Modality-Aware Contrastive and Uncertainty-Regularized Emotion Recognition

Anonymous authors
Paper under double-blind review

Abstract

Multimodal Emotion Recognition (MER) has attracted growing attention with the rapid advancement of human–computer interaction. However, different modalities exhibit substantial discrepancies in semantics, quality, and availability, leading to highly heterogeneous modality combinations and posing significant challenges to achieving consistent and reliable emotion understanding. To address this challenge, we propose the Modality-Aware Contrastive and Uncertainty-Regularized (MCUR) framework, which approaches MER from the perspective of representation consistency, aiming to enable robust emotion prediction across heterogeneous modality combinations. MCUR incorporates two core components: (1) Modality Combination-Based and Category-Based Contrastive Learning mechanism (MCB-CL), which encourages samples with the same emotion category and the same available modalities to be close in the representation space; and (2) Sample-wise Uncertainty-Guided Regularization (SUGR), which adaptively assigns sample-wise uncertain weights to samples to optimize training. Extensive experiments demonstrate that MCUR consistently outperforms existing methods, achieving average F1 gains of 2.2% on MOSI, 2.67% on MOSEI, and 4.37% on IEMOCAP.

1 Introduction

Multimodal emotion recognition (MER) aims to infer the emotional state of an utterance by integrating cues from language, vision, and audio. With the rapid progress of human-computer interaction technologies, MER has become a crucial research direction (Lian et al., 2025; Fang et al., 2025b; Zeng et al., 2023). However, real-world MER systems still face strong modality heterogeneity, as different modalities exhibit significant discrepancies in: (1) Semantics: text conveys explicit emotional polarity, while audio reflects intensity and rhythm, and visual cues express subtle affective nuances; (2) Availability: missing or degraded modalities naturally lead to diverse modality combinations, resulting in drastic input distribution shifts that further exacerbate the above inconsistencies and hinder cross-modal alignment and robust fusion.

Previous studies primarily focus on recovering missing information or transferring knowledge from complete modalities. For instance, reconstruction-based methods aim to infer missing modalities from the available ones (e.g., MPLMM (Guo et al., 2024), DiCMoR (Wang et al., 2023a), LNLN (Zhang et al., 2024)), while knowledge distillation (KD)-based methods (e.g., CorrKD (Li et al., 2024c), MMANet (Wei et al., 2023)) transfer knowledge from a teacher trained on complete modalities to a student exposed to partial inputs. Although these methods can improve cross-modal complementarity, they seldom impose explicit constraints on the structure of the representation space under different modality combinations. As the available evidence changes, the same emotional semantics may be mapped to different regions of the embedding space, leading to inconsistent representations and suboptimal performance under heterogeneous missing conditions. As illustrated in Figure 1, this inconsistency manifests in two forms: (1) intra-combination inconsistency, where samples with the same emotion label and modality combination are dispersed in the representation space; and (2) cross-combination prediction variation, where the same sample may yield different

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.

predictions under different available modality subsets (e.g., L+V vs. A+V). These observations motivate us to regulate representation consistency within comparable observations while avoiding uniform supervision across heterogeneous modality combinations.

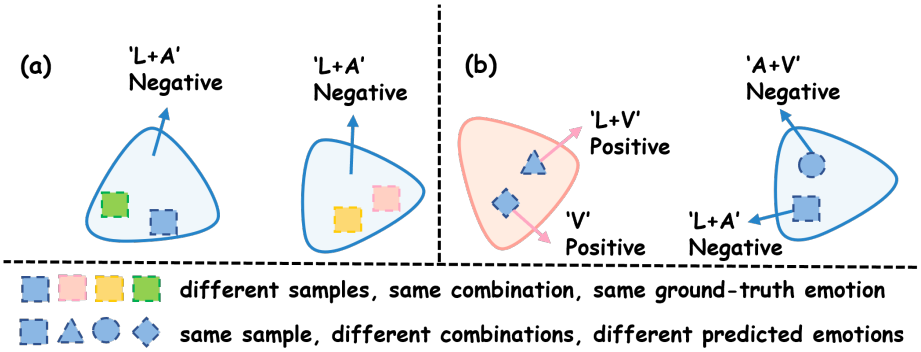

Figure 1: Illustration of the representation inconsistency challenges. (a) Intra-combination inconsistency. (b) Cross-combination prediction variation.

To tackle these challenges, we propose MCUR, a modality-aware contrastive and uncertainty-regularized framework that makes both representation-level and teacher-level supervision aware of the observed modality combination. MCUR consists of two key modules: (1) Modality Combination-Based and Category-Based Contrastive Learning (MCB-CL), which jointly models modality combinations and emotion categories to enforce contrastive constraints across modality subspaces, thereby enhancing both consistency and discriminability; (2) Sample-wise Uncertainty-Guided Regularization (SUGR), which quantifies the uncertainty discrepancy between teacher and student predictions to dynamically adjust task and distillation weights, achieving adaptive balance across heterogeneous inputs. Our key contributions are summarized as follows:

- We identify representation inconsistency and unreliable supervision as key challenges under heterogeneous modality combinations, and propose MCUR to address them jointly.
- We introduce MCB-CL, a contrastive learning mechanism that jointly considers modality combinations and label classes to enhance representation consistency and discriminability.
- We propose SUGR, an uncertainty-aware regularization strategy that adaptively weights teacher supervision according to teacher-student prediction uncertainty.

2 Related Work

2.1 MER with Missing Modalities

Multimodal emotion recognition (MER) aims to predict the emotion of a given utterance by integrating signals from language, video, and audio (Zhuang et al., 2024; Fang et al., 2025a; Zhuang et al., 2026c). Existing approaches to address MER with missing modalities can be broadly categorized into two groups: modality reconstruction-based methods and knowledge distillation (KD)-based methods. The former methods estimate the missing modality using the available ones, thereby recovering the original information (Wang et al., 2023a; Zhang et al., 2024; Zhuang et al., 2026c; Wang et al., 2023b; Lian et al., 2023; Zhuang et al., 2026b). For instance, IMDer (Wang et al., 2023b) employs a diffusion model to reconstruct missing modalities by comparing the original and reconstructed representations. MPLMM (Guo et al., 2024) pre-trains on complete datasets and uses several prompts to learn information from each modality, which is then applied to reconstruct the missing modalities. LNLN (Zhang et al., 2024), on the other hand, focuses on the language modality, leveraging it as the dominant source of information in incomplete scenarios. While these methods show

promise, they often treat modality reconstruction as an isolated task and struggle to produce robust, unified representations across diverse missing-modality conditions. In contrast, our proposed MCUR, is a KD-based framework that learns directly from fused representations without explicit modality reconstruction. Recent studies further explore uncertainty, modality conflict, calibrated representations, and dynamic modality selection under missing modalities (Qiu et al., 2026; Wang et al., 2026b; Jiang et al., 2026; He et al., 2026; Wang et al., 2026a; Liu et al., 2026; Du et al., 2026), while MCUR focuses on modality-aware supervision through positive-pair construction and teacher weighting. By jointly considering modality combinations and emotion categories, MCUR enhances both consistency and discriminability in learned features, and learns more robust representations across varying missing-modality scenarios.

2.2 Knowledge Distillation

Knowledge distillation (KD)-based methods (Hinton, 2015; Tung & Mori, 2019; Li et al., 2024b) aim to preserve relationships between samples while reconstructing fused representations. These approaches typically involve a teacher model, trained with full modality data, and a student model, which learns from datasets with missing modalities. The teacher guides the student by providing sample representations (Li et al., 2024a), inter-sample relationships (Li et al., 2024c), and logits (Li et al., 2024a; Zhuang et al., 2025), helping improve the student's performance. For example, MMANet (Wei et al., 2023) transfers inter-sample relationships and strengthens weaker modality representations. Although effective, these methods typically overlook consistency within the same emotion category under specific missing-modality patterns, as well as sample-wise differences in prediction uncertainty. In contrast, our proposed MCUR addresses these limitations by leveraging both modality combinations and emotion categories to enhance representation consistency and discriminability, and dynamically adjusting the loss weights based on the prediction uncertainty of each sample.

3 Methodology

The structure of the proposed MCUR is illustrated in Figure 2. MCUR involves two modules: Modality Combination-Based and Category-Based Contrastive Learning (MCB-CL) module and Sample-wise Uncertainty-Guided Regularization (SUGR) module.

3.1 Problem Definition

In the context of MER with missing modalities, there are m modalities in the dataset $\mathcal{D} = \{X_1, X_2, ..., X_m\}$, where $X_p \in \mathbb{R}^{t_p \times d_p}$ denotes the modality representations and $p \in \{1, 2, ..., m\}$ denotes the modality. Here d_p and t_p represent the dimensionality and sequence length for each modality, respectively. The goal is to train a robust model that can handle samples with various missing modalities simultaneously. Specifically, for modality p, we define $\delta_p \in \mathbb{R}^N$ to indicate the presence or absence of the modality. Here N is the size of the dataset. Each element $\delta_{p,i}$ in δ_p is either 0 (missing modality) or 1 (available modality), with the condition that $\delta_{1,i} + \delta_{2,i} + ... + \delta_{m,i} > 0$, meaning that each sample must have at least one modality available.

3.2 MCUR Overview

MCUR follows the standard KD paradigm (Wei et al., 2023; Li et al., 2024c; Hinton, 2015), employing a teacher-student architecture, as shown in Figure 2. The teacher model is pretrained using full modalities and the student model is trained with limited modalities. And the parameters in the teacher model are frozen when the student model is training. Given multimodal input $x = \{X_1, X_2, ..., X_m\}$, we randomly drop certain modalities (e.g., X_j, X_k) in the student model by setting $\delta_j = \delta_k = 0$ to construct controlled missing-modality conditions. All modalities are first passed through a 1D convolutional layer to capture temporal patterns and standardize their shape to $X_i \in \mathbb{R}^{T \times D}$, where T and D are fixed hyper-parameters. Each modality is then encoded via a modality-specific Perceiver encoder

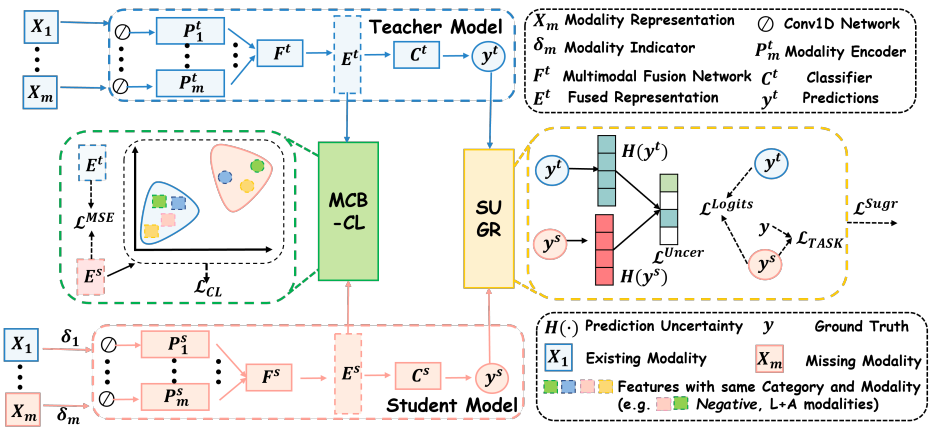

Figure 2: The structure of the MCUR framework. MCUR includes two key modules: Modality Combination-Based and Category-Based Contrastive Learning (MCB-CL), and Sample-wise Uncertainty-Guided Regularization (SUGR). MCB-CL brings samples with the same category and modality combination closer in the representation space, while SUGR adaptively weights teacher supervision according to prediction uncertainty.

P_j^w (Jaegle et al., 2021; Liang et al., 2023), where $j \in \{1, ..., m\}$ and $w \in \{s, t\}$ denotes student or teacher. Encoded features are fused by a transformer encoder (Vaswani et al., 2017) followed by an MLP (Hazarika et al., 2020) and a Variational Information Bottleneck (VIB) (Alemi et al., 2022; Gao et al., 2024; Kingma, 2013; Tishby et al., 2000), forming the fusion module F^w, which outputs the fused embedding $E^w \in \mathbb{R}^D$. These embeddings from both models are further aligned through the MCB-CL module to ensure representation consistency. Finally, the classifier C^w predicts logits $y^w \in \mathbb{R}^k$, which, along with ground truth $y \in \mathbb{R}^k$, are used by the SUGR module to dynamically weight supervision based on prediction uncertainty. Implementation details of teacher model, P_j^w and F^w are provided in the Appendix A.2.

3.3 MCB-CL

In this section, we present the Modality Combination-Based and Category-Based Contrastive Learning (MCB-CL) mechanism. Existing methods for handling missing modalities often focus on reconstructing the representations to improve predictions, typically expressed as $p(k|x_i)$, where x_i represents the ith sample's representation and k is its corresponding category (Huan et al., 2023). Some approaches also leverage CL (Li et al., 2024c; Zhuang et al., 2025), considering representations with the same label as positive pairs. However, under heterogeneous modality availability, same-label samples may provide different or conflicting evidence. Specifically, these methods fail to consider the modality combinations (c_i), which can provide valuable information, especially when certain modality combinations under-perform. We therefore additionally condition positive-pair construction on the observed modality combination. MCB-CL integrates both modality combinations and emotion classes into the CL framework. Specifically, MCB-CL improves $p(k|x_i)$ by incorporating modality combinations c_i alongside it. Inspired by the conditional structure in Bayes' rule, we rewrite $p(k|x_i)$ as:

$$p(k|x_i) = p(c_i|x_i)\frac{p(k|c_i, x_i)}{p(c_i|x_i, k)}. \tag{1}$$

Inspired by existing works (Peng et al., 2024; Shen et al., 2021; Tsai et al., 2020), we also apply CL to parametrize each term. First, for $p(c_i|x_i)$, which identifies samples within the batch that share the same modality combination as the current sample's representation, we use supervised multi-instance CL as:

$$p(c_i|x_i) \triangleq \frac{\sum_{j \in \mathcal{M}_i} exp(x_i x_j^\top / \tau)}{\sum_{l=1}^{B} exp(x_i x_l^\top / \tau)}. \tag{2}$$

Here B is the batch size, and τ is temperature hyper-parameter controlling the concentration level. $\mathcal{M}_i = \{j | x_j \in E^s, c_j = c_i, j \leq B\}$ represents the indices of samples in the current batch with the same modality combination as x_i.

Next, for $p(k|c_i, x_i)$, which focuses on identifying samples with the same category within a given modality combination, we also apply multi-instance supervised CL. To avoid introducing additional variables, we use j and l for samples during the search:

$$p(k|c_i, x_i) \triangleq \frac{\sum_{j \in \mathcal{N}_i} exp(x_i x_j^\top / \tau)}{\sum_{l \in \mathcal{M}_i} exp(x_i x_l^\top / \tau)}. \tag{3}$$

Here, $\mathcal{N}_i = \{j | x_j \in E^s, c_j = c_i \& k_j = k, j \leq B\}$ represents the indices of samples with both the same modality combination c_i and the same category k as x_i. $\mathcal{M}_i$ is the same as in Equation 2, but j is replaced by l here.

Finally, for $p(c_i|x_i, k)$, which aims to find modality combinations corresponding to a specific category, we again apply supervised CL, using j and l for samples during the search:

$$p(c_i|x_i, k) \triangleq \frac{\sum_{j \in \mathcal{N}_i} exp(x_i x_j^\top / \tau)}{\sum_{l \in \mathcal{S}_i} exp(x_i x_l^\top / \tau)}. \tag{4}$$

Here, $\mathcal{S}_i = \{l | x_l \in E^s, k_l = k, l < B\}$ is the indices of samples with the same category k as x_i. While the Bayesian identity provides a theoretical foundation for joint probability decomposition, the implicit coupling between these terms in a discrete contrastive setting may limit the model's ability to fine-tune the importance of different alignment goals.

To provide more explicit control over the modality-aware feature alignment and enhance optimization stability, we decouple the components of the Bayesian identity by introducing two hyperparameters: μ_1 and μ_2. Consequently, the final MCB-CL loss for student representations E^s is formulated as:

$$\mathcal{L}_{CL}^s = -\mathbb{E}_i \left[\log p(c_i|E_i^s) + \mu_1 \log p(k|c_i, E_i^s) - \mu_2 \log p(c_i|E_i^s, k) \right]. \tag{5}$$

Consistent with existing KD methods (Li et al., 2024c; Zhuang et al., 2025), we also preserve the consistency between the teacher's and student's representations, which is achieved by calculating the representation difference through:

$$\mathcal{L}_i^{MSE} = ||E_i^s - E_i^t||^2. \tag{6}$$

3.4 SUGR

In this section, we introduce the Sample-wise Uncertainty-Guided Regularization (SUGR) module to address observation-dependent prediction uncertainty. Since the teacher observes complete modalities while the student operates under incomplete observations, their prediction uncertainty can vary across modality combinations. SUGR uses this discrepancy to adaptively emphasizes samples where the absence of certain modalities leads to greater uncertainty and assigns higher weights. To quantify the prediction uncertainty, we adopt different measures for classification and regression tasks. For classification, we compute the entropy of the predicted probability distribution (Wei et al., 2023) :

$$H(y_i^w) = -\sum \sigma(y_i^w) \log(\sigma(y_i^w)), \tag{7}$$

where $\sigma(y_i^w)$ is the softmax probability output of the model. For regression, the Mean Squared Error (MSE) between the predicted value and the ground truth is applied:

$$H(y_i^w) = |y_i - y_i^w|^2. \tag{8}$$

The absolute uncertainty difference between the teacher and student uncertainty is then computed to measure the increased uncertainty:

$$\mathcal{L}_i^{Uncer} = |H(y_i^t) - H(y_i^s)|, \tag{9}$$

where a larger discrepancy implies higher uncertainty and leads to a greater penalty during optimization. To ensure reliable prediction, logits distillation is also adapted as in prior work

(Li et al., 2024c;a; Zhuang et al., 2025). For regression tasks, the MSE between teacher and student predictions is adapted:

$$\mathcal{L}_i^{Logits} = |y_i^s - y_i^t|^2. \tag{10}$$

For classification tasks, we use Decoupled Knowledge Distillation (DKD) (Zhao et al., 2022), which separately distills both target and non-target class probabilities:

$$DKD(y^s, y^t) = \alpha KL(b^t||b^s) + (1-\alpha)KL(\hat{p}^t||\hat{p}^s), \tag{11}$$

where b^t and b^s are the predicted probabilities of the target class, and $\hat{p}^t$ and $\hat{p}^s$ correspond to non-target classes. α is a hyper-parameter balancing the two components.

Additionally, the task-specific loss $\mathcal{L}_{TASK}$ is adapted for downstream tasks. For regression, we use the Mean Absolute Error:

$$\mathcal{L}_{TASK,i}^s = |y_i - y_i^s|. \tag{12}$$

For classification, we use standard cross-entropy loss:

$$\mathcal{L}_{TASK,i}^s = CrossEntropy(y_i, y_i^s). \tag{13}$$

Finally, all above loss terms are combined into the overall SUGR loss. The uncertainty difference modulates the influence of both the task-specific and the logits distillation loss:

$$\mathcal{L}_i^{Sugr} = \mathcal{L}_i^{Uncer} \cdot (\mathcal{L}_{TASK,i}^s + \mathcal{L}_i^{Logits}). \tag{14}$$

The overall training loss for MCUR framework is obtained through:

$$\mathcal{L}_{all} = \gamma \mathcal{L}_{CL}^S + \mathcal{L}^{MSE} + \zeta \mathcal{L}^{Sugr}, \tag{15}$$

where γ, and ζ are hyper-parameters.

4 Experiments

4.1 Datasets and Implementation Details

In line with prior work (Guo et al., 2024; Li et al., 2024c), three widely used MER datasets are used for evaluation: MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and IEMOCAP (Busso et al., 2008). For the MOSI and MOSEI datasets, we use accuracy (ACC) and F1 score computed for 'positive/negative' classification results as evaluation metrics. For IEMOCAP dataset, we calculate the weighted accuracy and F1 score of four emotion categories. We evaluate model performance under two protocols: fixed missing and random missing. In the fixed missing protocol, we test each model with one specific modality missing. For instance, in Table 1, the label 'L' indicates that only the language modality is available. In the random missing protocol, modalities are randomly dropped for each sample, as in prior work (Lian et al., 2023; Wang et al., 2023b;a). To quantify the extent of missing modalities, we used the missing rate (MR), defined as: $MR = 1 - \frac{\sum_{i=1}^{N} a_i}{N \times m}$. Here a_i represents the number of available modalities in the ith sample, N is the number of samples and m is the number of modalities in each sample. Given that each sample must contain at least one modality, we ensure that $a_i > 0$ and $MR \leq \frac{m-1}{m}$. The MR values are set between 0.1 and 0.7, with 0.7 approximating the maximum missing rate $\frac{2}{3}$. More implementation details are shown in Appendix B.

4.2 Comparison with State-of-the-art Methods

We compare MCUR with representative methods, including both modality reconstruction-based models (MPLMM (Guo et al., 2024), IMDer (Wang et al., 2023b), and LNLN (Zhang et al., 2024)) and KD-based models (CorrKD (Li et al., 2024c), MMANet (Wei et al., 2023)).

We use official implementations and apply the same training settings across all models. Each model is evaluated under 14 missing modality scenarios—seven fixed and seven random. Additional results on MOSI dataset are reported in Appendix E.

Results are presented in Table 1, and 'Avg' denotes the average performance over all scenarios, with the best scores highlighted in bold. MCUR consistently outperforms all baselines across most scenarios and datasets. Interestingly, KD models generally outperform reconstruction ones. This may be because reconstruction methods focus on recovering individual modalities, which can accumulate errors during fusion. In contrast, KD models directly guide the student to approximate the fused representations, reducing such error amplification. We also observe that for most baseline models, accuracy scores are notably higher than F1 scores, suggesting a tendency to overfit to frequent emotions. In contrast, MCUR yields more balanced accuracy and F1 results, indicating better representation consistency and more stable predictions.

Table 1: Performance comparison in datasets. 'ACC/F1' is reported.

Avail	CorrKD	MPLMM	IMDer	LNLN	MMANet	MCUR
Results on MOSEI						
L	85.66/85.64	84.87/84.94	85.31/85.01	84.98/85.03	85.50/85.34	**85.97/85.88**
V	63.65/48.51	61.39/53.49	62.33/57.04	62.47/51.44	62.96/58.75	**65.00/64.59**
A	62.85/48.51	62.80/48.49	62.85/48.51	62.85/48.51	63.29/51.54	**63.48/60.35**
L,V	86.08/86.02	85.09/85.09	85.31/85.31	85.80/85.80	85.44/85.27	**86.13/86.07**
L,A	85.69/85.68	84.84/84.89	85.33/85.00	85.53/85.54	85.55/85.43	**85.97/85.88**
A,V	64.00/55.83	62.66/50.85	61.97/53.23	62.82/49.42	64.34/61.41	**64.36/64.42**
L,A,V	85.99/85.97	85.20/85.13	85.86/85.84	86.02/85.98	84.98/84.99	**86.10/86.04**
0.1	83.16/82.96	83.02/82.78	**83.82/83.64**	83.71/83.48	81.89/81.77	83.46/83.39
0.2	**81.98/81.56**	80.49/79.97	80.96/80.37	81.40/80.88	80.21/79.91	81.26/81.17
0.3	79.06/78.25	79.28/78.45	79.00/78.34	79.00/78.06	78.23/77.56	**79.88/79.76**
0.4	77.02/75.72	75.98/74.36	73.20/70.89	77.41/76.03	76.97/75.99	**77.55/77.32**
0.5	75.18/73.54	73.45/71.21	69.90/66.38	73.69/71.14	73.91/72.16	**75.89/75.64**
0.6	70.64/67.56	71.79/68.84	72.15/69.27	71.41/67.79	72.29/69.72	**73.50/72.87**
0.7	70.61/67.51	69.21/65.52	65.58/59.33	70.53/66.24	70.83/67.94	**72.59/71.82**
Avg.	76.54/73.88	75.72/72.43	75.26/72.01	76.26/72.52	76.17/74.13	**77.23/76.80**
Results on IEMOCAP						
L	75.37/72.46	75.83/70.03	75.03/64.72	76.49/72.55	73.85/71.86	**79.58/77.89**
V	74.57/66.20	71.91/64.77	75.00/65.07	74.60/67.85	75.08/67.38	**75.40/70.23**
A	77.05/73.25	76.79/68.62	74.65/65.50	76.55/68.74	75.00/64.66	**79.42/76.84**
L,V	75.99/73.34	75.69/70.47	75.32/65.90	76.57/73.42	74.01/71.75	**79.64/78.10**
L,A	78.68/76.00	76.23/70.58	75.35/66.10	78.33/75.23	73.80/71.77	**81.50/80.23**
A,V	76.68/73.32	72.84/65.54	75.64/67.65	76.55/71.50	74.92/67.26	**79.48/77.47**
L,A,V	78.89/76.31	76.01/70.58	75.72/67.85	78.46/75.87	73.93/71.66	**81.74/80.72**
0.1	77.99/75.29	75.64/70.02	75.51/67.32	78.09/75.30	74.49/72.01	**81.69/80.51**
0.2	77.83/75.20	75.67/69.88	75.19/66.88	78.28/75.13	74.28/71.57	**80.44/79.09**
0.3	77.56/74.54	75.27/69.22	75.16/66.56	77.88/74.45	74.47/71.04	**80.49/79.01**
0.4	76.84/73.72	74.39/68.46	75.40/66.40	76.44/72.52	74.87/71.42	**80.68/78.98**
0.5	76.20/72.69	75.05/68.70	74.95/65.68	76.97/72.60	74.57/70.32	**78.81/76.56**
0.6	75.91/71.92	75.00/68.46	75.13/65.46	76.23/70.80	74.25/69.33	**78.60/76.12**
0.7	75.99/71.66	74.44/67.58	74.84/65.04	75.21/69.63	75.00/69.78	**78.14/75.42**
Avg.	76.82/73.28	75.05/68.78	75.21/66.15	76.90/72.54	74.47/70.13	**79.69/77.65**

4.3 Ablation Studies

To evaluate the contribution of each component in MCUR, we perform ablation studies, with results shown in Table 2. Specifically, we assess the impact of four loss terms: the contrastive learning loss ($\mathcal{L}_{CL}^{s}$), the uncertainty term ($\mathcal{L}^{Uncer}$), the logits distillation loss ($\mathcal{L}^{Logits}$), and the MSE loss ($\mathcal{L}^{MSE}$), across 14 missing modality settings. Removing any of these components leads to an average performance drop of over 0.88 in ACC and 1.0 in F1. The most critical components are the $\mathcal{L}_{CL}^{s}$ and $\mathcal{L}^{Uncer}$. In particular, excluding the uncertainty loss causes substantial performance declines: 3.89/3.47 on MOSI, 1.29/2.38 on MOSEI, and 3.12/2.5 on IEMOCAP, respectively. We also find that the impact of each

Table 2: Ablation results of MCUR components. Averaged performance is reported.

	MOSI	MOSEI	IEMOCAP	MOSEI		IEMOCAP	
Models	ACC/F1	ACC/F1	ACC/F1	Brier	NLL	Brier	NLL
Full	**74.5/74.0**	**77.2/76.8**	**79.7/77.7**	**0.179**	**0.541**	**0.159**	**0.492**
w/o $\mathcal{L}_{CL}^{s}$	71.6/70.5	75.4/75.5	77.9/75.1	0.183	0.547	0.174	0.530
w/o $\mathcal{L}^{Uncer}$	70.6/70.5	75.9/74.4	76.6/75.2	0.182	0.545	0.186	0.556
w/o $\mathcal{L}^{Logits}$	71.5/71.4	76.3/75.8	78.8/76.5	0.182	0.545	0.175	0.532
w/o $\mathcal{L}^{MSE}$	72.7/71.8	76.3/75.8	78.4/75.5	0.181	0.543	0.173	0.525

component varies by datasets. In MOSI and MOSEI, removing the MSE loss leads to the smallest degradation, suggesting that representation alignment is less critical. In IEMOCAP, the smallest drop comes from removing the logits distillation loss.

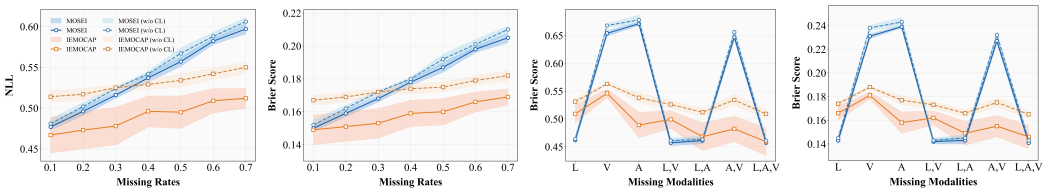

Figure 3: Uncertainty estimation under different missing modality conditions. (a) NLL under different MRs; (b) Brier under different MRs; (c) NLL under different modalitie; (d) Brier under different modalities. The Brier Score and NLL are used as evaluation metrics. Solid lines represent the full MCUR model, while dashed lines indicate the variant without contrastive learning. Orange curves show results on the IEMOCAP dataset, and blue curves correspond to the MOSEI dataset.

4.4 Quantitative Analysis

4.4.1 Effects of MCB-CL

To better understand the necessity of integrating combination into CL and to examine whether it mitigates intra-combination inconsistency, we conduct ablation studies on the MOSEI and IEMOCAP datasets. The prediction uncertainty is used as the measurement of inconsistency and the visualization of the uncertainty under different missing-modality conditions, with and without MCB-CL, is shown in Figure 3. Specifically, we adopt two standard uncertainty metrics: the Brier Score, which measures the calibration of probabilistic predictions, and the Negative Log-Likelihood (NLL), which quantifies the divergence between predicted probabilities and the ground truth.

Under the random missing protocol, uncertainty increases monotonically with the MR, regardless of whether MCB-CL is applied. On MOSEI, uncertainty grows rapidly but remains stable across runs, while on IEMOCAP, it increases more slowly yet exhibits larger variance. This difference stems from the task characteristics: MOSEI is a regression-based dataset with a thresholded decision boundary, whereas IEMOCAP involves discrete emotion categories that introduce greater intrinsic uncertainty. When MCB-CL is removed (dashed lines in the Figure 3), both datasets show a clear increase in uncertainty, with IEMOCAP displaying a larger rise than MOSEI. This demonstrates that MCB-CL enhances the robustness and consistency of representations. Interestingly, on IEMOCAP, incorporating MCB-CL increases the variance of uncertainty, but its maximum uncertainty remains lower than the minimum when MCB-CL is absent.

Under the fixed missing protocol, the overall trend remains similar, though the uncertainty distribution varies depending on which modality is missing. Consistent with the random setting, MOSEI exhibits smaller fluctuations, while IEMOCAP shows larger ones. However, the absolute uncertainty differs more markedly across missing-modality types. On MOSEI, maintaining the 'A', 'V', or 'A,V' modalities results in significantly higher uncertainty, re-

Table 3: Ablation of CL formulations in MCUR. Averaged 'ACC/F1' is reported.

Dataset	Vanilla	Traditional	1-instance	SupCon
MOSI	**74.46/73.99**	71.47/70.72	71.49/71.05	73.72/68.28
MOSEI	**77.23/76.80**	74.74/74.39	75.38/75.27	75.99/75.85

flecting the dominant predictive role of the 'L' modality (Zhuang et al., 2025). In contrast, on IEMOCAP, where different modalities contribute more evenly, the uncertainty variation across missing-modality conditions is less pronounced. Across both datasets, removing MCB-CL consistently increases uncertainty and amplifies intra-combination inconsistency. Notably, on MOSEI, MCB-CL leads to a larger reduction in uncertainty for modality combinations that originally exhibit higher inconsistency, such as 'V', 'A', and 'A,V'. This suggests that MCB-CL is particularly effective when modality representations are less aligned.

4.4.2 Uncertainty Analysis

To further investigate cross-combination inconsistency, we analyze the predictive uncertainty of different MCUR variants. As before, we use the Brier Score and NLL as evaluation metrics, and the results are summarized in Table 2. For a more comprehensive comparison, we report the average uncertainty across all 14 missing-modality settings, averaged over five random runs. We observe that the full MCUR model consistently exhibits the lowest uncertainty, while removing any component leads to a noticeable increase in uncertainty. On MOSEI, excluding the $\mathcal{L}_{CL}^{s}$ causes the largest rise in uncertainty, indicating that MCB-CL effectively strengthens representation alignment and improves stability, which aligns with the trends shown in Figure 3. Removing the uncertainty loss $\mathcal{L}^{Uncer}$ produces the second-largest degradation. On IEMOCAP, omitting $\mathcal{L}^{Uncer}$ results in the most significant increase in uncertainty, suggesting that uncertainty-aware weighting provides critical guidance.

4.4.3 Effects of Different CL Formulations

To investigate the efficacy of the proposed MCB-CL mechanism, we compare it against four contrastive learning variants: (1) Traditional CL (Radford et al., 2021), which treats each sample as a unique instance and all others as negatives, ignoring label information; (2) SupCon-CL (Khosla et al., 2020), which considers all samples within the same emotion category as positives; (3) Vanilla MCB-CL (Ours), which defines positives as samples sharing both the same emotion category and the identical modality-missing pattern. The average performance across 14 missing modality scenarios is summarized in Table 3; and (4) 1-instance MCB-CL, a simplified version that selects only a single positive sample per anchor. As shown in Table 3, Traditional CL yields the lowest performance, likely because it neglects semantic labels and inadvertently penalizes semantically related samples as negatives. While the 1-instance variant and SupCon-CL show notable improvements by incorporating label-level supervision, they both fall short of our proposed Vanilla MCB-CL. Specifically, our method achieves the superior results, underscoring that simultaneously accounting for category consistency and modality-specific patterns provides a more fine-grained and discriminative supervision signal for robust multimodal representation learning.

Additional results, including analysis of robustness, cross-dataset transfer, teacher bias on language modality, and more, are provided in the Appendix E.

5 Conclusion

In this paper, we presented MCUR, a unified framework that tackles MER from the perspective of representation consistency. MCUR integrates a MCB-CL mechanism to promote discriminative and consistent representations across heterogeneous modality configurations. In addition, the proposed SUGR module adaptively adjusts the supervision strength according to teacher-student prediction discrepancy, fostering more stable and reliable learning. Extensive experiments on three benchmark datasets demonstrate the effectiveness and robustness of MCUR under diverse modality conditions.

AI use statement

In this work, we used generative AI tools solely to improve the grammar, wording, and readability of the manuscript. We did not use generative AI tools for research ideation, hypothesis formulation, methodology or experiment design, mathematical derivations, implementation, data analysis, interpretation of results, or other research-related tasks. All AI-assisted edits were carefully reviewed and revised by the authors. The authors take full responsibility for the final content of this work, including all claims, analyses, and conclusions.

Ethics statement

This work relies exclusively on publicly available datasets released by prior publications. We did not collect new human-subject data, and no personally identifiable information was used. All datasets were used in accordance with their original licenses. No institutional ethics review was required. We adhere to the ICLR Code of Ethics.

Reproducibility statement

We provide the implementation and experimental details necessary to reproduce our results throughout the paper. The complete model architecture and formulations of MCB-CL and SUGR are described in Section 3 and Appendix A. The missing-modality protocols, training strategy, optimization settings, hyperparameter search spaces, and baseline reproduction procedures are detailed in Appendix B. Dataset statistics, feature extraction procedures, and evaluation metrics are provided in Appendix C, while the implementations and settings of all compared baselines are summarized in Appendix D. We evaluate MCUR on the publicly available MOSI, MOSEI, and IEMOCAP datasets under 14 missing-modality scenarios, including seven fixed and seven random missing settings. The main experiments use fixed random seeds, and additional experiments report results over five or ten random seeds with 95% confidence intervals to assess result stability. All experiments are conducted with PyTorch on a single NVIDIA RTX 3090 GPU, with the optimizer, learning rate, and model-specific hyperparameters explicitly specified in Appendix B. Additional ablations, robustness analyses, uncertainty evaluations, complexity measurements, and representation analyses are provided in Appendix E.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In International Conference on Learning Representations, 2022.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In 2016 IEEE winter conference on applications of computer vision (WACV), pp. 1–10. IEEE, 2016.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42:335–359, 2008.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In 2014 ieee international conference on acoustics, speech and signal processing (icassp), pp. 960–964. IEEE, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, 2019.

Siyi Du, Xinzhe Luo, Declan O'regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In International Conference on Learning Representations, volume 2026, pp. 116887–116913, 2026.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14314–14324, 2025a.

Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. Catch your emotion: Sharpening emotion perception in multimodal large language models. In Forty-second International Conference on Machine Learning, 2025b. URL https://openreview.net/forum?id=IYOksPHJKT.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26876–26885, 2024.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1726–1736, 2024.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM international conference on multimedia, pp. 1122–1131, 2020.

Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 30183–30193, June 2026.

Geoffrey Hinton. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

Ruohong Huan, Guowei Zhong, Peng Chen, and Ronghua Liang. Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. IEEE Transactions on Multimedia, 26:5753–5768, 2023.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In International conference on machine learning, pp. 4651–4664. PMLR, 2021.

Haoyu Jiang, Xiaoliang Chen, Duoqian Miao, Xiaolin Qin, Xianyong Li, and Yajun Du. Cica: Coupling confidence-aware pretraining with confidence-informed attention for robust multimodal sentiment analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 37842–37851, 2026.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in neural information processing systems, 33:18661–18673, 2020.

Diederik P Kingma. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 10074–10082, 2024a.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, et al. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. Advances in Neural Information Processing Systems, 37:28515–28536, 2024b.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12458–12468, 2024c.

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.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. OV-MER: Towards open-vocabulary multimodal emotion recognition. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=Y8lfuSoqQz.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Russ Salakhutdinov. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. Transactions on Machine Learning Research, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=ttzypy3kT7.

Rui Liu, Haolin Zuo, Zheng Lian, Hongyu Yuan, and Qi Fan. Hardness-aware dynamic curriculum learning for robust multimodal emotion recognition with missing modalities. In Proceedings of the 33rd ACM International Conference on Multimedia, pp. 5755–5764, 2025.

Xiaohao Liu, Xiaobo Xia, Jiaheng Wei, Shuo Yang, Xiu Su, See-Kiong Ng, and Tat-Seng Chua. Calibrated multimodal representation learning with missing modalities. In Forty-third International Conference on Machine Learning, 2026. URL https://openreview.net/forum?id=7fI30TOBtA.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2018.

Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. IEEE Transactions on Multimedia, 25:4121–4134, 2022.

Bo Peng, Zhen Fang, Guangquan Zhang, and Jie Lu. Knowledge distillation with auxiliary variable. In International Conference on Machine Learning, pp. 40185–40199. PMLR, 2024.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543, 2014.

Xihang Qiu, Yuhao Fang, Qing Zhou, Bin Zhai, Jialong Hong, Wanpeng Zhang, Yao Lu, Ye Zhang, and Chun Li. Beyond missing modalities: Hypergraph conditioned diffusion for uncertainty-aware multimodal emotion recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22953–22963, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PMLR, 2021.

Yuming Shen, Ziyi Shen, Menghan Wang, Jie Qin, Philip Torr, and Ling Shao. You never cluster alone. Advances in Neural Information Processing Systems, 34:27734–27746, 2021.

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

Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. Mice: Mixture of contrastive experts for unsupervised image clustering. In International conference on learning representations, 2020.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 1365–1374, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Xiaotao Wang, Yiyang Fang, Wenke Huang, Bin Yang, Guancheng Wan, and Mang Ye. API: Adaptive prototype imputation for incomplete multimodal sentiment analysis. In Forty-third International Conference on Machine Learning, 2026a. URL https://openreview.net/forum?id=NW2wCSQiId.

Yan Wang, Fuyuan Cao, and Xingwang Zhao. Conflict-aware adaptive cross-reconstruction for multimodal sentiment analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15880–15889, 2026b.

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22025–22034, 2023a.

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

Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20039–20049, 2023.

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 The Twelfth International Conference on Learning Representations, 2024.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259, 2016.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2236–2246, 2018.

Ying Zeng, Sijie Mai, Wenjun Yan, and Haifeng Hu. Multimodal reaction: Information modulation for cross-modal representation learning. IEEE Transactions on Multimedia, 26:2178–2191, 2023.

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

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp. 11953–11962, 2022.

Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. Glomo: Global-local modal fusion for multimodal sentiment analysis. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 1800–1809, 2024.

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

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. Advances in Neural Information Processing Systems, 38:142100–142129, 2026a.

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 Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 2281–2289, 2026b.

Yan Zhuang, Yanru Zhang, Jiawen Deng, and Fuji Ren. Intra-sample and intra-modal enhancement for multimodal sentiment analysis with missing modalities. IEEE Transactions on Multimedia, 28:1847–1859, 2026c.

A Additional Model Details

In this section, we provide the details of the proposed MCUR. The teacher and student models share the same architecture. As shown in Figure 4, the teacher model consists of two main components: pre-processing and the backbone model, which we will explain in the following subsections.

A.1 Pre-processing

Given a multimodal dataset with N samples $\mathcal{D} = \{X_1, X_2, ..., X_m\}$, where $X_p \in \mathbb{R}^{t_p \times d_p}$ denotes the modality representations and $p \in \{1, 2, ..., m\}$ denotes the modality. Here d_p and t_p represent the dimensionality and sequence length for each modality, respectively. To standardize the dimensionality of the representations, we apply a 1D convolutional network with kernel size 3 to process the input data, as follows:

$$X_p^w = W_3^w(X_p) \tag{16}$$

where $X_p^w \in \mathbb{R}^{T \times D}$, D and T represent the common dimensionality and sequence length of all modalities, respectively. $W_{3\times3}^w$ is the trainable weights, $w \in \{t, s\}$ indicates whether the representation is from the teacher (t) or student model (s).

A.2 Backbone Model

Once we obtain the modality representations X_p^w, the backbone model processes these representations to produce the final fused features and prediction logits. The backbone uses a Perceiver model (Jaegle et al., 2021), which learns modality-specific representations by attending to input features using cross-modal transformer blocks.

In the cross-modal transformer block, the modality representations X_p^w serve as keys and values, while learnable prompts $\hat{P_p^w} \in \mathbb{R}^{l_p \times D}$ act as queries. The attention mechanism is defined as:

$$P_p^w = \text{Softmax}(\frac{\hat{P_p^w} W_{Q_e}^w (W_h^w)^\top (X_p^w)^\top}{\sqrt{d_k}}) X_p^w W_{V_h}^w, \tag{17}$$

where $W_{Q_e}^w, (W_h^w)^\top, W_{V_h}^w \in \mathbb{R}^{D \times D}$ are trainable weights, and l_p is the length of the prompt. After applying the cross-modal transformer, the resulting enhanced modality representation P_p^w is obtained by averaging the outputs from the final transformer layer.

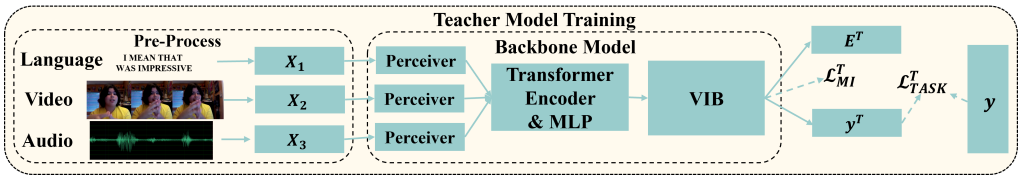

Figure 4: The structure of the teacher model.

Once we have the processed representations P_p^w from all m modalities, the representations P_p^w are then passed through a transformer encoder for interaction and then through a two-layer MLP for fusion, as follows:

$$E_f^w = W_2^w(W_1^w \mathrm{TE}^w([P_1^w, P_2^w, ..., P_m^w]) + b_1^w) + b_2^w, \tag{18}$$

where $\mathrm{TE}^w(\cdot)$ is the two-layer transformer encoder, E_f^w is the fused representation. W_1^w and W_2^w are trainable weights, b_1^w and b_2^w are biases.

Since the m modalities often contain redundant information, their combination can introduce noise that interferes with the effective transmission of information. To address this, the fused representation E_f^w is passed through a Variational Information Bottleneck (VIB) (Gao et al., 2024; Alemi et al., 2022; Tishby et al., 2000; Xiao et al., 2024; Mai et al., 2022) to reduce both redundancy and noise in the representations. VIB approximates the information bottleneck by learning an encoding E^w that is maximally informative about the target y, while being compressive in relation to the fused input E_f^w. This is achieved by optimizing the following objective function:

$$\mathcal{L}_{VIB}^w = I(E^w, y) - \beta I(E^w, E_f^w), \tag{19}$$

where E^w is the compressed representation of E_f^w and y is the ground truth. $I(\cdot)$ represents mutual information, and β is a Lagrange multiplier. In line with prior work (Gao et al., 2024), the task loss $\mathcal{L}_{TASK}$ is integrated with Kullback-Leibler divergence to measure the mutual information terms $I(E^w, y)$ and $I(E^w, E_f^w)$ as expressed by:

$$\mathcal{L}_{VIB}^w = \mathcal{L}_{TASK}^w(y^w, y) + \beta KL(p(e^w|e_f^w)||\mathcal{N}(0, \mathrm{I})), \tag{20}$$

where $p(e^w|e_f^w) \sim \mathcal{N}(\mu^w, (\sigma^w)^2\mathrm{I})$, and the mean μ^w and standard deviation σ^w are encoded using MLP layers through: $\mu^w = W_3^w e_f^w + b_3^w$ and $\sigma^w = W_4^w e_f^w + b_4^w$. $\mathcal{L}_{TASK}^w$ is introduced in Equations 12 and 13. Using the re-parameterization (Kingma, 2013) trick $e^w = \mu^w + \epsilon\sigma^w$, where $\epsilon \sim \mathcal{N}(0, \mathrm{I})$, the final representation e^w is used to produce the logits y^w through MLP layer:

$$y^w = W_5^w e^w + b_5^w. \tag{21}$$

Here W_3^w, W_4^w and W_5^w are trainable weights, b_3^w, b_4^w and b_5^w are biases in MLP. Note that all $w \in \{t, s\}$ indicates whether the elements are from the teacher (t) or student model (s).

B Additional Implementation Details

This section outlines the training setup and implementation details, including the experiment setups, teacher model training details, modality setup and training strategy, baseline reproduction details.

B.1 Missing Protocol Setups

In this work, we evaluate and compare different models under two missing modality protocols.

The fixed missing protocol is designed to simulate realistic situations where a specific modality is consistently unavailable (Zhuang et al., 2025; 2026a; Wei et al., 2023; Li et al., 2024c; Guo et al., 2024; Wang et al., 2023a). For instance, an audio sensor might fail, leaving only

textual and visual information available. Evaluating models under this setting allows us to analyze their robustness to modality-specific failures that may occur during deployment.

In contrast, the random missing protocol reflects more common conditions where missing data occurs unpredictably (Zhuang et al., 2026a; Wang et al., 2023a;b; Zhang et al., 2024). Factors such as poor signal quality or transmission errors can cause different modalities to be randomly missing across samples. We simulate this behavior by randomly dropping modalities in each sample to assess model performance under stochastic modality loss.

B.2 Teacher Model Training Details

The teacher model is trained using the following loss function:

$$\mathcal{L}_{all} = \mathcal{L}^t_{TASK} + \beta \mathcal{L}^t_{VIB}. \tag{22}$$

For student model, the total loss is defined as:

$$\mathcal{L}_{all} = \gamma \mathcal{L}^s_{CL} + \beta \mathcal{L}^s_{VIB} + \mathcal{L}^{MSE} + \zeta \mathcal{L}^{Sugr}, \tag{23}$$

where β, γ and ζ are hyper-parameters.

B.3 Training Configuration

We conducted all experiments using PyTorch with CUDA 11.5 on a single NVIDIA RTX 3090 GPU. For optimization, we used AdamW (Loshchilov & Hutter, 2018) on the MOSI and MOSEI datasets, and Adam (Kingma & Ba, 2014) for IEMOCAP. The teacher model was trained using complete modality inputs with a fixed β value of 0.01. We set the learning rate to 2e-5 and used a random seed of 5576 for both MOSI and MOSEI. For IEMOCAP, the learning rate was set to 1e-3 with a random seed of 1111. The student model was trained using the same random seeds as the teacher model. The μ_1 and μ_2 are set as 1.0 by default. During student model's training, we performed hyper-parameter tuning with the learning rate searched over {1e-5, 2e-5, 4e-5, 8e-4, 1e-3}. The trade-off hyper-parameters γ and ζ were searched over {0.1, 0.2, 0.5, 1.0, 2.0, 5.0} and {1.0, 2.0, 5.0, 10.0, 20.0, 100.0}, respectively. Finally, we trained MCUR on the MOSI dataset using a learning rate of 2e-5 with AdamW, $\gamma = 2.0$, $\beta = 0.01$, and $\zeta = 1.0$. For the MOSEI dataset, we used a learning rate of 2e-5, $\gamma = 0.2$, $\beta = 0.01$, and $\zeta = 5.0$, while for the IEMOCAP dataset, we set the learning rate to 8e-4 with Adam, $\alpha = 0.2$, $\beta = 0.01$, $\gamma = 0.1$, $\zeta = 100$ and $\tau = 0.2$.

B.4 Modality Setup and Training Strategy

All KD-based methods, including MCUR, share the same teacher model, which is pretrained using complete modality inputs. For reconstruction-based approaches, all modalities are also available during training to guide the recovery of missing ones. To simulate modality loss during training, we uniformly sample from seven predefined missing modality patterns, each with a probability of 1/7.

B.5 Baseline Reproduction Details

Most existing methods are trained under the setting in which separate models are learned for different missing modality configurations (Wang et al., 2023a;b; Zhang et al., 2024). While effective under controlled assumptions, this strategy often leads to limited robustness when modality availability varies in practice. To ensure a fair comparison and to better assess performance under conditions closer to real world scenarios, we reproduced all baseline methods using their official implementations whenever available.

For KD-based approaches such as CorrKD and MMANet, we adopted the same teacher model used in our MCUR framework to ensure consistency across methods. For approaches that require finetuning from pretrained checkpoints, including IMDer, we initialized the model using the official released checkpoints. All hyperparameters were set according to the configurations reported in the original papers and their corresponding open source code. When certain implementation details were not explicitly specified, such as the teacher model

configuration for knowledge distillation methods, we aligned these settings with those used in MCUR to maintain fairness and reproducibility. Since the official code for CorrKD is not publicly available, we reimplemented the method based on the descriptions provided in the original paper. We first verified that our implementation achieved results comparable to those reported under the original experimental settings, and then applied it to the missing modality scenarios considered in this work. In cases where reproduced results differed from reported numbers, we conducted careful re-tuning to ensure competitive performance.

In the reproducibility experiments presented in Table 12 in Section E.8, we only added or modified the code corresponding to the evaluated components. The hyper-parameter search space was kept consistent with that used in MCUR, ensuring a fair and reliable comparison: $\gamma \in \{0.1, 0.2, 0.5, 1.0, 2.0, 5.0\}$, $\zeta \in \{1.0, 2.0, 5.0, 10.0, 20.0, 100.0\}$.

B.6 Uncertainty Implementation Details

In this section, we introduce the way to measure the uncertainty in Table 2 and Figure 3. For classification tasks, uncertainty is computed as entropy over predicted class probabilities:

$$H(p) = -\sum_{c} p_c \log p_c, \tag{24}$$

where p_c is obtained via softmax function.

Although MOSEI is trained as a regression task, we convert outputs into probabilities for uncertainty evaluation: $p = \sigma(z)$ where where z is the model output and σ is the sigmoid function. Then the brier score is calculated as:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2. \tag{25}$$

This corresponds exactly to: (1) applying sigmoid to logits and (2) computing mean squared error with binary labels. For NLL score, we use the following formula:

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \sigma(z_i) + (1 - y_i)\log(1 - \sigma(z_i))\right]. \tag{26}$$

C Additional Details of Datasets and Evaluation Metrics

Datasets. In line with previous work (Guo et al., 2024; Li et al., 2024c), we evaluate the MCUR framework on three widely used MER datasets: MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and IEMOCAP (Busso et al., 2008). Both MOSI and MOSEI are regression tasks, consisting of user-generated monologue videos collected from social media platforms such as YouTube. MOSI contains 2,199 video clips, with 1,284 for training, 229 for validation, and 686 for testing. MOSEI includes 22,856 video clips, with 16,326 for training, 1,871 for validation, and 4,659 for testing. Each video clip is annotated with a emotion polarity score ranging from -3 (strong negative) to 3 (strong positive). IEMOCAP, on the other hand, is a multimodal dialogue dataset for classification task. It contains 4,453 video clips, with each clip labeled as one of the following four emotional classes: Neutral, Happy, Sad, and Angry.

Evaluation Metrics. Following common practice (Guo et al., 2024; Lian et al., 2023; Li et al., 2024c; Zhuang et al., 2025), for the MOSI and MOSEI datasets, we use accuracy (ACC) and F1 score computed for 'positive/negative' classification results as evaluation metrics. For IEMOCAP, we calculate the accuracy and F1 score for each of the four categories and then compute their weighted average for evaluation.

Feature Extraction. For the language modality, we use pre-trained BERT embeddings (Devlin et al., 2019) for the MOSI and MOSEI datasets, and GLoVe word embeddings (Pennington et al., 2014) for the IEMOCAP dataset. For the video modality, we extracted facial features using the Facet toolkit (Baltrušaitis et al., 2016) across all datasets. Audio features are extracted using the COVAREP toolkit (Degottex et al., 2014) for all datasets.

D Baselines

We compare MCUR with several state-of-the-art models, including both modality reconstruction-based and KD-based models, to provide a thorough evaluation.

D.1 KD-based Methods

CorrKD (Li et al., 2024c): CorrKD employs a sample-level contrastive distillation mechanism for knowledge transfer across samples, addressing missing semantics. It aligns feature distributions using category prototypes and optimizes emotion decision boundaries via response disentanglement and mutual information maximization.

MMANet (Wei et al., 2023): MMANet offers a three-component framework: a deployment network, a teacher network, and a regularization network. It introduces margin-aware distillation module, which adjusts sample contributions based on classification uncertainty, and modality-aware regularization module to refine weak modality representations.

D.2 Modality Reconstruction-based Methods

MPLMM (Guo et al., 2024): MPLMM utilizes three prompt types, including generative, missing-signal, and missing-type, to reconstruct missing modality features and enhance intra- and inter-modality learning. These prompts are used to reconstruct the missing modalities with available modalities.

IMDer Wang et al. (2023b): IMDer recovers missing data by mapping random noise to the distribution space of the missing modality. It leverages available modalities as prior conditions to refine the recovery process, ensuring semantic alignment.

LNLN Zhang et al. (2024): LNLN focuses on the language modality as the dominant emotion source. Its Dominant Modality Correction module and Dominant Modality-based Multimodal Learning modules enhance robustness against noise by improving dominant modality representations.

E Additional Results

In this section, we present additional experiments to further validate the effectiveness of MCUR.

First, we report the performance of MCUR under different missing-modality conditions on the MOSI dataset (Table 4). We then provide the mean and variance across multiple runs on all three datasets to assess result stability (Table 5). To further examine consistency, we report the mean and variance for each missing-modality setting over five random runs on the MOSEI and IEMOCAP datasets (Table 6). Next, we evaluate the robustness of MCUR in more realistic settings (Table 7) and analyze its time and space complexity across datasets (Table 8). We also report the mean and variance of prediction uncertainty for each missing-modality condition on all three datasets (Table 9).

In addition, we conduct a detailed hyper-parameter analysis to understand the sensitivity of MCUR (Tables 10 and 11). We further evaluate the model under more challenging conditions where noise and missing modalities occur simultaneously (Table 13). To better understand the behavior of MCUR, we analyze whether the student model inherits biases from the teacher model (Tables 14 and 15), and examine whether the learned representations are unified or dispersed. We also study the cross-sample generalization ability of MCUR (Table 16).

Finally, we visualize the training loss to illustrate the optimization process (Figure 5), and present visualizations of the learned representations (Figure 6).

Table 4: Performance comparison in MOSI. 'ACC/F1' is reported. Float values in 'Avail' are MR.

Avail	CorrKD	MPLMM	IMDer	LNLN	MMANet	MCUR
Results on MOSI						
L	83.38/83.22	82.77/82.59	83.23/83.30	**84.60/84.56**	**84.60**/84.52	83.99/83.97
V	53.81/53.10	42.23/25.07	42.23/25.07	42.23/25.07	55.79/54.02	**59.15/59.13**
A	43.45/27.93	42.84/26.64	42.23/25.07	42.23/25.07	55.18/54.02	**60.82/58.76**
L,V	83.54/83.35	83.23/83.03	82.93/83.00	84.60/84.56	82.32/82.02	**84.76/84.74**
L,A	83.38/83.20	82.77/82.60	84.15/84.17	84.60/84.56	82.16/81.80	**85.21/85.13**
A,V	56.86/56.93	42.99/26.71	42.23/25.07	42.23/25.07	**59.76/59.86**	58.54/56.78
L,A,V	83.54/83.34	83.23/83.04	83.38/83.42	84.60/84.56	81.86/81.28	**85.37/85.28**
0.1	81.25/81.19	80.95/80.93	81.25/81.36	80.64/80.73	80.49/80.02	**82.62/82.50**
0.2	77.90/77.89	74.85/75.00	75.76/75.88	74.85/74.98	78.96/78.76	**80.03/79.81**
0.3	73.63/73.76	71.80/71.88	75.30/75.36	71.04/70.91	74.54/74.55	**78.96/78.65**
0.4	71.04/71.21	65.09/64.64	67.23/66.69	69.51/69.18	71.19/71.22	**75.30/74.96**
0.5	67.84/67.94	62.96/62.05	57.32/54.04	61.43/59.61	**70.43/70.58**	70.27/69.76
0.6	61.59/61.17	58.38/55.68	58.69/55.16	58.69/55.94	66.62/66.79	**69.36/69.03**
0.7	60.21/59.24	53.20/49.01	59.30/55.98	57.47/54.17	63.87/63.98	**67.99/67.42**
Avg.	70.10/68.82	66.24/62.06	66.80/62.40	67.05/62.78	71.98/71.79	**74.46/73.99**

Table 5: Statistics of MCUR with 5 different random seeds (95% confidence intervals). Averaged 'ACC/F1' of all missing scenarios is reported.

	MOSI	MOSEI	IEMOCAP
ACC	73.50±0.57	77.05±0.18	79.56±0.10
F1	73.20±0.50	76.48±0.21	77.65±0.11

E.1 Results on MOSI Dataset

In this section, we analyze the performance of MCUR across three benchmark datasets. As shown in Table 4, MCUR achieves consistently better results on the MOSI dataset under most modality-missing conditions, improving the average ACC and F1 by 2.48 and 2.20, respectively.

Although randomness introduces stochasticity that causes some performance variation across datasets (Table 5), MCUR still delivers strong results overall. To better understand the effect of randomness, we report the mean and variance of prediction results under 14 different modality-missing scenarios on the MOSEI and IEMOCAP datasets (Table 6). We observe that the fluctuations in F1 are generally higher than those in accuracy, but the average variation across all missing-modality settings is much smaller than that in any single condition. This finding suggests that MCUR can effectively handle diverse missing-modality situations and maintain consistent, robust performance.

E.2 Robustness under simulated stochastic missing and noisy conditions

To comprehensively assess the robustness of the MCUR framework in more challenging scenarios, we conducted additional experiments in this section. Since modality loss in real-life situations is inherently uncertain, we simulated this uncertainty by running the trained model ten times on the test set with different random seeds. MCUR's average performance with 95% confidence intervals across 14 modality missing conditions in three datasets is reported in Table 7.

Our findings indicate that MCUR effectively handles various modality missing situations. The results across all three datasets showed low variance, highlighting MCUR's stability and robustness. Specifically, in the MOSEI dataset, the fluctuations in 'ACC/F1' scores were less than 0.2, even under the most challenging conditions. Although the MOSI dataset demonstrated slightly higher fluctuations, the performance variation did not exceed 0.5. These results further demonstrate the superior robustness and reliability of MCUR.

Table 6: Statistics of MCUR on datasets with 95% confidence intervals.

Results on MOSEI							
	L	V	A	L,V	L,A	A,V	L,A,V
ACC	85.8±0.3	64.7±0.6	63.3±0.2	86.1±0.2	85.8±0.3	64.9±0.5	86.0±0.1
F1	85.7±0.3	64.0±0.7	60.2±1.1	85.9±0.2	85.7±0.3	64.3±0.4	85.9±0.1
MR	0.1	0.2	0.3	0.4	0.5	0.6	0.7
ACC	83.7±0.2	81.8±0.4	79.6±0.3	77.6±0.3	75.4±0.9	72.8±0.6	71.5±1.1
F1	83.5±0.2	81.5±0.4	79.3±0.4	77.2±0.3	74.9±1.0	72.1±0.6	70.6±0.9
Results on IEMOCAP							
	L	V	A	L,V	L,A	A,V	L,A,V
ACC	78.9±0.5	75.4±0.1	79.8±0.3	79.3±0.4	81.6±0.5	80.0±0.7	81.8±0.1
F1	77.3±0.5	70.5±0.5	77.3±0.5	77.8±0.4	80.4±0.5	78.1±0.7	80.8±0.2
MR	0.1	0.2	0.3	0.4	0.5	0.6	0.7
ACC	81.6±0.4	81.1±0.7	80.4±0.5	79.9±0.5	78.6±0.4	78.3±0.5	77.7±0.6
F1	80.5±0.5	79.4±0.3	79.0±0.6	78.3±0.6	76.6±0.5	76.0±0.6	75.2±0.7

Table 7: Robustness of MCUR under simulated stochastic missing and noisy conditions with 10 different random seeds (95% confidence intervals). Averaged 'ACC/F1' of all missing scenarios is reported.

	MOSI	MOSEI	IEMOCAP
ACC	74.46±0.21	77.16±0.08	79.64±0.11
F1	74.04±0.22	76.73±0.09	77.60±0.11

E.3 Robustness to Class and Modality-Pattern Imbalance

Since MCB-CL defines positive pairs using both emotion labels and modality combinations, its effectiveness could potentially degrade when the joint label-modality distribution becomes highly imbalanced. We therefore conduct additional stress tests on IEMOCAP by independently introducing class imbalance and long-tail modality-pattern distributions.

For class imbalance, samples are downsampled according to their frequency rank r, with retained ratios of $1/(r+1)$ and $1/(r+1)^{1.5}$ for the severe and extreme settings, respectively. For modality-pattern imbalance, we consider the seven non-empty modality combinations and assign severe and extreme sampling probabilities as {0.80, 0.08, 0.04, 0.03, 0.02, 0.02, 0.01} and {0.90, 0.04, 0.02, 0.015, 0.01, 0.01, 0.005}, respectively.

Under severe class imbalance, MCUR obtains 75.73±2.59/69.99±3.96 on the full-modality setting and 74.52±0.86/68.27±2.17 averaged over the 14 missing-modality conditions. Under extreme class imbalance, the corresponding results are 77.50±2.99/74.82±3.64 and 75.79±1.64/71.84±2.36. These results show that MCUR remains effective under substantially skewed training distributions.

We further evaluate the effect of long-tail modality-pattern prevalence. Under the extreme long-tail setting, F1 changes from 76.13 to 75.90 without class downsampling, from 68.03 to 67.07 under severe class imbalance, and from 71.42 to 70.35 under extreme class imbalance. The corresponding maximum additional degradation is 1.07 F1 points.

To further examine whether contrastive supervision becomes sparse, we measure the joint label-modality distribution under the most skewed setting. The joint Gini coefficient reaches 0.856, while the effective number of joint cells decreases to 2.29 out of 28 possible cells. Nevertheless, the average number of valid same-label and same-pattern positives remains 10.00 per anchor, and only 9.55% of anchors have no valid positive in a mini-batch. For anchors without valid positives, the contrastive term is skipped while the task and SUGR losses remain active. These results suggest that MCB-CL remains operational even under highly asymmetric label-modality distributions.

Table 8: Complexity of the MCUR.

Datasets	#params	GPU Memory	Time
MOSI	112,627,841	5,542.6MB	686.5s
MOSEI	116,421,953	6,234.6MB	7,660.1s
IEMOCAP	624,654	1,166.6MB	1,194.8s

E.4 Cross-Combination Modality Conflict Analysis

We analyze the consistency of emotional predictions across different modality combinations to characterize the degree of modality conflict in incomplete multimodal settings. On IEMOCAP, we consider the seven observed modality combinations and compare their pseudo-label predictions for 1,500 utterances.

The analysis reveals substantial disagreement across modality combinations. Specifically, 88.73% of the utterances exhibit at least one pseudo-label disagreement among the seven combinations, while only 11.27% remain consistent across all combinations. The mean pairwise exact disagreement rate reaches 50.31%, and 32.80% of the utterances fall into the high-conflict group. The disagreement is also highly dependent on the modality pair: text-only and vision-only observations disagree on 68.93% of the utterances, while audio-only and vision-only observations disagree on 63.27%. In contrast, the disagreement between text-audio and full-modality observations is substantially lower at 22.20%.

These results indicate that samples sharing the same emotion label can nevertheless provide substantially different emotional evidence under different modality combinations. Consequently, treating all same-label samples as positive pairs may introduce noisy cross-combination alignments and over-regularize the representation space. MCB-CL instead promotes consistency within the same observation condition, which is more appropriate for incomplete multimodal learning where the available evidence varies across modality combinations.

E.5 Computation Overhead

To assess the computational efficiency of MCUR, we trained the model for 100 epochs on an NVIDIA RTX 3090 GPU and recorded its number of parameters, peak GPU memory usage, and training time across the three datasets (Table 8).

We found that MCUR requires the largest number of parameters and computational resources on the MOSEI dataset. However, even in this case, the peak GPU memory usage was only 6.2 GB, allowing MCUR to run comfortably on consumer-grade GPUs. Training on MOSEI completed in approximately two hours, demonstrating the practicality and efficiency of the proposed framework.

E.6 Additional Uncertainty Analysis

In this section, we present the mean and variance of prediction uncertainty for each missing-modality condition across all three datasets in Table 9, corresponding to the complete results of Figure 3.

Under the random missing protocol, uncertainty increases monotonically with the MR. On MOSEI, uncertainty rises quickly but remains stable across different runs, while on IEMOCAP, it grows more gradually yet exhibits higher variance. Under the fixed missing protocol, similar trends are observed, though the uncertainty distribution depends on which modality is missing. Consistent with the random setting, MOSEI shows smaller fluctuations, whereas IEMOCAP displays larger ones. The absolute uncertainty also varies more significantly across missing-modality types. On MOSEI, retaining the ‘A’, ‘V’, or ‘A,V’ modalities leads to noticeably higher uncertainty, reflecting the dominant predictive role of the ‘L’. In contrast, on IEMOCAP, where different modalities contribute more evenly, the variation in uncertainty across missing-modality conditions is less pronounced.

Table 9: Statistics of prediction uncertainty on datasets with 95% confidence intervals.

Results on MOSI					
	L	V	A	L,V	L,A
Brier	0.154±0.030	0.254±0.003	0.249±0.001	0.155±0.030	0.153±0.030
NLL	0.482±0.067	0.701±0.005	0.691±0.001	0.484±0.066	0.480±0.068
MR	0.1	0.2	0.3	0.4	0.5
Brier	0.163±0.026	0.174±0.023	0.183±0.020	0.193±0.016	0.203±0.016
NLL	0.502±0.057	0.526±0.050	0.546±0.045	0.569±0.035	0.589±0.036
MR	0.6	0.7	A,V	L,A,V	
Brier	0.213±0.012	0.219±0.012	0.250±0.001	0.154±0.030	
NLL	0.611±0.027	0.625±0.026	0.693±0.002	0.481±0.067	
Results on MOSEI					
	L	V	A	L,V	L,A
Brier	0.143±0.003	0.231±0.002	0.239±0.001	0.142±0.002	0.143±0.003
NLL	0.462±0.007	0.654±0.005	0.671±0.003	0.457±0.005	0.461±0.006
MR	0.1	0.2	0.3	0.4	0.5
Brier	0.150±0.002	0.159±0.002	0.168±0.002	0.178±0.002	0.187±0.003
NLL	0.477±0.004	0.496±0.005	0.516±0.004	0.537±0.005	0.557±0.006
MR	0.6	0.7	A,V	L,A,V	
Brier	0.198±0.002	0.205±0.003	0.227±0.002	0.141±0.002	
NLL	0.582±0.004	0.597±0.006	0.647±0.005	0.457±0.004	
Results on IEMOCAP					
	L	V	A	L,V	L,A
Brier	0.166±0.006	0.181±0.003	0.158±0.009	0.162±0.006	0.149±0.010
NLL	0.509±0.016	0.546±0.008	0.489±0.023	0.499±0.016	0.468±0.025
MR	0.1	0.2	0.3	0.4	0.5
Brier	0.149±0.009	0.151±0.009	0.153±0.009	0.159±0.008	0.160±0.008
NLL	0.467±0.022	0.473±0.023	0.478±0.023	0.493±0.019	0.495±0.020
MR	0.6	0.7	A,V	L,A,V	
Brier	0.166±0.006	0.169±0.005	0.155±0.009	0.146±0.010	
NLL	0.509±0.015	0.518±0.012	0.482±0.023	0.459±0.025	

E.7 Hyper-parameter Analysis

In this section, we conduct sensitivity analyses on four key hyper-parameters, namely γ, ζ, μ_1, and μ_2, on MOSEI and IEMOCAP. The average performance over 14 missing modality settings is reported in Tables 10 and 11.

Table 10: Performance of different γ and ζ values. Averaged 'ACC/F1' is reported.

γ	MOSEI	IEMOCAP	ζ	MOSEI	IEMOCAP
0.1	76.6/76.1	79.7/77.7	1	76.7/76.2	78.8/76.5
0.2	77.2/76.8	79.3/77.2	2	77.0/76.3	78.8/76.6
0.5	76.9/76.1	79.0/76.5	5	77.2/76.8	78.8/76.5
1	76.6/76.0	79.0/76.1	10	76.5/76.0	78.9/76.6
2	75.9/75.8	78.9/75.3	20	76.3/76.0	79.2/77.1
5	75.8/75.5	78.9/75.2	50	76.2/75.7	79.4/77.0
10	-	-	100	75.8/75.7	79.7/77.7

Sensitivity of γ and ζ. Table 10 shows that MCUR performs best with relatively small values of γ. On MOSEI, the best result is obtained at $\gamma = 0.2$, with an ACC/F1 of 77.2/76.8, while on IEMOCAP the best result appears at $\gamma = 0.1$, with 79.7/77.7. As γ increases, performance gradually declines on both datasets. This indicates that the alignment objective is helpful, but it should remain a supporting term rather than dominate the optimization. A moderate value of γ therefore provides the best balance between representation alignment and emotion classification.

The effect of ζ is more stable on MOSEI but more pronounced on IEMOCAP. On MOSEI, the results stay within a narrow range, with the best performance at $\zeta = 5$. On IEMOCAP, larger values of ζ generally bring better results, and the best performance is achieved at $\zeta = 100$, with an ACC/F1 of 79.7/77.7. This suggests that stronger weighting of the SUGR module is especially helpful when the data are more sensitive to uncertainty caused by missing modalities. Overall, MCUR remains stable across a wide range of ζ values, while still benefiting from a larger value on the more challenging dataset.

Taken together, these results show that MCUR is not highly sensitive to the exact choice of γ and ζ. The performance changes smoothly rather than sharply, which indicates that both the contrastive objective and the uncertainty guided regularization provide reliable guidance under different settings.

Sensitivity of μ_1 and μ_2. As shown in Table 11, performance consistently improves as μ_1 increases from 0.2 to 1.0, where the best results are achieved on both datasets. Importantly, the neighboring settings $\mu_1 = 0.8$ and $\mu_1 = 1.2$ remain highly competitive, indicating that the model is robust to moderate variations of this parameter. Compared with $\mu_1 = 1.0$, a smaller value ($\mu_1 = 0.8$) slightly weakens semantic supervision and leads to a minor performance drop, but helps retain more modality-related information. In contrast, a larger value ($\mu_1 = 1.2$) strengthens category supervision and maintains comparable accuracy, while slightly degrading F1, suggesting a reduced benefit from modality-aware alignment. Overall, $\mu_1 = 1.0$ provides the best tradeoff, while nearby values confirm stable behavior around this optimum.

A similar trend is observed for μ_2, where the best performance is obtained at $\mu_2 = 1.0$ on both datasets. The results at $\mu_2 = 0.8$ and $\mu_2 = 1.2$ remain close to the optimum, further demonstrating the robustness of MCUR. Specifically, a smaller value ($\mu_2 = 0.8$) relaxes the regularization and preserves more modality-specific details, at the cost of a slight performance decrease. A larger value ($\mu_2 = 1.2$) enforces stronger suppression of redundant information, which remains effective on IEMOCAP but leads to a mild degradation on MOSEI, indicating that excessive regularization may remove useful cues. These observations suggest that $\mu_2 = 1.0$ achieves a balanced tradeoff, while the model maintains stable performance within a reasonable range around this setting.

Table 11: Performance of different μ_1 and μ_2 values. Averaged 'ACC/F1' is reported.

μ_1	MOSEI	IEMOCAP	μ_2	MOSEI	IEMOCAP
0.2	76.43/76.21	78.75/76.28	0.2	76.77/76.12	79.27/77.07
0.4	76.90/76.28	78.85/76.33	0.4	76.89/76.19	79.34/77.49
0.6	76.38/76.23	78.86/76.31	0.6	76.96/76.34	79.01/77.15
0.8	77.01/76.43	79.46/77.63	0.8	77.06/76.42	79.64/77.65
1.0	77.23/76.80	79.69/77.65	1.0	77.23/76.80	79.69/77.65
1.2	77.03/76.42	79.70/77.58	1.2	76.77/76.38	79.44/77.54

E.8 Generalizability of the MCUR Components

To evaluate the generalizability of MCUR's core components, we integrated MCB-CL and SUGR into two representative architectures, CorrKD and MMANet. Evaluation across 14 missing modality configurations in Table 12 reveals consistent performance gains regardless of the base model. Specifically, incorporating either module into CorrKD yielded improvements of up to 1.53% on IEMOCAP. These enhancements were even more pronounced for MMANet, where MCB-CL and SUGR boosted performance by up to 4.14% and 4.54% on IEMOCAP, respectively, while also providing steady gains on MOSEI. These results underscore the 'plug-and-play' capability of our modules and their effectiveness in bolstering model robustness against modality incompleteness.

E.9 Results on Noisy and Incomplete Settings

Although MCUR is developed for missing-modality settings, its uncertainty-based supervision can also be evaluated under simultaneous missing and feature-level noisy inputs. To

Table 12: Generalizability of the MCUR components on KD baselines. Averaged 'ACC/F1' is reported.

	MOSEI			IEMOCAP		
Models	Vanilla	+MCB-CL	+SUGR	Vanilla	+MCB-CL	+SUGR
CorrKD	76.5/73.9	76.6/74.9	**76.8/75.3**	76.8/73.3	**78.1/74.8**	77.9/74.8
MMANet	76.2/74.1	**76.5/75.3**	76.3/75.1	74.5/70.1	**78.1**/74.3	78.1/**74.7**

verify this, we extend our experiments to scenarios where both missing modalities and noisy inputs are present.

Specifically, we inject Gaussian noise into each representation as in prior research Zhuang et al. (2026b) through:

$$embed = embed + intensity \cdot \mathcal{N}(0, \mathrm{I}), \tag{27}$$

where *intensity* controls the noise level.

The results are shown in Table 13. As noise increases, all models degrade, but MCUR consistently remains the most robust. Under strong noise, e.g, $intensity = 20$, MCUR achieves an F1 of 74.0, demonstrating clear noise resistance.

Table 13: Performance of MCUR on noisy datasets. Averaged 'ACC/F1' is reported.

Model	*intensity*	ACC/F1	*intensity*	ACC/F1	*intensity*	ACC/F1
MMANet	5	75.0/74.3	10	74.7/73.4	20	74.4/72.2
LNLN	5	75.7/71.4	10	75.1/71.1	20	73.7/69.7
MCUR	5	**75.9/74.6**	10	**75.4/74.4**	20	**74.6/74.0**

E.10 Fragmented Representation Analysis

A potential concern is that conditioning the contrastive objective on modality combinations in MCB-CL may lead to fragmented representations, rather than a unified multimodal space. However, strict modality invariance, i.e., forcing all modality combinations to collapse into a single representation, is not well-suited to missing-modality settings. Different modality subsets (e.g., text-only vs. audio-visual) provide inherently heterogeneous and incomplete observations. Enforcing full alignment across such conditions can obscure modality-specific cues and introduce dominance effects from stronger modalities.

This limitation is also reflected in prior approaches that aim to learn unified representations. Despite this objective, the resulting embeddings are often biased toward the language modality. As a consequence, their performance degrades substantially when text is unavailable (Table 1), suggesting that the learned representations are not truly modality-invariant in practice, but instead rely heavily on language signals.

In contrast, MCB-CL is designed to balance semantic consistency and modality awareness. It encourages representations to remain aligned at the semantic level across modality conditions, while preserving informative differences induced by varying modality subsets. This leads to a structured embedding space where global relationships are maintained without enforcing unnecessary collapse.

To further examine this property, we analyze the cosine similarity between teacher (full-modality) and student (missing-modality) representations. When text is present, the similarity is approximately 0.69. Under missing-text conditions (e.g., audio-visual only), similarity remains consistently high, ranging from 0.63 to 0.75, with all results statistically significant ($p < 0.001$). In addition, we observe reduced intra-class variance, indicating that the representations become more compact while preserving their global organization. These observations suggest that MCB-CL does not fragment the embedding space. Instead, it promotes alignment with the full-modality semantic structure while improving representation compactness.

Table 14: Effect of teacher quality and bias on student performance on MOSEI.

Teacher	T Avg F1	S Avg F1	ΔAvg	T Worst F1	S Worst F1	ΔWorst	BiasAmp
noise	71.35	75.34	**+3.99**	48.51	57.14	**+8.63**	0.6350
mr_30	72.23	75.50	**+3.26**	48.51	59.38	**+10.87**	0.5810
high_l	72.28	75.66	**+3.38**	48.51	59.47	**+10.96**	0.6133
high_l_2	72.48	75.31	**+2.83**	48.51	58.38	**+9.87**	0.6521

This behavior is also reflected in downstream performance. When text is available, the teacher model achieves slightly higher performance than the student (nearly 0.26% F1), which is expected given its access to complete information. In contrast, under audio-visual-only conditions, the student outperforms the teacher by a substantial margin (15.8% F1), indicating improved utilization of non-language modalities.

Overall, both representation-level analysis and predictive performance consistently indicate that MCB-CL enhances robustness to missing modalities while maintaining coherent semantic structure, rather than inducing fragmented representations.

E.11 Analysis of Transfer Bias from Teacher Model

As discussed in Section E.10, MCUR tends to rely more heavily on textual information in both the teacher and student models. We therefore examine how teacher quality and teacher bias affect student learning under missing-modality conditions.

To this end, we conduct controlled experiments in which only the teacher distribution is perturbed, while the student architecture, optimizer, training schedule, and all MCUR losses remain unchanged. Evaluation follows the same 14 missing-modality scenarios used in the main paper, including 7 fixed and 7 random settings.

To quantify language dominance, we define two auxiliary metrics. Let

$$S = \{L, V, A, L+V, L+A, A+V, L+A+V\}. \tag{28}$$

We define:

$$\text{LangDomGap} = F1(L) - \frac{1}{3}\big(F1(A) + F1(V) + F1(A+V)\big), \tag{29}$$

which measures the performance advantage of language over non-language settings. Larger values indicate stronger language dominance.

We also define:

$$\text{BiasAmp} = \frac{\text{LangDomGap}_{\text{student}}}{\text{LangDomGap}_{\text{teacher}} + 1\text{e}{-8}}, \tag{30}$$

where values below 1 indicate that the student attenuates the teacher's bias.

We consider four teacher settings with identical model architecture but different training perturbations: (1) noise, which injects Gaussian noise into teacher embeddings (*tense* = 5); (2) mr_30, which applies 30% random modality dropout during teacher training; (3) high_l, which increases the sampling probability of the language-only setting to $P(L) = 4/7$ and assigns 1/14 to each of the remaining six combinations; and (4) high_l_2, which biases sampling toward language-related combinations with probabilities $[1/7, 1/14, 1/14, 3/14, 3/14, 1/14, 3/14]$ over $[L, A, V, L+A, L+V, A+V, L+A+V]$. The results are shown in Table 14.

Across all four teacher settings, MCUR consistently improves average F1 by 2.8 to 4.0 points and worst-case F1 by 8.6 to 11.0 points. The gains are largest in non-language settings. For example, under noise, the student improves by 8.63 on V, 14.96 on A, and 15.86 on $A+V$; under mr_30, the corresponding gains are 10.86, 15.48, and 15.95. Similar patterns hold for high_l and high_l_2, indicating that the performance gains mainly come from better use of non-language modalities rather than increased reliance on text.

This trend is also reflected in the bias statistics. LangDomGap decreases from 0.3558-0.3732 for the teachers to 0.2164-0.2434 for the students, corresponding to a reduction of 34.8%-41.9%. BiasAmp remains below 1 in all cases, showing that the student attenuates rather

Table 15: Calibration comparison on non-language settings (A and $A{+}V$). Lower is better.

		A		A+V	
Teacher	Model	Brier ↓	NLL ↓	Brier ↓	NLL ↓
add_noise	Teacher	0.2702	0.7632	0.2712	0.7672
	Student	**0.2243**	**0.6398**	**0.2214**	**0.6337**
mr_30	Teacher	0.2331	0.6590	0.2305	0.6532
	Student	**0.2286**	**0.6495**	**0.2265**	**0.6449**
high_l	Teacher	0.2331	0.6590	0.2305	0.6533
	Student	**0.2289**	**0.6500**	**0.2286**	**0.6493**
high_l_2	Teacher	0.2331	0.6589	0.2305	0.6532
	Student	**0.2276**	**0.6473**	**0.2255**	**0.6429**

than amplifies the teacher's language bias. Similar reductions are also observed for ACC, with BiasAmp(ACC) consistently below 1.

We further analyze calibration on the non-language settings (A and $A + V$) in Table 15. Under the add_noise setting, MCUR yields substantial improvements in both Brier score and NLL. For example, on A and $A{+}V$, Brier decreases from 0.2702/0.2712 to 0.2243/0.2214, and NLL decreases from 0.7632/0.7672 to 0.6398/0.6337, respectively. Under the remaining teacher settings (mr_30, high_l, and high_l_2), MCUR consistently reduces both Brier score and NLL across A and $A + V$. Although the magnitude of improvement is smaller compared to the noisy setting, the trend is stable across all perturbations.

These results indicate that MCUR not only improves predictive performance but also produces more reliable probability estimates in non-language scenarios. This is particularly important in missing-modality settings, where uncertainty is inherently higher. Together with the performance and bias analyses, the calibration results further support that the observed gains arise from improved robustness rather than over-reliance on language signals.

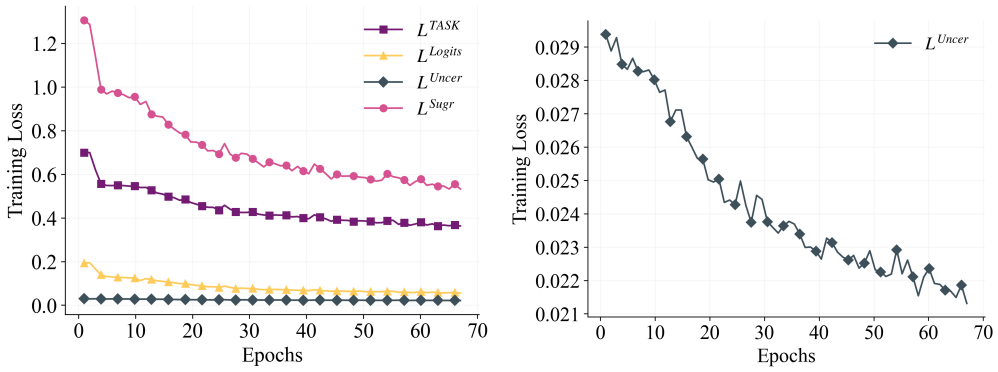

Figure 5: Analysis on the training convergence on IEMOCAP. (a) L^{Sugr} Components; (b) L^{Uncer}.

E.12 Cross-Dataset Generalization

In this section, we test the generalization ability of MCUR. Specifically, we train MCUR on MOSEI dataset and test on MOSI dataset for that these two datasets are all regression tasks. We use a strict transfer setting with no fine-tuning and only simple zero-padding to match the input shape. The results are shown in Table 16.

MCUR preserves teacher-level accuracy while improving F1 by 4.59 over the teacher and 2.87 over prior KD students, indicating that the learned consistency generalizes beyond a single dataset.

Table 16: Cross-dataset generalization results. Averaged 'ACC/F1' is reported.

Model	Student
Shared Teacher Model	**76.14**/69.95
CorrKD	62.76/48.74
MMANet	74.88/71.68
MCUR	75.46/**74.55**

E.13 Loss Analysis

To better understand the internal behavior of the SUGR module, we analyze how its loss components evolve during training. Specifically, we visualize the training dynamics of each loss on the IEMOCAP dataset, as shown in Figure 5. Here, all loss components represent the average values across all representations at each epoch.

Our observations are as follows: (1) The components of L^{Sugr} consistently decrease over epochs, indicating that each part of the SUGR module is well-optimized and contributes effectively to training. (2) In Figure 5(b), we examine L^{Uncer}, which captures uncertainty differences. While it shows minor fluctuations, the overall trend is downward. These fluctuations may result from inconsistencies in the missing modalities across epochs, causing stronger modalities to be lost at different rates in different batches. These results validate the design of the SUGR module and demonstrate its stable behavior during training under missing modality conditions.

E.14 Comparison with Recent Missing-Modality Methods

To further assess MCUR against recent missing-modality MER methods, we additionally compare it with TMDC (Zhuang et al., 2026b) and HARDY-MER (Liu et al., 2025) under the same IEMOCAP feature extraction pipeline, data split, missing-modality protocol, and evaluation metrics used in the main experiments. As shown in Table 20, MCUR achieves 79.79/77.65 in ACC/F1, compared with 75.56/74.18 for HARDY-MER and 74.61/64.95 for TMDC.

The compared methods follow different technical paradigms. TMDC focuses on modality denoising and complementation, while HARDY-MER adopts hardness-aware dynamic curriculum learning. In contrast, MCUR does not reconstruct or complement missing modalities, and does not introduce curriculum scheduling. Instead, it regulates the representation space and teacher supervision according to the observed modality combination. The comparison therefore complements the main experiments by demonstrating that MCUR remains competitive with more recent approaches under a matched evaluation protocol.

E.15 Visualization of the Representations

To better understand the impact of the MCB-CL and SUGR modules in MCUR, we visualized the representations tested on the IEMOCAP dataset with $MR = 0.7$, where each sample contained only one modality. The representations are obtained from MCUR, MCUR without MCB-CL, and MCUR without the SUGR, as shown in Figure 6.

Our observations reveal that removing either module leads to more scattered representations, with the absence of the SUGR module resulting in the most pronounced dispersion. This suggests that without the SUGR module, the uncertainty introduced by missing modalities is not effectively accounted for, which disrupts the clustering of representations. On the other hand, removing the MCB-CL module also leads to a less cohesive representation, likely because it eliminates the shared label information across combinations of the same modalities. When both modules are included, the resulting clusters are much more defined, highlighting the effectiveness of both modules in MCUR.

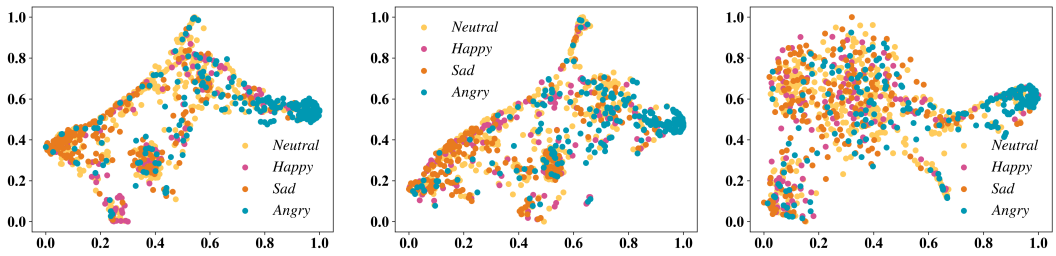

Figure 6: Visualization of MCUR and its variants on IEMOCAP dataset with MR=0.7. (a) w/o MCB-CL; (b) w/o SUGR; (c) Full MCUR.

F Discussion of Limitations and Societal Impacts

F.1 Discussion of Limitation

Although MCUR demonstrates strong performance across various scenarios, several limitations remain. First, the dataset constraint poses a challenge. Due to the lack of datasets that truly reflect real-world multimodal conditions, our experiments mirror prior work by simulating missing modalities by randomly discarding modality representations. This approximation may not fully capture the complexity and unpredictability of real-world data loss. Second, from a model design perspective, our approach, like other KD-based methods, requires training an additional teacher model. This increases computational cost and complicates deployment in resource-limited environments. Finally, the model's sensitivity to data variation is worth noting. Differences in background and scene characteristics across datasets, for instance, monologue settings in MOSI and MOSEI versus multi-person dialogues in IEMOCAP, as well as randomness in modality dropout, can lead to fluctuations in performance. In future work, we aim to explore models that can better handle real-world conditions with missing modalities, achieving greater efficiency, stability, and robustness without relying on simulated modality loss or additional teacher networks.

F.2 Discussion of Societal Impacts

Recognizing emotion from incomplete multimodal data improves the robustness and applicability of MER systems. However, this technology also poses potential societal risks. For instance, inferring emotions from partial, noisy or missing data may lead to inaccurate profiling, privacy breaches, or unintended surveillance without user consent.