1}{3}\sum_{m\in\{t,a,v\}} \mathcal{L}^m_{PRL}$ averages the probabilistic representation learning objective across the three modalities, and λ_1 and λ_2 balance the two auxiliary objectives. For classification and regression tasks, the task loss is defined as

$$\mathcal{L}_{TASK}(\widehat{y}, y) = \begin{cases} CrossEntropy(\widehat{y}, y), & classification \\ |\widehat{y} - y|, & regression \end{cases}. \tag{15}$$

4 EXPERIMENTS

Table 1: Comparison on CMU-MOSI and CMU-MOSEI. ♡ indicates results reported in the original papers. The best results are highlighted in **bold**.

Method	CMU-MOSI					CMU-MOSEI				
	Acc7%↑	*Acc2%*↑	*F1%*↑	*MAE*↓	*Corr*↑	*Acc7%*↑	*Acc2%*↑	*F1%*↑	*MAE*↓	*Corr*↑
MMRest♡ [CVPR'26]	48.98	86.78	86.89	0.683	0.802	54.72	86.88	86.89	0.531	0.769
PSA-MF♡ [AAAI'26]	46.50	86.43	86.19	0.686	0.807	55.00	86.30	86.28	0.521	0.774
DecAlign♡ [ICLR'26]	45.07	85.75	85.82	0.735	0.811	55.02	86.48	86.07	0.543	0.768
GLoMo [MM'24]	47.59	87.94	87.82	0.634	0.848	53.48	86.38	86.38	0.547	0.768
ITHP [ICLR'24]	46.42	88.09	88.09	0.652	0.848	52.40	86.74	86.68	0.537	0.773
KAN-MCP [MM'25]	46.57	88.55	88.54	0.658	0.841	53.18	86.99	87.01	0.547	0.772
TMDC [AAAI'26]	45.69	88.70	88.75	0.646	0.852	52.68	87.10	87.01	0.540	0.789
EUAR [MM'24]	47.01	87.33	87.35	0.654	0.838	53.97	86.35	86.39	0.535	0.779
PML [IJCAI'25]	47.74	87.94	87.94	0.649	0.841	52.96	86.44	86.46	0.545	0.767
HME [NeurIPS'25]	48.03	88.55	88.47	0.637	0.848	54.40	87.07	86.97	0.523	0.789
QA-MoE [ACL'26]	48.47	88.55	88.55	0.630	0.847	54.08	87.18	87.22	0.517	0.789
C^{2}UR (ours)	**49.49**	**89.92**	**89.93**	**0.618**	**0.855**	**55.46**	**87.54**	**87.49**	**0.505**	**0.795**

4.1 EXPERIMENTAL SETUP

We evaluate C^{2}UR on four MAC tasks: Multimodal Sentiment Analysis (MSA), Multimodal Humor Detection (MHD), Multimodal Sarcasm Detection (MSD), and Multimodal Emotion Recognition (MER). In addition to standard evaluation, we assess its robustness across a continuous reliability spectrum by jointly varying noise intensity and missing rate, using a single trained checkpoint.

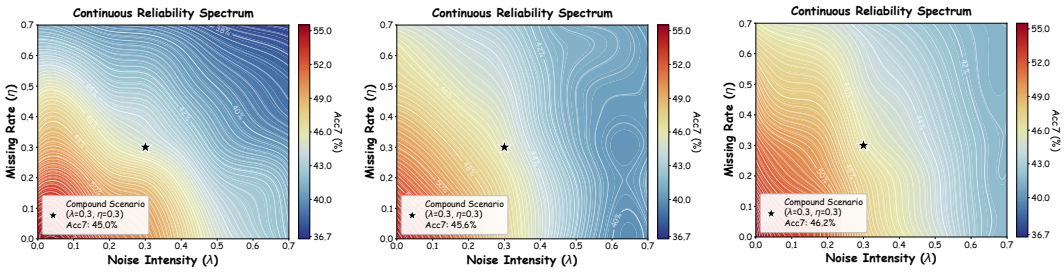
Datasets. For MSA, we conduct experiments on CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018). For MHD, MSD, and MER, we adopt UR-FUNNY (Hasan et al., 2019), MUStARD (Castro et al., 2019), and CHERMA (Sun et al., 2023), respectively. These datasets enable a comprehensive assessment of the applicability of C^{2}UR. Details are provided in Appendix B.1.

Baselines. We compare C^{2}UR with recent state-of-the-art MSA and multimodal learning methods. In particular, we include uncertainty-aware approaches EUAR (Gao et al., 2024a), PML (Hu et al., 2025), HME (Zhuang et al., 2025a), and QA-MoE (Zhu et al., 2026), with HME and QA-MoE also used in the continuous reliability spectrum evaluation. For MHD, MSD, and MER, we adopt representative task-specific baselines. Detailed descriptions of all baselines are provided in Appendix B.2.

Due to space limitations, details on feature extraction, implementation, and evaluation metrics are provided in Appendix B. Additional results on MHD, MSD, and MER are presented in Appendix C.

4.2 QUANTITATIVE RESULTS

Obs.❶ C^{2}UR achieves state-of-the-art performance across four MAC tasks. As shown in Table 1, C^{2}UR consistently outperforms the compared methods across all metrics on CMU-MOSI and CMU-MOSEI, demonstrating its effectiveness in both sentiment polarity classification and fine-grained intensity prediction. In particular, it surpasses recent uncertainty-aware methods. Additional results on MHD, MSD, and MER in Appendix C.1–C.3 further show its superiority over task-specific baselines, demonstrating its applicability across diverse affective tasks. By selectively rectifying uncertain representations with reliable cross-modal evidence while preserving task-relevant cues, C^{2}UR more effectively exploits multimodal information. These results highlight the benefit of leveraging informative evidence from uncertain modalities rather than simply suppressing them.



(a) HME (b) QA-MoE (c) C^{2}UR (ours)

Figure 3: **Continuous Reliability Spectrum.** Comparison of C^{2}UR with two state-of-the-art baselines, HME (Zhuang et al., 2025a) and QA-MoE (Zhu et al., 2026), across continuous combinations of noise intensity (λ) and missing rate (η). C^{2}UR preserves a broader high-performance region and exhibits smoother, more graceful degradation as input reliability decreases. The star ($\star$) marks the compound-degradation setting ($\lambda = 0.3$, $\eta = 0.3$).

Obs.❷ C^{2}UR maintains robust performance across the continuous reliability spectrum. In real-world scenarios, modality noise and missingness may occur simultaneously, leading to compound and continuously varying reliability degradation. Following the recommendation of Zhu et al. (2026), we therefore evaluate a single trained checkpoint over the entire reliability spectrum grid without further tuning. Figure 3 visualizes Acc7 under jointly varying noise intensity λ and missing rate η, comparing C^{2}UR with the uncertainty-aware methods HME (Zhuang et al., 2025a) and QA-MoE (Zhu et al., 2026). C^{2}UR maintains a broader high-performance region under mild-to-moderate degradation and degrades more smoothly as input reliability decreases. It also achieves the best performance at the compound-degradation setting marked by $\star$. These results suggest that consistency-guided rectification better preserves useful multimodal evidence under jointly noisy and missing inputs, improving robustness across varying reliability conditions without additional optimization. Detailed results are provided in Appendix C.4.

4.3 Ablation Study

Table 2: Ablation results on CMU-MOSI and CMU-MOSEI. Each entry reports Acc7 and MAE.

Dataset	w/o CMUE	w/o UUD	w/o CUR	w/o CG	w/o UG	w/o UB	C^{2}UR
CMU-MOSI	48.61 / 0.641	45.55 / 0.655	46.72 / 0.641	49.20 / 0.644	45.84 / 0.652	43.80 / 0.675	**49.49 / 0.618**
CMU-MOSEI	54.19 / 0.528	54.17 / 0.524	54.00 / 0.512	53.84 / 0.523	54.40 / 0.512	52.92 / 0.525	**55.46 / 0.505**

To examine the contributions of C^{2}UR, we ablate key components of MUM, CUR, and DEF: (i) 'w/o CMUE' replaces predicted utilities with uniform weights and removes utility supervision; (ii) 'w/o UUD' disables utility–uncertainty decoupling by replacing utility with normalized reliability; (iii) 'w/o CUR' bypasses posterior rectification; (iv) 'w/o CG' removes consistency gating; (v) 'w/o UG' removes the target-uncertainty gate; and (vi) 'w/o UB' removes the utility-weighted original-evidence branch. Table 2 reports results. Full metrics and additional ablation results are provided in Appendix C.5.

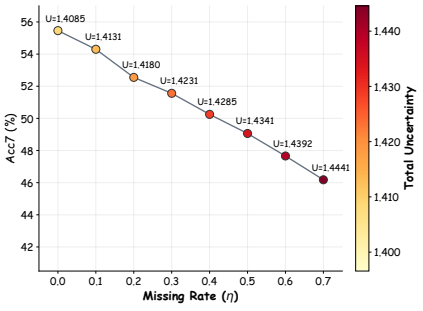


Figure 4: Uncertainty under modality missingness on CMU-MOSEI. Acc7 at varying missing rates η. Colors and annotations indicate the mean total uncertainty before rectification.

Obs.❸ The components of C^{2}UR are complementary and jointly contribute to its performance. The complete model outperforms all listed variants in both Acc7 and MAE. Removing UB causes the largest Acc7 degradation on both datasets, indicating that rectified consensus alone cannot replace modality-specific task evidence. Removing CMUE also consistently degrades performance, showing the importance of task-specific utility estimation over uniform modality weighting. Disabling UUD

further weakens fine-grained prediction, particularly on CMU-MOSI. This supports distinguishing a modality's task contribution from its reliability: uncertain evidence may remain informative, and reliability alone may assign it insufficient influence. Removing CUR degrades both metrics, consistent with the benefit of refining uncertain representations using auxiliary evidence rather than only adjusting their fusion weights. Removing UG and CG further degrades performance, indicating that effective rectification requires both target-uncertainty gating and cross-modal consistency. Overall, these results support the complementary roles of utility estimation, selective posterior rectification, and preservation of original evidence.

4.4 Further Analysis

We provide additional analyses in the appendix, including parameter sensitivity analysis (Appendix C.6), testing stability analysis (Appendix C.7), and model complexity analysis (Appendix C.8).

Obs.❹ Estimated uncertainty reflects declining reliability. To assess whether uncertainty can serve as a reliability indicator, we vary the missing rate η on CMU-MOSEI to simulate declining multimodal data quality. As shown in Figure 4, the mean total uncertainty before rectification rises steadily with missingness, while Acc7 decreases. This trend suggests that the estimated uncertainty reflects the loss of reliable information, supporting its use as a reliability indicator for posterior rectification and evidence fusion.

Obs.❺ Cross-modal rectification reduces uncertainty while preserving predictive utility. To examine whether C^{2}UR benefits uncertain yet informative modalities, we select CMU-MOSI samples with both high text modality uncertainty and high task utility. As shown in Figure 5, rectification consistently lowers text modality uncertainty and reduces absolute prediction error for most selected samples. Although uncertainty reduction follows from the posterior update, the accompanying error reduction suggests that rectification provides useful predictive evidence rather than merely increasing confidence. One possible explanation is that reliable acoustic and visual cues help resolve ambiguity in the text representation, while the original evidence branch retains its task-relevant information. However, lower uncertainty does not improve every prediction, suggesting that the benefit also depends on how informative the auxiliary evidence is for each sample. Overall, these results support selectively refining uncertain yet informative representations while preserving their predictive value.

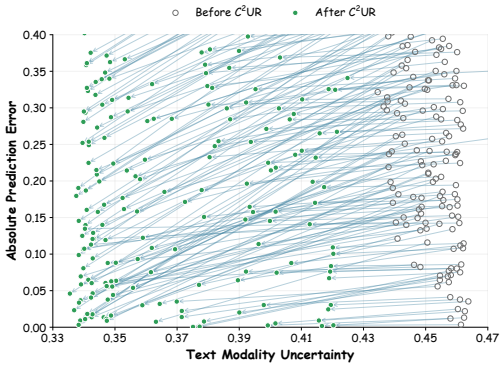


Figure 5: Text modality uncertainty rectification on CMU-MOSI. Gray circles and green dots indicate states before and after rectification; arrows connect the same sample.

5 Conclusion

In this paper, we propose C^{2}UR, a multimodal affective computing framework that decouples modality task utility from representation uncertainty. By adaptively rectifying uncertain yet informative representations with reliable cross-modal consensus and combining original and rectified evidence through dual-branch fusion, C^{2}UR reduces uncertainty while preserving valuable task information. Experiments across four MAC tasks demonstrate state-of-the-art performance, with further evaluations showing strong robustness under combined noise and modality missingness.

Limitations. The effectiveness of cross-modal rectification depends on the availability of informative auxiliary evidence. When multiple modalities are severely degraded or provide consistent but misleading evidence, rectification may offer limited benefits or introduce bias. Moreover, removal-based utility supervision incurs additional training overhead. Future work will explore more efficient utility estimation and more robust consensus modeling under severe multimodal degradation.

REFERENCES

Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *NAACL*, pp. 370–379, 2019.

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 59–66, 2018.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *ICML*, pp. 5608–5628, 2024.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *ACL*, pp. 4619–4629, 2019.

Jili Chen, Yihua Zhong, Qionghao Huang, Changqin Huang, Fan Jiang, Xiaodi Huang, and Xun Wang. Ucmib-pns: balancing sufficiency and necessity with probabilistic causality and cross-modal uncertainty in multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 17(1): 92–106, 2026.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019.

Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *ACM MM*, pp. 9650–9659, 2024a.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *CVPR*, pp. 26876–26885, 2024b.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *ACL*, pp. 1726–1736, 2024.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-Philippe Morency, and Soujanya Poria. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *ICMI*, pp. 6–15, 2021a.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *EMNLP*, pp. 9180–9192, 2021b.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *EMNLP-IJCNLP*, pp. 2046–2056, 2019.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

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

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Peng Hu, Yang Qin, Yuanbiao Gou, Yunfan Li, Mouxing Yang, and Xi Peng. Probabilistic multimodal learning with von mises-fisher distributions. In *IJCAI*, pp. 5390–5398, 2025.

Changqin Huang, Jili Chen, Qionghao Huang, Shijin Wang, Yaxin Tu, and Xiaodi Huang. Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114:102725, 2025.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, volume 30, 2017.

Kyeonghun Kim and Sanghyun Park. Aobert: All-modalities-in-one bert for multimodal sentiment analysis. *Information Fusion*, 92:37–45, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, volume 30, 2017.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pp. 22214–22224, 2023a.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, Mingyang Sun, Ziyun Qian, Dongliang Kou, and Lihua Zhang. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *NeurIPS*, volume 37, pp. 28515–28536, 2024.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *CVPR*, pp. 6631–6640, 2023b.

Yong Li, Yuanzhi Wang, Yi Ding, Shiqing Zhang, Ke Lu, and Cuntai Guan. Decoupled hierarchical distillation for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6912–6928, 2026.

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.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

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 *ACL*, pp. 2247–2256, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Miaosen Luo, Yuncheng Jiang, and Sijie Mai. Towards explainable fusion and balanced learning in multimodal sentiment analysis. In *ACM MM*, pp. 1997–2006, 2025.

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*, pp. 2554–2562, 2021.

Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25: 4121–4134, 2023.

Sijie Mai, Ying Zeng, and Haifeng Hu. Learning by comparing: Boosting multimodal affective computing through ordinal learning. In *WWW*, pp. 2120–2134, 2025.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *WACV*, pp. 3930–3940, 2022.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *ICLR*, 2026.

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 *ACL*, pp. 2359–2369, 2020.

Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *ICCV*, pp. 6902–6911, 2019.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *ACL*, pp. 658–670, 2023.

Xiaolin Tang, Kai Yang, Hong Wang, Jiahang Wu, Yechen Qin, Wenhao Yu, and Dongpu Cao. Prediction-uncertainty-aware decision-making for autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*, 7(4):849–862, 2022.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

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 *ACL*, pp. 6558–6569, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017.

Yidan Wang, Zongheng Wang, Hongjie Xing, Chunguo Li, and Xiaoxiao Liu. Multi-metric representation learning strategy based on clustering for fine-grained multimodal sentiment analysis. In *CVPR*, pp. 37864–37873, 2026.

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *ICCV*, pp. 22025–22034, 2023a.

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *NeurIPS*, volume 36, pp. 17117–17128, 2023b.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *ICLR*, 2024.

Heng Xie, Kang Zhu, Zhengqi Wen, Jianhua Tao, Xuefei Liu, Ruibo Fu, and Changsheng Li. Psa-mf: Personality-sentiment aligned multi-level fusion for multimodal sentiment analysis. In *AAAI*, volume 40, pp. 2191–2199, 2026.

Zhi Xu, Dingkang Yang, Mingcheng Li, Yuzheng Wang, Zhaoyu Chen, Jiawei Chen, Jinjie Wei, and Lihua Zhang. Debiased multimodal understanding for human language sequences. In *AAAI*, volume 39, pp. 14450–14458, 2025.

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

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*, volume 35, pp. 10790–10797, 2021.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pp. 1103–1114, 2017.

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 *ACL*, pp. 2236–2246, 2018.

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 *EMNLP*, pp. 756–767, 2023a.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *NeurIPS*, volume 37, pp. 55943–55974, 2024a.

Jiaxin Zhang. Modern monte carlo methods for efficient uncertainty quantification and propagation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5):e1539, 2021.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *ICML*, pp. 41753–41769, 2023b.

Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. Learning multitask commonness and uniqueness for multimodal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence*, 5:1349–1361, 2024b.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *ACL-IJCNLP*, pp. 2608–2618, 2021.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Fuji Ren. Kebr: Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis. In *ACM MM*, pp. 5732–5741, 2024.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *ACL*, pp. 22123–22138, 2025.

Yitong Zhu, Yuxuan Jiang, Guanxuan Jiang, Bojing Hou, Peng Yuan Zhou, Ge Lin, and Yuyang Wang. Qa-moe: Towards a continuous reliability spectrum with quality-aware mixture of experts for robust multimodal sentiment analysis. In *ACL*, pp. 31689–31703, 2026.

Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. Glomo: Global-local modal fusion for multimodal sentiment analysis. In *ACM MM*, pp. 1800–1809, 2024.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. Hyper-modality enhancement for multimodal sentiment analysis with missing modalities. In *NeurIPS*, volume 38, pp. 142100–142129, 2025a.

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 *ICCV*, pp. 4626–4636, 2025b.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. Tmdc: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In *AAAI*, volume 40, pp. 2281–2289, 2026.

Appendix Contents

The Use of Large Language Models (LLMs) **15**

A More Technical Details **15**

A.1 Architecture of Modality-specific Encoders 15

A.2 Probabilistic Representation Learning 15

A.3 Closed-form Derivation of Consistency-guided Uncertainty Rectification 17

B Experimental Details **20**

B.1 Datasets Information 20

B.2 Baselines 20

B.3 Feature Extraction 22

B.4 Implementation Details 23

B.5 Evaluation Metrics 24

C More Experimental Results **25**

C.1 Results on Multimodal Humor Detection 25

C.2 Results on Multimodal Sarcasm Detection 26

C.3 Results on Multimodal Emotion Recognition 26

C.4 Results on Continuous Reliability Spectrum 27

C.5 Detailed Ablation Study 27

C.6 Parameter Sensitivity Analysis 28

C.7 Testing Stability Analysis 29

C.8 Model Complexity Analysis 30

THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, large language models (LLMs) were used solely for language polishing to improve clarity and fluency. All scientific content, including the experimental design, results, and conclusions, is the responsibility of the authors. No LLM was involved in the ideation, analysis, or interpretation of the research, and all LLM-assisted content was carefully reviewed and verified by the authors.

A MORE TECHNICAL DETAILS

A.1 ARCHITECTURE OF MODALITY-SPECIFIC ENCODERS

Following prior work (Tsai et al., 2019), we first employ modality-specific one-dimensional convolutional projection layers to map heterogeneous input features from different modalities into a unified representation space and align their feature dimensions:

$$\hat{\boldsymbol{U}}_m = W_{1D}^m(\boldsymbol{U}_m),$$

where W_{1D}^m denotes the one-dimensional convolutional projection layer for modality m, $\hat{\boldsymbol{U}}_m \in \mathbb{R}^{l_m \times d}$ denotes the projected feature sequence, and d denotes the embedding dimension shared across all modalities.

We then feed the projected feature sequence into a modality-specific Transformer (Vaswani et al., 2017) encoder to capture intra-modal temporal dependencies. Specifically, we introduce positional encoding to preserve positional information within the sequence, and formulate the sequence representation of modality m as:

$$\boldsymbol{S}_m = \mathcal{F}_{\phi_m}(\hat{\boldsymbol{U}}_m + PE(l_m, d)),$$

where $\mathcal{F}_{\phi_m}$ denotes the modality-specific Transformer encoder parameterized by ϕ_m, and $PE(l_m, d)$ denotes the positional encoding with sequence length l_m and dimension d.

Finally, we apply global max pooling to aggregate the sequence-level representation into a fixed-dimensional modality representation:

$$\boldsymbol{X}_m = GMP(\boldsymbol{S}_m),$$

where $\boldsymbol{X}_m \in \mathbb{R}^d$ denotes the initial representation of modality m, and $GMP(\cdot)$ denotes the global max pooling operation.

A.2 PROBABILISTIC REPRESENTATION LEARNING

This subsection provides additional details on the Gaussian parameterization, reparameterized sampling procedure, uncertainty estimation, and training objective of the probabilistic representation learning module introduced in the main text.

Gaussian Parameterization. For each modality $m \in \{t, a, v\}$, given the unimodal representation $\boldsymbol{X}_m \in \mathbb{R}^d$ produced by its modality-specific encoder, we employ mean and variance parameterization networks shared across the three modalities to map $\boldsymbol{X}_m$ to a probabilistic estimate of a common task-relevant latent variable:

$$\boldsymbol{\mu}_m = f_\mu(\boldsymbol{X}_m), \qquad \boldsymbol{\sigma}_m^2 = \mathrm{Softplus}(f_\sigma(\boldsymbol{X}_m)) + \epsilon, \tag{16}$$

where $f_\mu(\cdot)$ and $f_\sigma(\cdot)$ denote the mean and variance parameterization networks shared across the three modalities, respectively, and $\epsilon > 0$ is a small constant introduced for numerical stability. The Softplus function ensures that the variance of each latent dimension remains positive, i.e., $\boldsymbol{\sigma}_m^2 \in \mathbb{R}_+^d$.

To reduce the parameter and computational overhead of probabilistic modeling, we adopt a diagonal covariance assumption and model the modality-specific latent variable $\boldsymbol{Z}_m \in \mathbb{R}^d$ as

$$q(\boldsymbol{Z}_m \mid \boldsymbol{X}_m) = \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \qquad \boldsymbol{\Sigma}_m = \mathrm{diag}\big(\boldsymbol{\sigma}_m^2\big). \tag{17}$$

Here, $\boldsymbol{\mu}_m$ represents the modality-specific central estimate of the common task-relevant information, while $\boldsymbol{\Sigma}_m$ characterizes the dispersion of the corresponding latent distribution across different dimensions. Unlike a deterministic point representation, this probabilistic formulation represents

each modality as a distribution, allowing its representation uncertainty to be explicitly quantified through the posterior variance.

To enable differentiable optimization of the stochastic latent variable, we sample from the posterior distribution using the reparameterization trick (Kingma & Welling, 2013):

$$\boldsymbol{Z}_m = \boldsymbol{\mu}_m + \boldsymbol{\sigma}_m \odot \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{18}$$

where $\odot$ denotes element-wise multiplication, $\boldsymbol{\xi}$ is sampled from a standard normal distribution, and $\boldsymbol{I}$ denotes the identity matrix. This reparameterization isolates the sampling randomness in $\boldsymbol{\xi}$, which is independent of the model parameters, thereby allowing gradients to propagate through the mean and variance parameterization networks.

Modality Uncertainty Estimation. Based on the probabilistic representation defined above, we quantify the uncertainty of each modality using the overall dispersion of its posterior distribution. Specifically, we first compute the average posterior variance across all latent dimensions:

$$\bar{\sigma}_m^2 = \frac{1}{d} \sum_{r=1}^{d} \sigma_{m,r}^2 = \frac{1}{d} \operatorname{Tr}(\boldsymbol{\Sigma}_m), \tag{19}$$

where $\sigma_{m,r}^2$ denotes the posterior variance of modality m in the r-th latent dimension. Since $\boldsymbol{\Sigma}_m$ is diagonal, its trace equals the sum of the variances across all latent dimensions.

We then map the average posterior variance to a bounded modality uncertainty score using a monotonic exponential transformation:

$$u_m = 1 - \exp\left(-\bar{\sigma}_m^2\right) = 1 - \exp\left(-\frac{1}{d} \operatorname{Tr}(\boldsymbol{\Sigma}_m)\right). \tag{20}$$

Since $\bar{\sigma}_m^2 \geq 0$, we have $u_m \in [0, 1)$. Moreover, since $1 - \exp(-x)$ is strictly monotonically increasing for $x \geq 0$, this transformation preserves the ordering induced by posterior dispersion while mapping the unbounded average variance to a common bounded interval. Such a bounded formulation facilitates subsequent joint modeling with other modality-specific attributes. When the latent distribution is more concentrated, $\bar{\sigma}_m^2$ is small and u_m approaches zero. Conversely, u_m increases as the latent distribution becomes more dispersed. Therefore, u_m provides a compact measure of the uncertainty associated with the current unimodal probabilistic representation.

Probabilistic Representation Learning Objective. Probabilistic parameterization alone does not guarantee that the learned latent representations retain sufficient task-discriminative information for downstream sentiment prediction. We therefore impose task supervision on the probabilistic representations of all modalities using a unimodal prediction head $g_u(\cdot)$ shared across modalities. For modality m, after obtaining the stochastic latent representation $\boldsymbol{Z}_m$ through the reparameterization procedure above, we define the probabilistic representation learning objective as

$$\mathcal{L}_{PRL}^m = \mathbb{E}_{q(\boldsymbol{Z}_m|\boldsymbol{X}_m)} \left[\mathcal{L}_{TASK}\left(g_u(\boldsymbol{Z}_m), y\right)\right] + \beta D_{KL}\left(q(\boldsymbol{Z}_m \mid \boldsymbol{X}_m) \,\|\, \mathcal{N}(\mathbf{0}, \boldsymbol{I})\right), \tag{21}$$

where $\mathcal{L}_{TASK}(\cdot)$ denotes the task loss associated with the target sentiment prediction task, $g_u(\cdot)$ is the unimodal task prediction head shared across the three modalities, and β is a hyperparameter controlling the strength of the KL regularization. The shared prediction head encourages latent representations from different modalities to be organized within a common task-semantic space, thereby making their probabilistic evidence directly comparable and suitable for subsequent cross-modal aggregation through a Product-of-Experts (Hinton, 2002).

During training, the expectation in the task loss is approximated using Monte Carlo samples obtained through reparameterization. Under the diagonal Gaussian posterior adopted in our model, the KL divergence from the standard normal prior admits the following closed-form expression:

$$D_{KL}\left(q(\boldsymbol{Z}_m \mid \boldsymbol{X}_m) \,\|\, \mathcal{N}(\mathbf{0}, \boldsymbol{I})\right) = \frac{1}{2} \sum_{r=1}^{d} \left[\mu_{m,r}^2 + \sigma_{m,r}^2 - \log\left(\sigma_{m,r}^2\right) - 1\right], \tag{22}$$

where $\mu_{m,r}$ and $\sigma_{m,r}^2$ denote the mean and variance of modality m in the r-th latent dimension, respectively. This closed-form expression avoids the need for additional numerical estimation of

the KL divergence during training. The probabilistic representation learning objectives for the three modalities are then aggregated as

$$\mathcal{L}_{PRL} = \frac{1}{3} \sum_{m \in \{t,a,v\}} \mathcal{L}_{PRL}^{m}. \tag{23}$$

In this objective, the task loss encourages stochastic representations sampled from each modality-specific distribution to preserve the discriminative information required for sentiment prediction, while the KL regularization term constrains the posterior distribution from deviating excessively from the standard normal prior and regularizes the probabilistic latent space. Through the shared Gaussian parameterization networks and shared task prediction head, different modalities are consistently modeled as probabilistic estimates within a common task-relevant latent space. Their posterior variances can then be used to quantify modality uncertainty, providing a probabilistic foundation for the subsequent reliability-weighted cross-modal Product-of-Experts and adaptive posterior correction.

A.3 Closed-form Derivation of Consistency-guided Uncertainty Rectification

This subsection derives the closed-form Gaussian solutions for the reliability-weighted cross-modal consensus and adaptive posterior rectification introduced in Section 3.3. We further show that rectification does not increase the target modality's posterior covariance or the associated variance-based uncertainty score.

For each modality $m \in \{t, a, v\}$, we denote its probabilistic representation as

$$q_m(\boldsymbol{z}) \equiv q(\boldsymbol{z} \mid \boldsymbol{X}_m) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \qquad \boldsymbol{\Lambda}_m = \boldsymbol{\Sigma}_m^{-1}, \tag{24}$$

where $\boldsymbol{\Lambda}_m$ denotes the precision matrix. Since each variance is parameterized by Softplus with a positive stability constant, $\boldsymbol{\Sigma}_m$ is positive definite and thus invertible. For target modality m, let $\mathcal{A}_m = \{t, a, v\} \setminus \{m\}$ denote the set of auxiliary modalities. The reliability weights and rectification coefficient are

$$\rho_r = 1 - u_r = \exp\left(-\frac{1}{d} \operatorname{Tr}(\boldsymbol{\Sigma}_r)\right), \qquad r \in \mathcal{A}_m, \tag{25}$$

$$\gamma_m = \widehat{I}_m \cdot u_m \cdot C_m \geq 0. \tag{26}$$

For a given input, these quantities are scalars independent of the integration variable $\boldsymbol{z}$ and can therefore be treated as fixed coefficients in the following derivations.

Closed-form Cross-modal Consensus. The reliability-weighted consensus distribution is defined as

$$q_{-m}(\boldsymbol{z}) = \frac{1}{\mathcal{Z}_{-m}} \prod_{r \in \mathcal{A}_m} q_r(\boldsymbol{z})^{\rho_r}. \tag{27}$$

The log density of each Gaussian expert is

$$\begin{aligned} \log q_r(\boldsymbol{z}) &= -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}_r| - \frac{1}{2} (\boldsymbol{z} - \boldsymbol{\mu}_r)^\top \boldsymbol{\Lambda}_r (\boldsymbol{z} - \boldsymbol{\mu}_r) \\ &= -\frac{1}{2} \boldsymbol{z}^\top \boldsymbol{\Lambda}_r \boldsymbol{z} + \boldsymbol{z}^\top \boldsymbol{\Lambda}_r \boldsymbol{\mu}_r + c_r, \end{aligned} \tag{28}$$

where c_r collects terms independent of $\boldsymbol{z}$. Substituting this expression into the weighted product yields

$$\begin{aligned} \log q_{-m}(\boldsymbol{z}) &= -\log \mathcal{Z}_{-m} + \sum_{r \in \mathcal{A}_m} \rho_r \log q_r(\boldsymbol{z}) \\ &= -\frac{1}{2} \boldsymbol{z}^\top \left(\sum_{r \in \mathcal{A}_m} \rho_r \boldsymbol{\Lambda}_r \right) \boldsymbol{z} + \boldsymbol{z}^\top \left(\sum_{r \in \mathcal{A}_m} \rho_r \boldsymbol{\Lambda}_r \boldsymbol{\mu}_r \right) + c_{-m}. \end{aligned} \tag{29}$$

Define the consensus precision and precision-weighted mean as

$$\boldsymbol{\Lambda}_m^{\mathrm{c}} = \sum_{r \in \mathcal{A}_m} \rho_r \boldsymbol{\Lambda}_r, \qquad \boldsymbol{h}_m^{\mathrm{c}} = \sum_{r \in \mathcal{A}_m} \rho_r \boldsymbol{\Lambda}_r \boldsymbol{\mu}_r. \tag{30}$$

Since $\rho_r > 0$ and $\mathbf{\Lambda}_r \succ \mathbf{0}$, we have $\mathbf{\Lambda}_m^{\mathrm{c}} \succ \mathbf{0}$. Completing the square gives

$$\begin{aligned}\log q_{-m}(\boldsymbol{z}) &= -\frac{1}{2}\boldsymbol{z}^\top \mathbf{\Lambda}_m^{\mathrm{c}} \boldsymbol{z} + \boldsymbol{z}^\top \boldsymbol{h}_m^{\mathrm{c}} + c_{-m} \\ &= -\frac{1}{2}\left(\boldsymbol{z} - (\mathbf{\Lambda}_m^{\mathrm{c}})^{-1}\boldsymbol{h}_m^{\mathrm{c}}\right)^\top \mathbf{\Lambda}_m^{\mathrm{c}} \left(\boldsymbol{z} - (\mathbf{\Lambda}_m^{\mathrm{c}})^{-1}\boldsymbol{h}_m^{\mathrm{c}}\right) + \widetilde{c}_{-m}.\end{aligned} \tag{31}$$

After normalization, the consensus distribution is Gaussian:

$$q_{-m}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_m^{\mathrm{c}}, \mathbf{\Sigma}_m^{\mathrm{c}}), \tag{32}$$

with

$$\mathbf{\Sigma}_m^{\mathrm{c}} = (\mathbf{\Lambda}_m^{\mathrm{c}})^{-1} = \left(\sum_{r\in\mathcal{A}_m} \rho_r \mathbf{\Sigma}_r^{-1}\right)^{-1}, \tag{33}$$

$$\boldsymbol{\mu}_m^{\mathrm{c}} = \mathbf{\Sigma}_m^{\mathrm{c}} \boldsymbol{h}_m^{\mathrm{c}} = \mathbf{\Sigma}_m^{\mathrm{c}} \sum_{r\in\mathcal{A}_m} \rho_r \mathbf{\Sigma}_r^{-1} \boldsymbol{\mu}_r. \tag{34}$$

Here, the superscript c denotes consensus parameters. In the three-modality setting, $\mathcal{A}_m = \{j, k\}$, giving

$$\mathbf{\Lambda}_m^{\mathrm{c}} = \rho_j \mathbf{\Sigma}_j^{-1} + \rho_k \mathbf{\Sigma}_k^{-1}, \tag{35}$$

$$\boldsymbol{\mu}_m^{\mathrm{c}} = \mathbf{\Sigma}_m^{\mathrm{c}} \left(\rho_j \mathbf{\Sigma}_j^{-1} \boldsymbol{\mu}_j + \rho_k \mathbf{\Sigma}_k^{-1} \boldsymbol{\mu}_k\right). \tag{36}$$

Closed-form Posterior Rectification. We incorporate the consensus into the target posterior as

$$q_m^*(\boldsymbol{z}) = \frac{1}{\mathcal{Z}_m^*} q_m(\boldsymbol{z}) q_{-m}(\boldsymbol{z})^{\gamma_m}. \tag{37}$$

Expanding the two Gaussian factors yields

$$\begin{aligned}\log q_m^*(\boldsymbol{z}) &= -\log \mathcal{Z}_m^* + \log q_m(\boldsymbol{z}) + \gamma_m \log q_{-m}(\boldsymbol{z}) \\ &= -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_m)^\top \mathbf{\Lambda}_m (\boldsymbol{z} - \boldsymbol{\mu}_m) - \frac{\gamma_m}{2}(\boldsymbol{z} - \boldsymbol{\mu}_m^{\mathrm{c}})^\top \mathbf{\Lambda}_m^{\mathrm{c}} (\boldsymbol{z} - \boldsymbol{\mu}_m^{\mathrm{c}}) + c_m^* \\ &= -\frac{1}{2}\boldsymbol{z}^\top \left(\mathbf{\Lambda}_m + \gamma_m \mathbf{\Lambda}_m^{\mathrm{c}}\right) \boldsymbol{z} + \boldsymbol{z}^\top \left(\mathbf{\Lambda}_m \boldsymbol{\mu}_m + \gamma_m \mathbf{\Lambda}_m^{\mathrm{c}} \boldsymbol{\mu}_m^{\mathrm{c}}\right) + \widetilde{c}_m^*.\end{aligned} \tag{38}$$

Completing the square yields

$$q_m^*(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_m^*, \mathbf{\Sigma}_m^*), \tag{39}$$

where

$$\mathbf{\Sigma}_m^* = (\mathbf{\Lambda}_m + \gamma_m \mathbf{\Lambda}_m^{\mathrm{c}})^{-1} = \left(\mathbf{\Sigma}_m^{-1} + \gamma_m \sum_{r\in\mathcal{A}_m} \rho_r \mathbf{\Sigma}_r^{-1}\right)^{-1}, \tag{40}$$

$$\boldsymbol{\mu}_m^* = \mathbf{\Sigma}_m^* \left(\mathbf{\Lambda}_m \boldsymbol{\mu}_m + \gamma_m \mathbf{\Lambda}_m^{\mathrm{c}} \boldsymbol{\mu}_m^{\mathrm{c}}\right) = \mathbf{\Sigma}_m^* \left(\mathbf{\Sigma}_m^{-1} \boldsymbol{\mu}_m + \gamma_m \sum_{r\in\mathcal{A}_m} \rho_r \mathbf{\Sigma}_r^{-1} \boldsymbol{\mu}_r\right). \tag{41}$$

Equivalently, substituting the consensus definition gives

$$q_m^*(\boldsymbol{z}) \propto q_m(\boldsymbol{z}) \left[\frac{1}{\mathcal{Z}_{-m}} \prod_{r\in\mathcal{A}_m} q_r(\boldsymbol{z})^{\rho_r}\right]^{\gamma_m}, \tag{42}$$

$$\propto q_m(\boldsymbol{z}) \prod_{r\in\mathcal{A}_m} q_r(\boldsymbol{z})^{\gamma_m \rho_r}. \tag{43}$$

Thus, the target expert retains unit weight, while each auxiliary expert contributes with effective weight $\gamma_m \rho_r$. Here, ρ_r weights each auxiliary expert according to its estimated reliability, whereas γ_m controls the overall influence of cross-modal evidence on the target posterior.

Element-wise Form under Diagonal Covariances. Since all modality covariances are diagonal, both the consensus and rectified covariances remain diagonal. Let $s \in \{1, \dots, d\}$ index the latent dimensions, and define the consensus precision in dimension s as

$$a_{m,s} = \sum_{r \in \mathcal{A}_m} \frac{\rho_r}{\sigma_{r,s}^2} > 0. \tag{44}$$

The consensus parameters can then be computed element-wise:

$$(\sigma_{m,s}^{\mathrm{c}})^2 = a_{m,s}^{-1} = \left(\sum_{r \in \mathcal{A}_m} \frac{\rho_r}{\sigma_{r,s}^2} \right)^{-1}, \tag{45}$$

$$\mu_{m,s}^{\mathrm{c}} = \frac{\sum_{r \in \mathcal{A}_m} \rho_r \mu_{r,s} / \sigma_{r,s}^2}{\sum_{r \in \mathcal{A}_m} \rho_r / \sigma_{r,s}^2}. \tag{46}$$

Similarly, the rectified parameters are

$$(\sigma_{m,s}^*)^2 = \left(\frac{1}{\sigma_{m,s}^2} + \gamma_m a_{m,s} \right)^{-1} = \frac{\sigma_{m,s}^2}{1 + \gamma_m a_{m,s} \sigma_{m,s}^2}, \tag{47}$$

$$\mu_{m,s}^* = \frac{\mu_{m,s}/\sigma_{m,s}^2 + \gamma_m \sum_{r \in \mathcal{A}_m} \rho_r \mu_{r,s}/\sigma_{r,s}^2}{1/\sigma_{m,s}^2 + \gamma_m a_{m,s}} = \frac{\mu_{m,s} + \gamma_m a_{m,s} \sigma_{m,s}^2 \mu_{m,s}^{\mathrm{c}}}{1 + \gamma_m a_{m,s} \sigma_{m,s}^2}. \tag{48}$$

Defining

$$\alpha_{m,s} = \frac{\gamma_m a_{m,s} \sigma_{m,s}^2}{1 + \gamma_m a_{m,s} \sigma_{m,s}^2} \in [0, 1), \tag{49}$$

the mean update can be expressed as

$$\mu_{m,s}^* = (1 - \alpha_{m,s}) \mu_{m,s} + \alpha_{m,s} \mu_{m,s}^{\mathrm{c}}. \tag{50}$$

Thus, each rectified mean coordinate interpolates between the target estimate and the auxiliary consensus, with the interpolation strength determined by the rectification coefficient, target variance, and auxiliary consensus precision. These element-wise expressions also avoid explicit matrix inversion.

Proof of Non-increasing Posterior Uncertainty. We first establish the covariance ordering. Since $\gamma_m \geq 0$ and $\mathbf{\Lambda}_m^{\mathrm{c}} \succ \mathbf{0}$,

$$(\mathbf{\Sigma}_m^*)^{-1} - \mathbf{\Sigma}_m^{-1} = \gamma_m \mathbf{\Lambda}_m^{\mathrm{c}} = \gamma_m \sum_{r \in \mathcal{A}_m} \rho_r \mathbf{\Sigma}_r^{-1} \succeq \mathbf{0}. \tag{51}$$

For completeness, define

$$\boldsymbol{B}_m = \gamma_m \mathbf{\Sigma}_m^{1/2} \mathbf{\Lambda}_m^{\mathrm{c}} \mathbf{\Sigma}_m^{1/2} \succeq \mathbf{0}. \tag{52}$$

The rectified covariance can be factorized as

$$\begin{aligned} \mathbf{\Sigma}_m^* &= \left(\mathbf{\Sigma}_m^{-1} + \gamma_m \mathbf{\Lambda}_m^{\mathrm{c}} \right)^{-1} \\ &= \left[\mathbf{\Sigma}_m^{-1/2} (\boldsymbol{I} + \boldsymbol{B}_m) \mathbf{\Sigma}_m^{-1/2} \right]^{-1} \\ &= \mathbf{\Sigma}_m^{1/2} (\boldsymbol{I} + \boldsymbol{B}_m)^{-1} \mathbf{\Sigma}_m^{1/2}. \end{aligned} \tag{53}$$

Since all eigenvalues of $\boldsymbol{B}_m$ are nonnegative, $(\boldsymbol{I} + \boldsymbol{B}_m)^{-1} \preceq \boldsymbol{I}$. Therefore,

$$\mathbf{\Sigma}_m - \mathbf{\Sigma}_m^* = \mathbf{\Sigma}_m^{1/2} \left[\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{B}_m)^{-1} \right] \mathbf{\Sigma}_m^{1/2} \succeq \mathbf{0}, \tag{54}$$

which proves $\mathbf{\Sigma}_m^* \preceq \mathbf{\Sigma}_m$.

Under the diagonal covariance assumption, this result also follows directly for each dimension:

$$\sigma_{m,s}^2 - (\sigma_{m,s}^*)^2 = \sigma_{m,s}^2 - \frac{\sigma_{m,s}^2}{1+\gamma_m a_{m,s}\sigma_{m,s}^2} = \frac{\gamma_m a_{m,s}(\sigma_{m,s}^2)^2}{1+\gamma_m a_{m,s}\sigma_{m,s}^2} \geq 0. \tag{55}$$

Summing over the latent dimensions gives

$$\mathrm{Tr}(\mathbf{\Sigma}_m^*) = \sum_{s=1}^{d}(\sigma_{m,s}^*)^2 \leq \sum_{s=1}^{d}\sigma_{m,s}^2 = \mathrm{Tr}(\mathbf{\Sigma}_m). \tag{56}$$

Applying the same uncertainty mapping as for the original posterior yields

$$u_m^* = 1 - \exp\left(-\frac{1}{d}\,\mathrm{Tr}(\mathbf{\Sigma}_m^*)\right) \leq 1 - \exp\left(-\frac{1}{d}\,\mathrm{Tr}(\mathbf{\Sigma}_m)\right) = u_m. \tag{57}$$

When $\gamma_m = 0$, rectification reduces to the original posterior, i.e., $q_m^* = q_m$ and $u_m^* = u_m$. When $\gamma_m > 0$, the positive definite auxiliary precision further gives $\mathbf{\Sigma}_m^* \prec \mathbf{\Sigma}_m$ and $u_m^* < u_m$.

B Experimental Details

B.1 Datasets Information

We evaluate C^{2}UR on five benchmarks spanning four tasks: Multimodal Sentiment Analysis (MSA), Multimodal Humor Detection (MHD), Multimodal Sarcasm Detection (MSD), and Multimodal Emotion Recognition (MER). Details of each dataset are provided below.

CMU-MOSI (Zadeh et al., 2016) contains over 2,000 online video segments and is widely used for MSA. Each segment carries a sentiment intensity annotation on a seven-point scale from -3 (strongly negative) to 3 (strongly positive).

CMU-MOSEI (Zadeh et al., 2018) provides a larger and more diverse benchmark for MSA, with over 22,000 video segments covering approximately 250 topics and more than 1,000 YouTube speakers. Annotations include six emotion categories and sentiment intensity scores ranging from -3 to 3. We use the sentiment annotations in our experiments.

UR-FUNNY (Hasan et al., 2019) supports MHD using TED talk recordings from 1,741 speakers. Each instance consists of a target utterance, termed a punchline, and its preceding context, with textual, acoustic, and visual information. Humor labels are derived from audience-laughter markers in the transcripts, while negative examples contain no such markers. Following previous studies (Huang et al., 2025; Xu et al., 2025), we use version 2, which comprises 7,614 training, 980 validation, and 994 test samples.

MUStARD (Castro et al., 2019) is an MSD benchmark containing 690 video clips from *Friends*, *The Big Bang Theory*, *The Golden Girls*, and *Sarcasmaholics*. Each clip is assigned a binary sarcasm label and includes the target utterance together with the preceding dialogue as context.

CHERMA (Sun et al., 2023) contains 28,717 Chinese utterances collected from diverse media sources for MER. Its seven emotion categories comprise Ekman's six basic emotions and a neutral category. Each utterance has three unimodal labels and one multimodal label. The dataset is partitioned into training, validation, and test sets in a ratio of $6:2:2$.

B.2 Baselines

We summarize the baselines used across the four MAC tasks below.

MMRest (Wang et al., 2026) learns global and cluster-specific distance metrics to distinguish fine-grained sentiments and reduce overlap between sentiment classes. Its final prediction combines metric-based projections with a score derived from unimodal and multimodal representations.

PSA-MF (Xie et al., 2026) incorporates personality information into sentiment representation learning. A hierarchical fusion mechanism then integrates multimodal features at multiple levels to support personalized sentiment prediction.

DecAlign (Qian et al., 2026) separates modality-specific and shared representations to address cross-modal heterogeneity. It combines prototype-guided optimal transport with distribution matching based on maximum mean discrepancy, followed by Transformer-based fusion.

GLoMo (Zhuang et al., 2024) combines local and global information for multimodal prediction. Modality-specific mixture-of-experts layers select and aggregate local representations, while a global-guided fusion module integrates them with global features.

ITHP (Xiao et al., 2024) adopts an information-theoretic hierarchy in which auxiliary modalities act as detectors of informative content in a primary modality. Information bottleneck constraints guide the extraction of task-relevant representations.

KAN-MCP (Luo et al., 2025) integrates Kolmogorov–Arnold networks for interpretable cross-modal modeling with information-bottleneck-based feature compression and denoising. Its learning strategy also addresses imbalance among modality contributions.

TMDC (Zhuang et al., 2026) addresses noisy and missing modalities through two sequential training stages. It first extracts denoised modality-specific and shared representations from complete inputs, then uses these representations to compensate for missing information.

EUAR (Gao et al., 2024a) introduces uncertainty-aware experts and routing to mitigate noisy multimodal inputs. An uncertainty-based training objective encourages the router to assign samples to experts with lower estimated uncertainty, improving feature extraction under noise.

PML (Hu et al., 2025) represents modality features with von Mises–Fisher distributions to model uncertainty in directional representations. It combines prototypical contrastive learning with reliability-aware fusion to improve class discrimination and manage conflicting modality evidence.

HME (Zhuang et al., 2025a) enriches available modalities with relevant semantic information from other samples without explicitly reconstructing missing inputs. Uncertainty-aware fusion adaptively combines the original and enhanced representations.

QA-MoE (Zhu et al., 2026) considers noise and modality missingness within a continuous reliability spectrum. It estimates modality reliability through self-supervised aleatoric uncertainty and uses these estimates to guide expert routing under varying input quality.

C-MFN (Hasan et al., 2019) extends the Memory Fusion Network by incorporating preceding context into punchline processing. It models modality-specific temporal patterns and cross-modal interactions for humor prediction; its punchline-only variant omits contextual information.

MulT (Tsai et al., 2019) uses directional pairwise cross-modal attention to capture interactions between asynchronous modality sequences. This mechanism models long-range cross-modal dependencies without requiring explicit temporal alignment.

MISA (Hazarika et al., 2020) maps each modality into shared and private representation spaces. The shared space reduces modality discrepancies, while the private spaces retain modality-specific characteristics for subsequent fusion.

Self-MM (Yu et al., 2021) derives unimodal pseudo-labels from multimodal sentiment supervision. Joint training with these labels encourages each modality to learn more discriminative representations.

MMIM (Han et al., 2021b) preserves task-relevant information by hierarchically maximizing mutual information between modality pairs and between unimodal inputs and the fused representation. These objectives are optimized jointly with the prediction task.

FDMER (Yang et al., 2022) disentangles shared and modality-specific features using common and private encoders. Adversarial learning encourages cross-modal consistency while retaining complementary information, which is integrated through cross-modal attention.

DMD (Li et al., 2023b) decomposes modality representations into modality-relevant and modality-exclusive components. A dynamic graph supports adaptive knowledge distillation across modalities.

AGM (Li et al., 2023a) alleviates modality competition by adaptively modulating training gradients. It estimates modality contributions through a mono-modal perspective and balances their optimization across different fusion architectures.

AtCAF (Huang et al., 2025) combines text debiasing with counterfactual cross-modal attention to learn causality-aware representations. These mechanisms reduce spurious influences during multimodal fusion.

SuCI (Xu et al., 2025) treats subject identity as a confounding factor and applies causal intervention to reduce subject-related spurious correlations. This encourages representations that generalize across subjects.

DHMD (Li et al., 2026) uses self-regression to separate modality-invariant and modality-specific information. It then combines graph-based coarse-grained distillation with dictionary-based fine-grained distillation to improve cross-modal knowledge transfer.

CIM-MTL (Akhtar et al., 2019) jointly learns sentiment and emotion prediction through a multitask framework. Context-level inter-modal attention captures complementary information and accounts for unequal modality contributions.

MAG-XLNet (Rahman et al., 2020) incorporates acoustic and visual information into a pretrained XLNet model through a multimodal adaptation gate. The gate conditions textual representations on nonverbal signals during fine-tuning.

BBFN (Han et al., 2021a) performs fusion and separation over pairs of modality representations. A gated Transformer regulates information exchange to mitigate modality imbalance.

MuLOT (Pramanick et al., 2022) combines self-attention for within-modality modeling with optimal transport for cross-modal alignment. Multimodal attention subsequently integrates the aligned features for sarcasm and humor detection.

MIL (Zhang et al., 2024b) jointly models sarcasm and sentiment through shared and task-specific representations. Cross-modal target attention supports multimodal fusion and information exchange between the two tasks.

TFN (Zadeh et al., 2017) learns both within-modality dynamics and cross-modal interactions through tensor fusion. Its joint representation captures interactions among textual, acoustic, and visual features for sentiment prediction.

LMF (Liu et al., 2018) uses low-rank tensor factorization to construct compact multimodal representations. This reduces the computational cost of tensor-based fusion while retaining its capacity to model cross-modal interactions.

EFT follows the early-fusion Transformer (Vaswani et al., 2017) implementation used in Sun et al. (2023). It integrates modality features before joint Transformer processing to learn a multimodal representation.

LFT follows the late-fusion Transformer (Vaswani et al., 2017) implementation used in Sun et al. (2023). It processes modalities separately with Transformer encoders before combining their outputs for prediction.

PMR (Lv et al., 2021) introduces a shared message hub that exchanges information with individual modalities through cross-modal attention. Repeated exchanges progressively refine both the shared message and modality representations. For CHERMA, the compared variant uses both unimodal and multimodal supervision.

LFMIM (Sun et al., 2023) preserves modality independence through separately supervised unimodal Transformer modules. Information flows from these modules to a multimodal Transformer, which aggregates their representations under multimodal supervision.

B.3 Feature Extraction

We extract textual, acoustic, and visual features using dataset-specific pretrained models and feature extraction tools.

Textual Features. For CMU-MOSI and CMU-MOSEI, we use DeBERTa (He et al., 2021) to obtain contextualized textual representations. For UR-FUNNY and MUStARD, we adopt BERT (Devlin et al., 2019), with the preceding context and target utterance concatenated using the `[SEP]` token. For CHERMA, we employ a Chinese BERT-base model to encode Chinese utterances.

Acoustic Features. For CMU-MOSI, CMU-MOSEI, UR-FUNNY, and MUStARD, we extract acoustic descriptors using COVAREP (Degottex et al., 2014), including Mel-frequency cepstral coefficients, pitch, and other voice characteristics. These features form temporal sequences that capture changes in vocal behavior. For CHERMA, acoustic representations are obtained from a pretrained wav2vec model.

Visual Features. For CMU-MOSI and CMU-MOSEI, we use Facet to extract facial behavior features, including facial action units, landmarks, and head pose. For UR-FUNNY and MUStARD, we use MTCNN (Zhang et al., 2016) for face detection and OpenFace (Baltrušaitis et al., 2018) to extract facial action units, head pose, and gaze features. For CHERMA, faces are aligned using MTCNN and encoded by a ResNet-18 model (He et al., 2016) pretrained on RAF-DB.

B.4 Implementation Details

Table 3: Hyperparameter settings of C^{2}UR across datasets.

Hyperparameter	CMU-MOSI	CMU-MOSEI	UR-FUNNY	MUStARD	CHERMA
Batch size	48	128	128	128	400
Training epochs	50	10	10	20	20
Warm-up	✓	✓	✓	✓	✓
Initial learning rate	2×10^{-5}	4×10^{-5}	9×10^{-6}	2×10^{-5}	2×10^{-5}
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Latent dimension d	352	448	192	464	512
KL regularization coefficient β	$1e-5$	$1e-5$	$1e-5$	$1e-5$	$1e-5$
PRL loss weight λ_1	0.7	0.2	0.1	0.6	0.3
CMUE loss weight λ_2	0.5	0.7	0.6	0.7	0.9

We implement C^{2}UR in PyTorch 2.7.0 with CUDA 12.8 and conduct experiments on an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU. We use AdamW (Loshchilov & Hutter, 2019) with learning-rate warm-up across all datasets. Table 3 summarizes the selected dataset-specific configurations.

For MSA, we conduct 30 random hyperparameter search trials and select the configuration achieving the lowest validation MAE. The KL regularization coefficient β is selected from $\{1e-6, 1e-5, 1e-4, 1e-3, 1e-2\}$, while each auxiliary loss weight, λ_1 and λ_2, is selected from $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. The selected value of β is $1e-5$ across all datasets, while the loss weights vary by dataset.

For a fair comparison, we implement GLoMo (Zhuang et al., 2024), ITHP (Xiao et al., 2024), KAN-MCP (Luo et al., 2025), TMDC (Zhuang et al., 2026), EUAR (Gao et al., 2024a), PML (Hu et al., 2025), HME (Zhuang et al., 2025a), and QA-MoE (Zhu et al., 2026) using the same feature extraction pipeline as C^{2}UR. Each baseline undergoes 30 random search trials within the hyperparameter ranges reported in its original paper, with configurations selected using the same validation criterion.

Continuous Reliability Spectrum. Following the continuous reliability spectrum evaluation of QA-MoE (Zhu et al., 2026), we assess robustness to modality missingness, modality noise, and their joint occurrence. We characterize these degradations using the missing rate η and noise intensity λ, respectively. Consistent with Section 3, degradation is applied to the input feature sequences $\boldsymbol{U}_m^i \in \mathbb{R}^{l_m \times d_m}$, $m \in \{t, a, v\}$, before modality-specific encoding.

Modality Missingness. For each evaluation sample, we randomly mask a subset of modalities while ensuring that at least one modality remains available. Let $b_m^i \in \{0, 1\}$ indicate whether modality m is available for sample i, and let $M_i = \sum_{m \in \{t,a,v\}} b_m^i$ denote the number of available modalities, with $M_i \geq 1$. The missing rate is defined as

$$\eta = 1 - \frac{\sum_{i=1}^{N_{\text{eval}}} M_i}{3N_{\text{eval}}}, \tag{58}$$

where N_{eval} denotes the number of evaluation samples. A missing modality is replaced with an all-zero feature sequence of the same shape as its original input. Since each sample retains at least

one of the three modalities, the maximum missing rate is $2/3$, at which each sample has exactly one available modality. We use 0.7 as a rounded label for this endpoint in the reported figures and tables.

Modality Noise. We introduce feature-level Gaussian corruption by mixing each input feature sequence with a noise matrix:

$$\widetilde{\boldsymbol{U}}_m^i = (1-\lambda)\boldsymbol{U}_m^i + \lambda \boldsymbol{E}_m^i, \qquad [\boldsymbol{E}_m^i]_{pq} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{59}$$

where $\boldsymbol{E}_m^i$ has the same shape as $\boldsymbol{U}_m^i$. The same noise intensity λ is used across modalities, with larger values introducing stronger corruption. Here, λ controls evaluation-time input degradation and is distinct from the training loss weights λ_1 and λ_2.

Joint Degradation and Evaluation. To simulate the simultaneous occurrence of modality noise and missingness, we combine Gaussian corruption and modality masking as

$$\boldsymbol{U}_m^{\deg,i} = b_m^i \left[(1-\lambda)\boldsymbol{U}_m^i + \lambda \boldsymbol{E}_m^i\right]. \tag{60}$$

Applying the availability mask to the entire corrupted sequence ensures that missing modalities remain zero regardless of λ. The degraded sequences are then processed by the modality-specific encoders and the subsequent MUM, CUR, and DEF modules.

We evaluate all combinations of noise intensity and missing rate: $\lambda \in \{0.0, 0.1, \ldots, 0.7\}$, $\eta \in \{0.0, 0.1, \ldots, 0.6, 2/3\}$, yielding an 8×8 grid of 64 degradation conditions, with the endpoint $\eta = 2/3$ labeled as 0.7. The setting $(\lambda, \eta) = (0, 0)$ corresponds to clean and complete inputs; $\lambda = 0$ isolates modality missingness, whereas $\eta = 0$ isolates modality noise. Settings with both parameters positive evaluate compound degradation. For each method, we use the checkpoint trained on clean and complete inputs that produces the standard results reported in the main paper. The same checkpoint is then evaluated over the entire grid without retraining, degradation-aware training, or additional parameter tuning. This protocol is applied consistently to C^{2}UR, HME (Zhuang et al., 2025a), and QA-MoE (Zhu et al., 2026). This protocol assesses robustness across varying input reliability conditions using a fixed model. Figure 3 visualizes the resulting Acc7 performance, and Table 7 reports the discrete results for C^{2}UR.

B.5 Evaluation Metrics

We adopt task-specific metrics to evaluate C^{2}UR. For MSA on CMU-MOSI and CMU-MOSEI, we report Acc7, Acc2, weighted F1, MAE, and Pearson correlation. Accuracy and F1 scores are expressed as percentages. Higher values indicate better performance for all metrics except MAE.

Seven-class Accuracy (Acc7). This metric evaluates fine-grained sentiment classification. Predicted and ground-truth scores are clipped to $[-3, 3]$ and rounded to the nearest integer to obtain seven sentiment categories:

$$\text{Acc7} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\widehat{c}_i = c_i), \tag{61}$$

where $\widehat{c}_i$ and c_i denote the predicted and ground-truth categories, respectively, and $\mathbf{1}(\cdot)$ is the indicator function.

Binary Accuracy (Acc2). Following prior work (Xiao et al., 2024; Luo et al., 2025), we evaluate sentiment polarity on non-neutral samples, distinguishing negative ($y < 0$) from positive ($y > 0$) sentiment:

$$\text{Acc2} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{62}$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Weighted F1-score (F1). Using the same non-neutral samples as Acc2, we compute the F1-score for each class and average the scores according to class frequency:

$$\text{F1} = \sum_{c \in \mathcal{C}} \frac{N_c}{N_{\text{eval}}} \frac{2P_c R_c}{P_c + R_c}, \tag{63}$$

where $\mathcal{C}$ is the set of sentiment classes, N_c is the number of ground-truth samples in class c, and $N_{\text{eval}} = \sum_{c \in \mathcal{C}} N_c$. Here, P_c and R_c denote class-specific precision and recall.

Mean Absolute Error (MAE). We measure sentiment intensity prediction error by averaging the absolute differences between predictions and annotations:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\widehat{y}_i - y_i|, \tag{64}$$

where $\widehat{y}_i$ and y_i are the predicted and ground-truth sentiment scores.

Pearson Correlation (Corr). This metric measures the linear association between predicted and annotated sentiment intensities:

$$\mathrm{Corr} = \frac{\sum_{i=1}^{N}(\widehat{y}_i - \overline{\widehat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\widehat{y}_i - \overline{\widehat{y}})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{65}$$

where $\overline{\widehat{y}}$ and $\bar{y}$ are the mean predicted and ground-truth scores, respectively.

For MHD on UR-FUNNY and MSD on MUStARD, we report binary classification accuracy following prior work (Xu et al., 2025; Huang et al., 2025; Li et al., 2026). These tasks distinguish humorous from non-humorous utterances and sarcastic from non-sarcastic utterances, respectively. For MER on CHERMA, we report F1-scores for the seven emotion categories and the overall F1-score, following Sun et al. (2023).

C More Experimental Results

C.1 Results on Multimodal Humor Detection

Table 4: Comparison on UR-FUNNY. ♡ denotes results reported in the original papers, while the remaining baseline results are taken from (Xu et al., 2025).

Method	Context	Punchline	Accuracy (↑)
C-MFN (Hasan et al., 2019) [EMNLP'19]	✗	✓	64.47
C-MFN (Hasan et al., 2019) [EMNLP'19]	✓	✓	65.23
MulT (Tsai et al., 2019) [ACL'19]	✗	✓	66.65
MISA (Hazarika et al., 2020) [ACM MM'20]	✗	✓	67.34
Self-MM (Yu et al., 2021) [AAAI'21]	✗	✓	68.77
MMIM (Han et al., 2021b) [EMNLP'21]	✗	✓	69.53
FDMER (Yang et al., 2022) [ACM MM'22]	✗	✓	70.55
DMD (Li et al., 2023b) [CVPR'23]	✗	✓	68.70
AGM♡ (Li et al., 2023a) [ICCV'23]	✗	✓	66.07
AtCAF♡ (Huang et al., 2025) [Inf. Fusion'25]	✓	✓	72.13
SuCI (Xu et al., 2025) [AAAI'25]	✗	✓	70.92
DHMD♡ (Li et al., 2026) [TPAMI'26]	✓	✓	71.30
C^{2}UR (ours)	✓	✓	**73.34**

Table 4 compares C^{2}UR with representative methods on UR-FUNNY. C^{2}UR achieves an accuracy of 73.34%, outperforming the strongest baseline, AtCAF (Huang et al., 2025), by 1.21 percentage points and DHMD (Li et al., 2026) by 2.04 percentage points. These results demonstrate the effectiveness of C^{2}UR beyond sentiment analysis.

Humor may be conveyed through complementary signals such as wording, vocal emphasis, and facial expressions, whose predictive value need not coincide with their reliability. One possible explanation for the improvement is that utility–uncertainty decoupling allows C^{2}UR to retain informative cues even when their representations are uncertain. Reliable auxiliary evidence can then refine these representations, while the original-evidence branch preserves modality-specific information. This combination may help exploit subtle humor cues that would otherwise receive insufficient weight.

Table 5: Comparison on MUStARD. ♡ denotes results reported in the original papers, while the remaining baseline results are taken from (Pramanick et al., 2022).

Method	Context	Target	Accuracy (↑)
CIM-MTL (Akhtar et al., 2019) [NAACL'19]	✗	✓	67.14
C-MFN (Hasan et al., 2019) [EMNLP'19]	✗	✓	67.14
C-MFN (Hasan et al., 2019) [EMNLP'19]	✓	✓	70.00
MISA (Hazarika et al., 2020) [ACM MM'20]	✗	✓	66.18
MAG-XLNet (Rahman et al., 2020) [ACL'20]	✓	✓	74.72
BBFN (Han et al., 2021a) [ICMI'21]	✓	✓	71.42
MuLOT (Pramanick et al., 2022) [WACV'22]	✗	✓	74.52
MuLOT (Pramanick et al., 2022) [WACV'22]	✓	✓	76.82
MIL$^{\heartsuit}$ (Zhang et al., 2024b) [TAI'24]	✓	✓	76.30
ITHP$^{\heartsuit}$ (Xiao et al., 2024) [ICLR'24]	✓	✓	75.30
DHMD$^{\heartsuit}$ (Li et al., 2026) [TPAMI'26]	✓	✓	78.50
C^{2}UR (ours)	✓	✓	**80.88**

C.2 Results on Multimodal Sarcasm Detection

As reported in Table 5, C^{2}UR achieves an accuracy of 80.88% on MUStARD, exceeding DHMD (Li et al., 2026) and MuLOT (Pramanick et al., 2022) by 2.38 and 4.06 percentage points, respectively. Its advantage over the compared methods supports the applicability of the proposed framework to sarcasm detection.

Sarcasm presents a particular challenge because disagreement between verbal and nonverbal cues can itself carry task-relevant information. Consequently, indiscriminately enforcing cross-modal agreement may obscure useful evidence. C^{2}UR instead regulates posterior rectification according to auxiliary prediction consistency and retains the original modality evidence through a separate branch. We hypothesize that this design helps balance the use of corroborating evidence with the preservation of informative cross-modal differences, contributing to the observed improvement.

C.3 Results on Multimodal Emotion Recognition

Table 6: Comparison on CHERMA in terms of F1-score. ♢ denotes results obtained from our experiments, while the remaining baseline results are taken from (Sun et al., 2023).

Method	Happiness (↑)	Sadness (↑)	Fear (↑)	Anger (↑)	Surprise (↑)	Disgust (↑)	Neutrality (↑)	Overall (↑)
TFN	74.91	75.56	66.15	74.41	66.29	43.34	65.60	68.37
LMF	74.52	75.83	66.73	74.55	65.08	45.70	65.64	68.23
EFT	74.98	76.88	67.32	74.85	66.73	47.48	64.60	68.72
LFT	75.07	76.29	66.80	74.88	66.67	47.74	65.97	69.05
MulT	76.18	76.88	67.36	74.85	68.18	46.96	65.26	69.24
PMR	75.68	76.46	67.97	75.43	67.37	48.93	66.59	69.53
LFMIM	76.60	77.83	69.44	75.32	69.83	50.20	68.24	70.54
HME$^{\diamondsuit}$	80.64	79.82	72.46	75.86	69.01	48.78	67.74	71.49
TMDC$^{\diamondsuit}$	81.21	79.96	72.12	74.34	68.89	48.92	68.53	71.43
QA-MoE$^{\diamondsuit}$	80.75	81.05	72.36	76.02	67.83	48.01	69.48	71.91
C^{2}UR (ours)	**81.25**	**81.29**	**73.41**	**76.59**	**70.76**	**50.55**	**69.88**	**72.86**

We compare C^{2}UR with ten baselines on CHERMA: TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), early-fusion Transformer (EFT) and late-fusion Transformer (LFT) as implemented in (Sun et al., 2023), MulT (Tsai et al., 2019), PMR (Lv et al., 2021), LFMIM (Sun et al., 2023), HME (Zhuang et al., 2025a), TMDC (Zhuang et al., 2026), and QA-MoE (Zhu et al., 2026). Following Sun et al. (2023), the compared PMR variant uses both unimodal and multimodal labels for training.

Table 6 presents the results on CHERMA. C^{2}UR achieves the highest F1-score among the compared methods in all seven emotion categories, with an overall F1-score of 72.86%, surpassing the second-

best method, QA-MoE (Zhu et al., 2026), by 0.95 percentage points. Compared with QA-MoE, the gains are particularly pronounced for surprise and disgust, reaching 2.93 and 2.54 percentage points, respectively. The consistent improvements across all categories suggest that the benefits are not restricted to a single emotion.

Different emotions and individual samples may rely on different combinations of linguistic, acoustic, and visual evidence. The gains are consistent with the motivation for estimating sample-specific modality utility separately from representation uncertainty. This separation enables informative modalities to contribute even when their evidence is uncertain, while consensus-guided rectification incorporates reliable auxiliary information. Combined with the original-evidence branch, this mechanism offers a possible explanation for the improved recognition across diverse emotion categories.

C.4 Results on Continuous Reliability Spectrum

Table 7: Discrete Evaluation Grid (Acc7%). The table reports the exact performance of a single C^{2}UR checkpoint across varying degradation levels. The blue-shaded cell denotes the compound-degradation setting ($\lambda = 0.3$, $\eta = 0.3$) analyzed in Figure 3.

MR (η)	**Noise Intensity** (λ)							
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.0	**55.46**	53.03	51.54	48.31	46.26	44.71	42.69	41.89
0.1	54.30	52.08	50.96	47.75	45.88	44.30	42.58	41.89
0.2	52.55	50.94	49.71	47.21	45.70	44.20	42.47	41.76
0.3	51.56	50.10	48.91	**46.20**	44.97	43.76	42.39	41.74
0.4	50.25	48.63	47.99	45.55	44.48	43.38	42.15	41.74
0.5	49.06	47.56	47.12	44.82	44.02	43.01	41.93	41.59
0.6	47.66	46.61	46.03	44.39	43.53	42.62	41.83	41.55
0.7	46.18	45.19	44.86	43.51	42.82	42.21	41.72	41.48

Table 7 provides the complete numerical results underlying the C^{2}UR reliability landscape in Figure 3. We evaluate Acc7 on CMU-MOSEI over a discrete grid of noise intensity λ and modality missing rate η, both ranging from 0 to 0.7 in increments of 0.1. All settings use the same trained checkpoint without additional tuning.

C^{2}UR achieves 55.46% Acc7 under clean and complete inputs ($\lambda = 0, \eta = 0$). Performance generally declines as either noise or missingness increases, with combined degradation posing a greater challenge. At the highlighted compound-degradation setting ($\lambda = 0.3, \eta = 0.3$), C^{2}UR achieves 46.20% Acc7, compared with 48.31% under noise alone ($\lambda = 0.3, \eta = 0$) and 51.56% under missingness alone ($\lambda = 0, \eta = 0.3$). Even at the most severe evaluated combination ($\lambda = 0.7, \eta = 2/3$, with the latter labeled as 0.7), it maintains 41.48% Acc7.

Together with the comparisons against HME and QA-MoE in Figure 3, these results show that C^{2}UR maintains a broader high-performance region and degrades more smoothly as input reliability decreases. A possible explanation is that utility-aware evidence preservation retains informative cues, while consistency-guided rectification selectively incorporates reliable auxiliary evidence. These findings support the applicability of a single C^{2}UR checkpoint across varying combinations of noise and modality missingness.

C.5 Detailed Ablation Study

Table 8 reports all variants on CMU-MOSI and CMU-MOSEI. The following names distinguish component removal from loss removal and follow the notation in Section 3.

Within MUM, 'w/o CMUE' fixes $\widehat{I}_m = 1/3$ and removes $\mathcal{L}_{CMUE}$; 'w/o UUD' (utility–uncertainty decoupling) instead sets $\widehat{I}_m = (1 - u_m)/\sum_j (1 - u_j)$ and removes the same loss. 'w/o $\mathcal{L}_{CMUE}$' retains the learned utility predictor but removes its removal-based supervision. 'w/o $\mathcal{L}_{PRL}$' removes the entire PRL objective while retaining Gaussian parameterization and uncertainty computation; 'w/o KL' removes only its prior regularizer. Within CUR, 'w/o CUR' uses the original posterior without

Table 8: Detailed ablation results of C^{2}UR on CMU-MOSI and CMU-MOSEI.

Method	CMU-MOSI					CMU-MOSEI				
	Acc7%↑	*Acc2%*↑	*F1%*↑	*MAE*↓	*Corr*↑	*Acc7%*↑	*Acc2%*↑	*F1%*↑	*MAE*↓	*Corr*↑
w/o CMUE	48.61	88.40	88.39	0.641	0.847	54.19	85.41	85.55	0.528	0.777
w/o UUD	45.55	88.09	88.11	0.655	0.848	54.17	86.60	86.69	0.524	0.788
w/o CUR	46.72	87.63	87.62	0.641	0.848	54.00	86.57	86.51	0.512	0.793
w/o CG	49.20	87.48	87.48	0.644	0.848	53.84	86.33	86.25	0.523	0.791
w/o RW	46.72	87.48	87.48	0.640	0.845	55.16	87.40	87.36	0.509	0.796
w/o UG	45.84	88.40	88.40	0.652	0.847	54.40	86.99	86.94	0.512	0.792
w/o UB	43.80	86.56	86.63	0.675	0.843	52.92	86.85	86.82	0.525	0.777
w/o CB	48.76	87.18	87.18	0.641	0.852	53.46	86.93	86.80	0.531	**0.797**
w/o $\mathcal{L}_{PRL}$	48.91	88.09	88.03	0.636	0.848	54.99	86.85	86.81	0.516	0.790
w/o KL	47.88	88.70	88.70	0.625	0.853	55.05	87.10	87.15	0.515	0.792
w/o $\mathcal{L}_{CMUE}$	49.05	88.24	88.24	0.634	0.849	55.29	87.27	87.23	0.521	0.791
C^{2}UR	**49.49**	**89.92**	**89.93**	**0.618**	**0.855**	**55.46**	**87.54**	**87.49**	**0.505**	0.795

rectification; 'w/o CG' (consistency gating) sets $C_m = 1$; 'w/o RW' (reliability weighting) sets auxiliary exponents $\rho_j = 1$ while retaining Gaussian precision weighting; and 'w/o UG' (uncertainty gating) uses $\gamma_m = \widehat{I}_m C_m$. Within DEF, 'w/o UB' (utility branch) sets $\boldsymbol{F}_u = \mathbf{0}$, while 'w/o CB' (consensus branch) sets $\boldsymbol{F}_c = \mathbf{0}$, preserving the final fusion dimensions.

Utility Modeling and Probabilistic Supervision. Removing CMUE or UUD worsens every reported metric on both datasets. The particularly pronounced Acc7 degradation of 'w/o UUD' on CMU-MOSI suggests that equating utility with reliability can overlook uncertain but discriminative cues. Uniform utility also weakens performance, indicating that treating all modalities as equally useful is insufficient. Retaining an unsupervised utility predictor recovers some performance relative to uniform weighting, especially on CMU-MOSEI, but remains inferior to the complete model. This is consistent with removal-based supervision providing task-specific guidance beyond that learned indirectly through fusion. Removing either $\mathcal{L}_{PRL}$ or KL also degrades all metrics. These outcomes suggest complementary roles for unimodal supervision and distribution regularization.

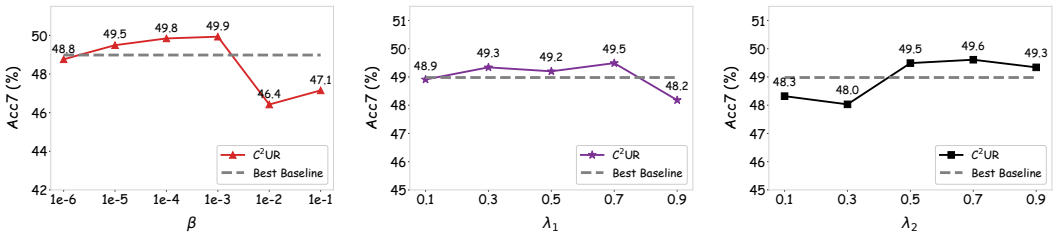
Selective Cross-modal Rectification. Bypassing CUR consistently harms performance, supporting the use of auxiliary evidence to refine modality posteriors. Removing UG produces a performance drop on CMU-MOSI, plausibly because relatively certain targets receive unnecessary correction. Removing CG also degrades all metrics, suggesting that confident auxiliary evidence should not be accepted without considering predictive agreement. RW has a more dataset-dependent effect: its removal substantially harms CMU-MOSI but produces smaller changes on CMU-MOSEI, including a slightly higher Corr. One possible explanation is that the remaining precision weighting already provides useful reliability information, making the additional scalar weighting less critical in some settings.

Complementarity of the Two Evidence Branches. Removing UB causes the largest Acc7 decline on both datasets; on CMU-MOSI it also raises MAE from 0.618 to 0.675. This suggests that consensus aggregation can lose modality-specific distinctions preserved by the original-evidence branch. Removing CB likewise degrades classification and MAE, with the largest MAE increase among the MOSEI variants, supporting the complementary contribution of rectified evidence. The full model achieves the best results on all five MOSI metrics and four MOSEI metrics, supporting the combined design.

C.6 Parameter Sensitivity Analysis

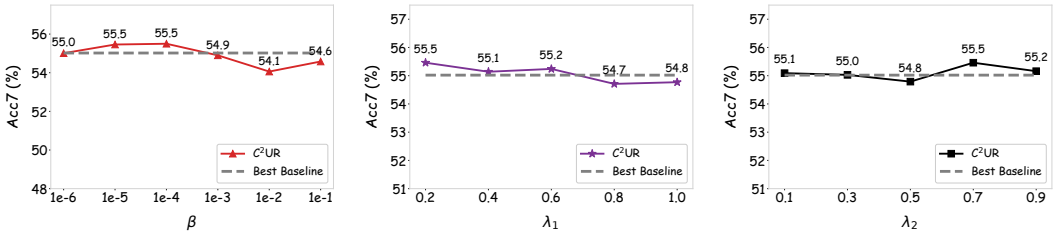
We examine the sensitivity of C^{2}UR to three hyperparameters: the KL regularization coefficient β, the PRL loss weight λ_1, and the CMUE loss weight λ_2. Figures 6 and 7 report Acc7 on CMU-MOSI and CMU-MOSEI, respectively. Dashed lines indicate the strongest baseline performance on each dataset.

(1) KL regularization coefficient β. The coefficient β controls how strongly the modality distributions are regularized toward the standard Gaussian prior. On CMU-MOSI, Acc7 increases from



(a) KL regularization coefficient β. (b) PRL loss weight λ_1. (c) CMUE loss weight λ_2.

Figure 6: Parameter sensitivity analysis of β, λ_1, and λ_2 on CMU-MOSI.



(a) KL regularization coefficient β. (b) PRL loss weight λ_1. (c) CMUE loss weight λ_2.

Figure 7: Parameter sensitivity analysis of β, λ_1, and λ_2 on CMU-MOSEI.

48.76% to 49.93% as β rises from 10^{-6} to 10^{-3}, but drops at larger values. On CMU-MOSEI, the highest Acc7 of 55.50% occurs at $\beta = 10^{-4}$, with weaker performance under stronger regularization. These results suggest that a relatively small KL penalty provides useful regularization, whereas an excessive constraint may restrict task-discriminative information. The default value $\beta = 10^{-5}$ achieves competitive performance on both datasets.

(2) PRL loss weight λ_1. The weight λ_1 balances probabilistic representation learning against the final prediction objective. On CMU-MOSI, C^{2}UR exceeds the strongest baseline at the tested values from 0.3 to 0.7, reaching 49.49% at 0.7 before declining to 48.18% at 0.9. On CMU-MOSEI, the best result is 55.46% at 0.2, and performance remains above the baseline at 0.4 and 0.6. These trends suggest that suitable unimodal supervision benefits representation learning, while assigning excessive weight to the auxiliary objective may compromise the final multimodal prediction objective. The different preferred values also indicate the need for dataset-specific balancing.

(3) CMUE loss weight λ_2. The weight λ_2 controls supervision of the modality utility predictor. On CMU-MOSI, performance is stronger at the tested values from 0.5 to 0.9 than at 0.1 and 0.3, reaching its highest Acc7 of 49.61% at 0.7. On CMU-MOSEI, Acc7 varies within a relatively narrow range of 54.79%–55.46%, with the highest value also obtained at 0.7. This pattern suggests that sufficiently weighted utility supervision helps estimate modality contributions, although the response is not monotonic. The default values are 0.5 and 0.7 for CMU-MOSI and CMU-MOSEI, respectively, following validation-based selection.

Overall, C^{2}UR performs competitively across several neighboring parameter settings, although overly strong KL regularization or PRL weighting can reduce performance. Notably, several alternative settings even outperform the configuration reported in Table 1. For example, setting $\beta = 10^{-3}$ on CMU-MOSI achieves 49.93% Acc7, compared with the reported 49.49%, while $\beta = 10^{-4}$ on CMU-MOSEI achieves 55.50%, compared with 55.46%. These results indicate that C^{2}UR is not overly sensitive to specific hyperparameter choices and remains stable within a reasonably broad range of configurations.

C.7 Testing Stability Analysis

Table 9 reports the performance of C^{2}UR across five random seeds on CMU-MOSI and CMU-MOSEI. We report the mean and 95% confidence interval for each metric, calculated as $\bar{x} \pm t_{0.975,4} s/\sqrt{5}$,

Table 9: Performance of C^{2}UR across five random seeds, reported as means with 95% confidence intervals.

Dataset	*Acc7%*↑	*Acc2%*↑	*F1%*↑	*MAE*↓	*Corr*↑
CMU-MOSI	49.75 ± 0.67	89.44 ± 0.88	89.45 ± 0.89	0.620 ± 0.009	0.856 ± 0.002
CMU-MOSEI	55.29 ± 0.39	87.39 ± 0.27	87.32 ± 0.30	0.510 ± 0.008	0.791 ± 0.006

where $\bar{x}$ and s denote the sample mean and sample standard deviation across the five runs, respectively, and $t_{0.975,4} \approx 2.776$. The narrow confidence intervals suggest stable performance across random seeds on both datasets. Moreover, even the worst observed value of each metric across the five runs remains competitive with the baselines, indicating that the performance gains are not confined to favorable random seeds.

C.8 Model Complexity Analysis

Table 10: Comparison of model complexity on the CMU-MOSI dataset with batch size 48.

Method	FLOPs	Memory	Parameters	Per-Epoch Training Time
GLoMo (Zhuang et al., 2024)	12.25 G	8.88 GB	191.92 M	2.96 s
ITHP (Xiao et al., 2024)	12.03 G	8.65 GB	184.88 M	2.19 s
KAN-MCP (Luo et al., 2025)	12.98 G	14.32 GB	184.76 M	6.54 s
EUAR (Gao et al., 2024a)	12.91 G	15.44 GB	212.28 M	3.12 s
TMDC (Zhuang et al., 2026)	13.12 G	9.36 GB	202.26 M	2.64 s
PML (Hu et al., 2025)	12.14 G	8.75 GB	188.27 M	2.47 s
HME (Zhuang et al., 2025a)	12.85 G	12.01 GB	228.53 M	7.05 s
QA-MoE (Zhu et al., 2026)	14.26 G	10.39 GB	232.72 M	3.64 s
C^{2}UR (ours)	12.51 G	8.74 GB	198.91 M	2.93 s

We compare C^{2}UR with representative baselines in terms of FLOPs, memory usage, parameter count, and per-epoch training time on CMU-MOSI with a batch size of 48. As shown in Table 10, C^{2}UR requires 12.51G FLOPs and 198.91M parameters, placing it within the range of the compared methods. Its memory usage is 8.74GB, the second lowest among all methods, and its training time of 2.93s per epoch is comparable to GLoMo (Zhuang et al., 2024).

Compared with the uncertainty-aware baselines EUAR (Gao et al., 2024a), HME (Zhuang et al., 2025a), and QA-MoE (Zhu et al., 2026), C^{2}UR uses fewer FLOPs and parameters, consumes less memory, and trains faster. In particular, its per-epoch training time is substantially lower than that of HME (7.05s) and KAN-MCP (Luo et al., 2025) (6.54s). Although C^{2}UR is not the least expensive model across all metrics, these results, together with its predictive performance, demonstrate a favorable balance between effectiveness and computational cost.