UCMIB-PNS: Balancing Sufficiency and Necessity With Probabilistic Causality and Cross-Modal Uncertainty in Multimodal Sentiment Analysis

Jili Chen, Yihua Zhong, Qionghao Huang, Changqin Huang, *Member, IEEE*, Fan Jiang, Xiaodi Huang, *Senior Member, IEEE*, and Xun Wang, *Member, IEEE*

***Abstract*—Multimodal sentiment analysis aims to accurately identify sentiment orientations by integrating information from multiple modalities such as text, audio, and video. However, a key challenge in multimodal fusion is effectively balancing the sufficiency and necessity of information across modalities. Traditional models often fail to qualify and capture this balance due to the presence of noise and redundant information in multimodal data, leading to suboptimal performance in sentiment analysis. To address this issue, we propose a novel multimodal sentiment analysis method called UCMIB-PNS, which is guided by information bottleneck and probabilistic causality. The method employs an Uncertain Cross-Modal Information Bottleneck (UCMIB) module to reduce redundant information within modalities and maximize discriminative information. The UCMIB utilizes codebooks to dynamically record the distributions of samples and employs random sampling to conduct uncertain modeling across different modalities. It integrates uncertainty-aware contrastive learning and KL divergence for dynamic comparison and compression of information from different modalities. Moreover, UCMIB-PNS uses differentiable Probability of Necessity and Sufficiency (PNS) estimators to estimate and re-weight the sufficiency and necessity of modalities by constructing several counterfactual scenarios through end-to-end learning. Experiments conducted on four publicly available multimodal sentiment analysis datasets demonstrate that UCMIB-PNS achieves optimal performance on both clean and noisy data. Extended experiments further validate the method's robustness under different types of noise.**

***Index Terms*—Causal inference, information bottleneck, multimodal fusion, multimodal sentiment analysis.**

I. Introduction

MULTIMODAL sentiment analysis (MSA) integrates text, audio, and video frames to detect emotions, providing a more comprehensive and accurate understanding than single-modality approaches. It has been widely applied in social media monitoring, customer service, advertising evaluation, and healthcare diagnostics [1], [2], [3], [4], [5], [6], [7]. Most MSA research concentrated on the integration of different modalities. In early work, tensor-based fusion techniques [8], [9], [10] were employed to achieve interactions among different modalities. To achieve more accurate sentiment insights, some studies employed cross-modal attention [11], [12], [13], [14], [15], [16], [17], [18] or Transformer-based models [19], [20], [21], [22], [23] to integrate different modalities, while others design self-supervised learning methods [24], [25], [26] to reduce the gap between modalities to facilitate modality alignment and fusion.

Although previous methods have achieved improvements in accuracy, they are based on the maximization of likelihood. Therefore, baselines tend to search for sufficient information in different modalities to minimize the loss function while neglecting necessary information. In MSA task, text is considered the dominant modality due to its rich semantics, while vision and audio typically serve as auxiliary modalities [11], [12], [13]. Thus, the information in text often exhibits higher sufficiency compared to that in vision and audio. As shown in Fig. 1, the sentiment can be discerned solely from the words "too" and "good" without needing vision or audio **(Sufficiency)**. However, without the frowning expression and the disdainful tone of voice, it is easy to misinterpret the emotion as positive, which means that information from visual and audio aspects is necessary **(Necessity)**. Furthermore, certain information, such

Received 25 April 2025; revised 1 September 2025; accepted 2 September 2025. Date of publication 8 September 2025; date of current version 10 March 2026. This work was supported in part by the National Key R&D Program of China under Grant 2024YFC3308200 and in part by the National Natural Science Foundation of China under Grant 62337001, Grant 62037001, and Grant 62207028. Recommended for acceptance by S. Zhao. *(Corresponding authors: Qionghao Huang; Changqin Huang.)*

Jili Chen and Qionghao Huang are with the Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua 321017, China (e-mail: irelia@zjnu.edu.cn; qhhuang@m.scnu.edu.cn).

Yihua Zhong is with the Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai 200062, China (e-mail: yiwer007@zjnu.edu.cn).

Changqin Huang is with the College of Education, Zhejiang University, Hangzhou 310027, China, and also with the Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua 321017, China (e-mail: cqhuang@zju.edu.cn).

Fan Jiang is with the School of Education Science, Guangdong Polytechnic Normal University, Guangzhou 510640, China (e-mail: fanjiang@zjnu.edu.cn).

Xiaodi Huang is with the School of Computing, Mathematics and Engineering, Charles Sturt University, Albury, NSW 2640, Australia (e-mail: xhuang@csu.edu.au).

Xun Wang is with the School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 314423, China (e-mail: wx@mail.zjgsu.edu.cn).

Our code is available at https://github.com/TheShy-Dream/UCMIB-PNS.

This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2025.3606964, provided by the authors.

Digital Object Identifier 10.1109/TAFFC.2025.3606964

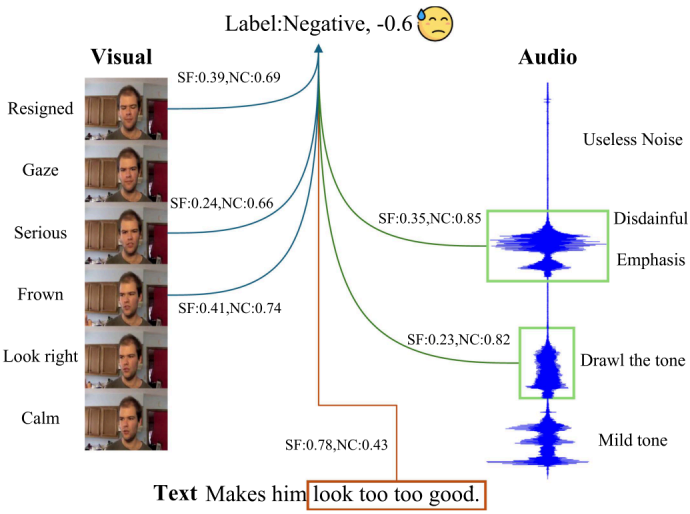


Fig. 1. Each modality contains distinct sentimental attributes, which encompass varying degrees of sufficiency (SF) and necessity (NC). Our model can accurately estimate the weights of sufficiency and necessity for different modalities. For example, "Serious SF:0.24, NC:0.66" indicates the serious expression has a sufficiency level of 0.24 and a necessity level of 0.66 for prediction. The text modality, due to its rich semantics, has a high degree of sufficiency. In contrast, other modalities contain a large amount of modality-specific information and thus have a high degree of necessity.

as ambiguous visual cues and background noise, is neither sufficient nor necessary and can adversely affect the model's predictions. Although many methods have been employed to preserve modality-specific information [25], [26], [27], [28], their model accuracy falls short of expectations due to the difficulty in balancing the sufficiency and necessity of multimodal features.

To address the aforementioned issues, one primary question is: *Q1: Can we quantify the notions of sufficiency and necessity, and reduce information that is redundant, being neither sufficient nor necessary?* The sufficiency refers to the ability of a feature to independently support accurate predictions or conclusions without requiring additional information. In contrast, necessity indicates that a feature is indispensable for achieving accurate results, even though it may not be sufficient on its own. Quantifying sufficiency and necessity in multimodal data is challenging due to the complex multimodal context, noise, and redundant information, which can obscure critical modality contributions and distort quantifications. When we quantify the sufficiency and necessity in multimodal data, it will provide a more solid foundation for addressing the second question: *Q2: How can we balance sufficiency and necessity to achieve more effective fusion?* Balancing sufficiency and necessity for effective fusion is challenging due to the dynamic interplay between modalities, where optimizing one may compromise the other. Besides, it is difficult to design models that can adaptively prioritize modalities based on context without introducing bias or information loss.

To address the two aforementioned problems, the basic idea is to prioritize features that are either sufficient for prediction on their own or necessary for the model's accuracy, as they cannot be omitted. This process is further refined by employing estimators of probability necessity and sufficiency (PNS), which re-weight different modalities to effectively select features that exhibit a high causal impact on the outcome. In light of this, we embrace the information bottleneck theory and causal inference and propose the **U**ncertain **C**ross-**M**odal **I**nformation **B**ottleneck and **P**robability of **N**ecessity and **S**ufficiency (**UCMIB-PNS**) guided MSA model, which reduces redundant information using cross-modal information bottoleneck and can perceive and balance the sufficiency and necessity of different modalities for multimodal sentiment analysis. To address *Q1*, we employ the **U**ncertain **C**ross-**M**odal **I**nformation **B**ottleneck **(UCMIB)** module to compress the multimodal representations, which reduces neither sufficient nor necessary information redundancy while preserving sentimental cues to the greatest extent. To enhance accuracy in noise perception, the UCMIB modules store the distribution of negative samples in dynamically updated codebooks and conduct contrastive learning based on uncertainty modeling. Then, guided by a counterfactual framework, we use causal inference to quantify the sufficiency and necessity of different modalities for the prediction outcomes through the probabilities of sufficient and necessary causes. To address *Q2*, we then reassign the attentive weights of different modalities according to the dynamic weights of sufficiency and necessity. By accurately controlling and weighting the sufficiency and necessity, the proposed UCMIB-PNS achieves the state-of-the-art results on four publicly available MSA datasets. Moreover, UCMIB-PNS demonstrates robustness against various noisy scenarios compared to its counterparts. Overall, the contributions of this paper can be summarized as follows:

- We propose a novel fusion method, UCMIB-PNS, which dynamically quantifies the sufficiency and necessity of modalities during the fusion process, thereby obtaining more useful modality representations.
- Based on the variational information bottleneck, we design the UCMIB module. The UCMIB module effectively mitigates information redundancy during multimodal fusion. Moreover, building on the counterfactual framework, we have developed a PNS estimator to quantify sufficiency and necessity. We also propose a reweighting method to balance these measures across different modalities.
- Our proposed UCMIB-PNS achieves optimal metrics on four publicly available datasets (CMU-MOSI [29], CMU-MOSEI [30], UR-FUNNY [31], and CH-SIMS [32]). Moreover, we have demonstrated that even under various noisy conditions, the performance of UCMIB-PNS consistently outperforms that of its counterparts.

II. Related Work

In this section, we have meticulously organized the related work of this paper, encompassing multimodal sentiment analysis, information bottleneck theory, and causal inference.

A. Multimodal Sentiment Analysis

Multimodal sentiment analysis is a technique that detects and interprets emotions by combining and analyzing information from various modalities like text, voice, and facial expressions.

Early multimodal sentiment analysis networks were based on tensor fusion networks [8], [9], [10], which predominantly employed tensor operations such as tensor concatenation, sequential processing, and low-rank fusion to integrate different modalities. Driven by the development of attention mechanisms and the widespread application of Transformers [33], Tsai et al. introduced cross-modal attention [34] to enhance interaction between different modalities. After that, more and more MSA studies [11], [12], [13], [14], [15], [16], [17] have adopted and refined cross-modal attention for the multimodal fusion. Other studies utilized self-supervised methods and auxiliary tasks [20], [24], [25], [26], [27], [35] to mitigate the gap between modalities, thereby facilitating modality alignment and fusion. Although they have achieved improvements in accuracy, their reliance on the maximum likelihood of different modalities for modality fusion overlooks the evaluation of the necessity of modal information. On the one hand, this leads to the neglect of modality-specific information, and on the other hand, it makes the model more vulnerable to noise interference. Recent studies [36], [37], [38] have pointed out that traditional multimodal methods are highly susceptible to noise and low-quality data, significantly limiting the applicability of multimodal models. Building upon prior work, this paper proposes a new modality fusion mechanism designed to reduce information redundancy while attempting to preserve discriminative features across modalities.

B. Information Bottleneck

The information bottleneck aims to extract the most relevant information from an input signal by balancing the trade-off between compression (minimizing irrelevant details) and preserving meaningful information. It focuses on retaining only the essential features that are useful for prediction or decision-making. The information bottleneck is widely applied in deep learning [39], [40], such as in model interpretability [41], [42], [43], [44], generalization [45], [46], [47] and compression [48], [49], [50], etc. In multimodal scenarios, variational information bottleneck (VIB) enhances modality alignment and fusion by balancing shared and complementary information across modalities, improving robustness to noise or missing data [51], [52]. Besides, VIB supports disentanglement of modality-invariant and modality-specific features, enabling better interpretability and generalization [37]. For example, Gao et al. [53] employed VIB to address redundancy issues in multimodal fusion, thereby enhancing the performance of multimodal learning. Cui et al. [54] utilized the information bottleneck to address the issues of modality-noise and modality-gap in multimodal entity and relation extraction. Inspired by their work, we design an uncertainty-based cross-modal information bottleneck to maximize the retention of useful information while reducing neither sufficient nor necessary information (noise), thereby optimizing the representation learning of modalities.

C. Causal Inference

In recent years, the integration of causal inference and deep learning has emerged as a promising research direction, addressing the limitations of traditional deep learning models that often rely solely on correlation rather than causation. This fusion aims to enhance model interpretability, generalizability, and robustness. Some work [55], [56], [57], [58] used front-door or back-door adjustment to reduce the bias of the model originating from the dataset and enhance the reliability of the model. In addition, some work [59], [60], [61], [62] employed a counterfactual framework to guide the model towards unbiased outputs from the perspective of causal effect estimation. Causal inference-based representation learning was widely used for feature selection and invariant learning [63], [64], [65], [66]. Causal inference is an important tool for measuring feature importance. Yang et al. [67] employed causal probability to evaluate the importance of image features, addressing the out-of-distribution generalization problem. Inspired by prior work, we extend the probability of causality to the multimodal scenario to dynamically assess the importance of different modalities, thereby facilitating more reliable modality fusion.

III. Methodology

In this section, we first introduce the task definition of multimodal sentiment analysis (MSA), the prerequisite knowledge, and finally, the architecture of UCMIB-PNS.

A. Task Definition

Multimodal sentiment analysis is the task of determining the sentiment score by integrating information from multiple modalities, such as text, audio, and visual data. Given a multimodal input sample indexed by i that consists of multiple modalities, e.g., text I_t, audio I_a, and visual I_v, the goal of multimodal sentiment analysis is to predict the sentiment label Y_i. The task can be formally defined as:

$$Y_i = f_\theta \left(I_t \in \mathbb{R}^{l_t \times d_t}, I_a \in \mathbb{R}^{l_a \times d_a}, I_v \in \mathbb{R}^{l_v \times d_v} \right), \quad (1)$$

where f_θ denotes the MSA model parameterized by θ. The lengths of the text, audio, and visual data are denoted as l_t, l_a, and l_v, respectively, while their corresponding dimensions are d_t, d_a, and d_v.

B. Preliminaries and Causal Assumptions

1) Information Bottleneck: The core idea of the Information Bottleneck (IB) is to balance the trade-off between compression and prediction. Formally, given an input variable X and a target variable Y, the IB method seeks to find a compressed representation Z of X that maximizes the mutual information between Z and Y, while minimizing the mutual information between Z and X. This trade-off is captured by maximizing the following objective function: $I(Z;Y) - \beta I(Z;X)$, where $I(\cdot;\cdot)$ denotes mutual information, β is a Lagrange multiplier that controls the trade-off between compression $I(Z;X)$ and prediction $I(Z;Y)$. Unlike the original IB, which relies on labeled targets Y to define relevance, MIB leverages mutual information across views as a supervision signal. This not only enables the removal of redundant information and the retention of discriminative cross-modal cues, but also reduces the dependence on labeled data, making it

applicable to unsupervised or weakly supervised scenarios. Taking view V_1 as an example, let Z_1 denote the latent representation learned from V_1, and let V_2 represent the complementary view. The objective is to learn Z_1 such that it retains predictive cues from V_2 while minimizing redundant information specific to V_1. Simply maximizing $I(V_1; Z_1)$ is not desirable, as it would encourage Z_1 to preserve all information from V_1, including view-specific details and noise that do not contribute to cross-view prediction. To address this, $I(V_1; Z_1)$ is decomposed into two components: $I(V_1; Z_1) = I(V_1; Z_1|V_2) + I(V_2; Z_1)$. Here, the first term $I(V_1; Z_1|V_2)$ represents the redundant information that should be minimized, while the second term $I(V_2; Z_1)$ denotes the predictive information that needs to be maximized. The optimization of view Z_1 can be achieved by using a relaxed Lagrangian objective:

$$\max_{Z_1} I(V_2; Z_1) - \beta_1 I(V_1; Z_1|V_2), \tag{2}$$

where β_1 is the trade-off parameter for V_1. The essence of multimodality lies in obtaining a comprehensive understanding of content through perception from different views. Therefore, we employ MIB to reduce neither sufficient nor necessary noise across different modalities while maximizing the retention of discriminative information.

2) Necessary and Sufficient Cause: To quantify the sufficient and necessary information across different modalities, we introduce the probabilities of causations from causal inference.

Definition 3.1. Probability of Necessity, PN [68]. Assume that X and Y are binary variables within a causal model M. Let x and y represent the propositions $X = true$ and $Y = true$, respectively, and let x' and y' denote their complements. PN is defined as the expression:

$$\begin{aligned} \text{PN} &\triangleq P(Y_{x'} = false \mid X = true, Y = true) \\ &\triangleq P(y'_{x'} \mid x, y), \end{aligned} \tag{3}$$

where y_x is a counterfactual notation, it indicates the potential response of Y when X is forcibly set to x through an action $do(X = x)$ [69]. PN represents the probability that event y would not have occurred if event x had not happened, given that both x and y did in fact occur.

Definition 3.2. Probability of Sufficiency, PS [68]. In correspondence with PN, PS provides the probability that setting event x would result in producing event y in a situation where both x and y are actually absent. The formula can be expressed as:

$$\begin{aligned} \text{PS} &\triangleq P(Y_x = true \mid X = false, Y = false) \\ &\triangleq P(y_x \mid x', y'), \end{aligned} \tag{4}$$

PS measures he capacity of x to produce y. The probability that the actively triggering event X will induce the occurrence of event Y, when conditioning on the absence of both events X and Y.

Definition 3.3. Probability of Necessity and Sufficiency, PNS [68]. Based on the definitions of PN and PS, we can derive the probability of necessity and sufficiency:

$$\text{PNS} \triangleq P(y_x, y'_{x'}), \tag{5}$$

PNS represents the likelihood that event y would be influenced by x in both necessary and sufficient ways, thereby capturing the dual aspects (necessity and sufficiency) of the event x in causing event y.

Lemma 3.1. The three causal probabilities (PN, PS, and PNS) satisfy the following equation, the detailed proof is provided in Appendix B available online.

$$\begin{aligned} \text{PNS} &= P(y'_{x'}, x, y) + P(y_x, x', y') \\ &= P(x, y)PN + P(x', y')PS. \end{aligned} \tag{6}$$

C. Ucmib-Pns

Based on the aforementioned theories, we propose UCMIB-PNS, the core idea of which is to first conduct unimodal extraction, use modal fusion to construct two multimodal views, and then use an Uncertain Cross-Modal Information Bottleneck (UCMIB) module to compress different multimodal views, remove multimodal noise, finally quantify and balance the sufficiency and necessity of different information through a PNS estimator. The overall architecture is shown in Fig. 2.

1) Unimodal Feature Extraction: Following [24], [26], we employ the BERT-base model to encode textual content, use the COVERAP [70] to encode audio features, and extract visual features using the Facet [71] toolkit. For audio and visual modalities, to enrich their representational flexibilities, we utilize a three-layer Transformer and a single-layer unidirectional LSTM [72] to encode audio and visual content:

$$\begin{aligned} F_t &= BERT(I_t), \\ F_m &= LSTM(Transformer(I_m)), m \in \{a, v\}, \end{aligned} \tag{7}$$

we denote the modality representations after unimodal feature extraction as F_v, F_t, and F_a, respectively.

2) Multimodal Fusion: We adopt a text-modality-dominant dual-branch design, driven by the well-recognized central role of text in multimodal sentiment analysis. As the primary carrier of sentiment expression, text typically conveys richer and more explicit sentimental cues than other modalities, providing fine-grained semantic information essential for accurate sentiment interpretation [11], [12], [18], [62], [73], [74], [75]. From the perspective of model learning efficiency and stability, pairwise fusion methods that combine all modality pairs lead to an exponential increase in the number of fusion branches as the number of modalities grows, which significantly escalates computational complexity and memory consumption. Moreover, integrating noisy signals from less reliable modalities in multiple pairwise combinations can amplify noise propagation across branches, resulting in redundancy and instability during training. By focusing on text-audio and text-vision branches separately, our model limits this combinatorial explosion of fusion views, thereby simplifying the fusion architecture. This design reduces noise accumulation and mitigates overfitting risks, ultimately improving convergence speed and robustness.

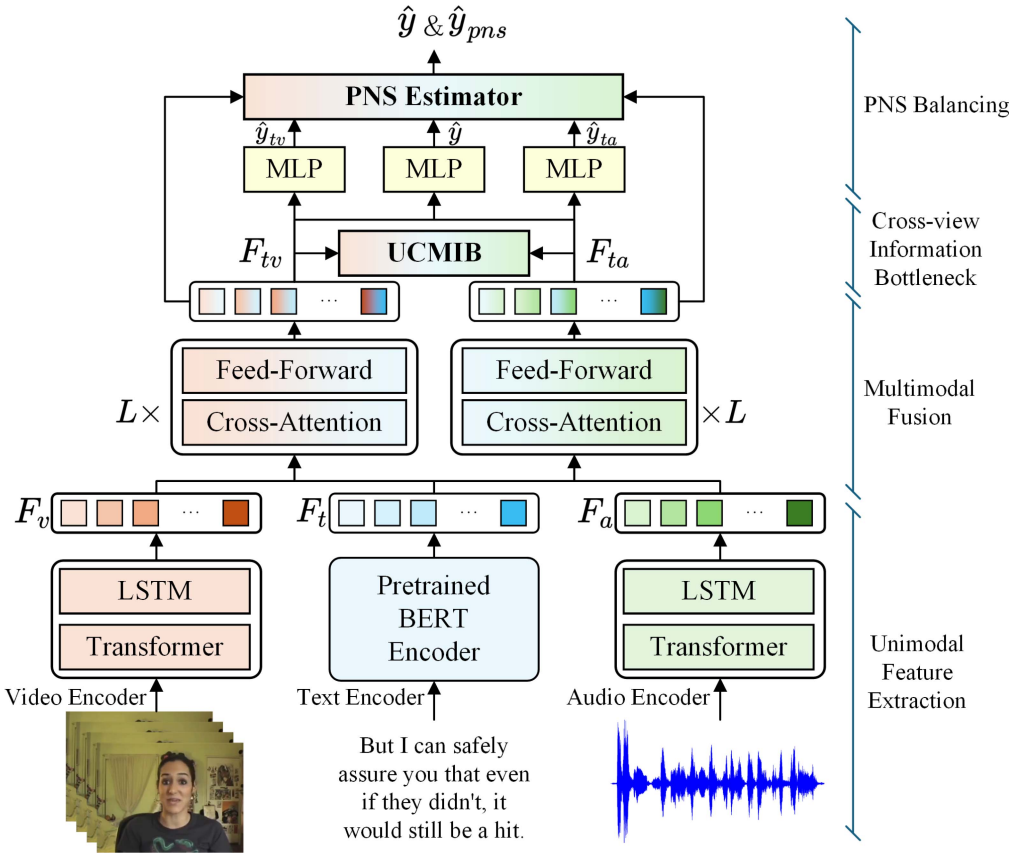


Fig. 2. Pipeline of UCMIB-PNS. The pipeline mainly consists of four parts: unimodal feature extraction, multimodal fusion, cross-view information bottleneck, and PNS balancing.

To integrate different modalities and obtain two comprehensive views, we construct an L-layer Transformer with cross-attention [34] for modality fusion:

$$\hat{h}_{tm}^{i} = MHA(LN(h_{tm}^{i-1}), LN(F_m)) + LN(h_{tm}^{i-1}),\ m \in \{a, v\},$$
$$h_{tm}^{i} = FFN(LN(\hat{h}_{tm}^{i})) + LN(\hat{h}_{tm}^{i}),\ m \in \{a, v\}, \tag{8}$$

where $\hat{h}_{tm}^{i}, h_{tm}^{i}$ represent the intermediate fusion representation and final fusion representation of modality m and modality t at layer i, respectively. Note that $h_{tm}^{0} = F_t$. $LN(\cdot)$ denotes the layer normalization [76] calculation, $FFN(\cdot)$ represents the calculation of the feed-forward layer, and $MHA(\cdot,\cdot)$ denotes the multi-head attention calculation where the first argument is used as the query vector, and the second argument is used as both the key and value vectors. This attention mechanism adopts a text-query design: in each fusion layer, the text representation is used as the query vector, while non-textual modalities (audio or vision) serve as the key and value vectors [13], [18], [34], [75]. This setup allows the model to selectively attend to complementary cues from other modalities, because textual information typically carries the most explicit and reliable sentiment signals, whereas audio and visual cues provide auxiliary but less direct affective evidence. Using text as the query ensures that the fused representation remains anchored to textual semantics, while incorporating keys and values from other modalities facilitates the integration of additional sentiment-relevant information. Furthermore, the inclusion of residual connections and layer normalization mitigates the risk of gradual information dilution, preserving the dominance of textual cues across layers. As the representation evolves into higher-order cross-modal features, using text as the query provides the most effective strategy for final prediction. The output of the multimodal fusion stage consists of two branches: the text-audio branch and the text-vision branch, which are denoted as $F_{ta} = h_{ta}^{L}$ and $F_{tv} = h_{tv}^{L}$, respectively. In the next subsection, we will treat these two branches as two views of sentiment for deeper multi-view information processing.

3) Uncertain Cross-Modal Information Bottleneck: We perform multimodal fusion before applying noise reduction (via UCMIB) to better capture and leverage cross-modal interactions. This design also helps to preserve modality-specific information and align with the information bottleneck principle. Reducing noise first would risk losing valuable information, ignoring cross-modal dependencies, and complicating the model design. The fusion-first approach ensures that the model first integrates complementary multimodal cues and then refines the resulting representations to focus on task-relevant information, making it more effective for sentiment analysis.

When fusing multimodal data, unnecessary noise and redundant information can affect the quantification of modality sufficiency and necessity, as well as the results of modality fusion. To address the issue, we design the Uncertain Cross-Modal Information Bottleneck (UCMIB) module, as shown in Fig. 3.

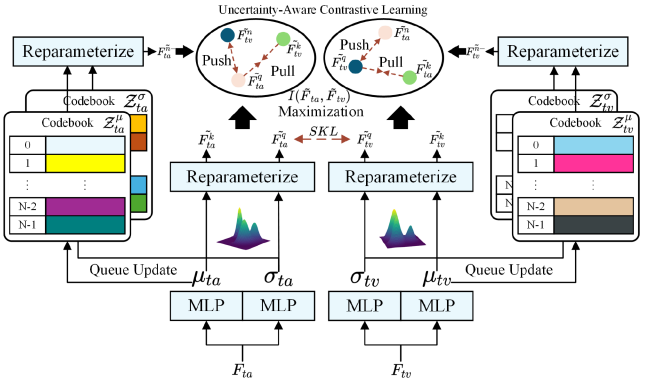


Fig. 3. The Uncertain Cross-modal Information Bottleneck module (UCMIB). The UCMIB module treats the two bimodal representations as two views and models them as multivariate Gaussian distributions. It uses symmetric KL divergence to measure the distributional differences between these views and employs uncertainty-aware contrastive learning (UACL) to maximize the mutual information between different views while reducing redundant information. Two queue-based codebooks are utilized to store the distributions of more negative samples.

Inspired by (2), the objective of UCMIB is to optimize multi-view learning using the multimodal information bottleneck with the following loss function (The detailed derivation is provided in Appendix A) available online:

$$\mathcal{L}_{MIB}(\theta,\psi,\beta) = -I(Z_1;Z_2) + \beta D_{SKL}\left(p_\theta(Z_1|V_1)\|p_\psi(Z_2|V_2)\right), \quad (9)$$

where $I(Z_1;Z_2)$ denotes the mutual information between the two latent representations Z_1 and Z_2, β is a hyperparameter, D_{SKL} denotes the symmetric Kullback-Leibler calculation, $p_\theta(Z_1|V_1)$ and $p_\psi(Z_2|V_2)$ denote the conditional distributions of the latent representations Z_1 and Z_2 given the input views V_1 and V_2, parameterized by θ and ψ, respectively. We use the output of the text-audio branch $V_1 = F_{ta}$ and the output of the text-vision branch $V_2 = F_{tv}$ as the two input views for the UCMIB module.

Existing work [53] has pointed out that distributional representations are conducive to capturing modality uncertainty, and that such uncertainty has a significant effect on denoising. Therefore, we employ Multi-Layer Perceptrons (MLPs) to map each view into a multivariate Gaussian distribution:

$$\mu_{tm} = f_{\theta^m}(F_{tm}),\ \sigma_{tm} = f_{\psi^m}(F_{tm}), \quad (10)$$

where μ_{tm} and σ_{tm} denote the learned mean and covariance of the multivariate Gaussian distribution using function f_{θ^m} and f_{ψ^m}. By using the re-parameterization [77] trick, we can sample the views from the multivariate Gaussian distribution as uncertain latent representations $\tilde{F_{tm}}$, i.e. $Z_1 = \tilde{F_{ta}}$, $Z_2 = \tilde{F_{tv}}$:

$$\tilde{F_{tm}} = \mu_{tm} + \sigma_{tm}\epsilon, \epsilon \sim \mathcal{N}(0,\mathbf{I}). \quad (11)$$

To optimize the first term in (9), we employ uncertainty-aware contrastive learning (UACL) to reduce irrelevant information, concentrating mutual information of Z_1 and Z_2 on what's discriminative for sentiment. Specifically, we perform reparameterization on the multivariate Gaussian distribution twice: one serves as the query point $\tilde{F^q_{tm}}$, i.e., the anchor point, and the other serves as the key point $\tilde{F^k_{tm}}$, i.e., the augmented point. $\tilde{F^q_{tm}}$ and $\tilde{F^k_{tm}}$ originate from the same posterior distribution of the same modality pair, representing semantically equivalent but stochastically perturbed views of the same sample. UACL pulls these two points closer in the representation space to maintain invariance under uncertainty perturbation. In contrast, the anchor $\tilde{F^q_{tm}}$ and a negative sample $\tilde{F^{n-}_{tm'}}$, which originates from a different instance and belongs to another modality branch ($m' \neq m$), encode distinct semantics and are therefore pushed apart in the representation space.

Unlike traditional contrastive learning [13], [27] that searches for negative samples within the batch, we construct a dynamic codebook to store the means and variances of negative samples, thereby achieving a more comprehensive capability for compressing redundant information and ensuring Z_1 and Z_2 stay aligned on sentiment cues. The codebook is built with two goals: maintaining diversity and coverage of negative samples and ensuring stability during updates. It is implemented as a fixed-size queue with N embeddings, each initialized with random Gaussian samples. After each iteration, the mean and variance of the current batch are enqueued, while the oldest entries are dequeued, maintaining a dynamic and memory-efficient repository of negative sample distributions. For the next batch, uncertain negative representations $\tilde{F^{n-}_{tm}}$ are sampled from these stored Gaussian distributions using re-parameterization ((11)).

To optimize the second term in (9), we employ the symmetric KL divergence from different views of the same sample. Overall, UCMIB uses the following loss function for optimization:

$$\mathcal{L}_{UCMIB} = -\sum_{m}^{m\in\{a,v\}} \log \frac{\exp\left(\tilde{F^q_{tm}}\cdot\tilde{F^k_{tm}}/\tau\right)}{\sum_{\tilde{F^{n-}_{tm}}\in\tilde{F^n_{tm}}}\exp\left(\tilde{F^q_{tm}}\cdot\tilde{F^{n-}_{tm}}/\tau\right)} + \beta \sum_{m_1\neq m_2}^{m_1,m_2\in\{a,v\}} D_{SKL}\left(\tilde{F^q_{tm_1}}\|\tilde{F^q_{tm_2}}\right), \quad (12)$$

where τ is a temperature coefficient. Compared with traditional contrastive learning, UCMIB not only maximizes the mutual information but also captures the latent noise and uncertainty in the modalities from the perspective of uncertainty. This lays the foundation for the sufficiency and necessity of modality information in the next subsection.

4) Probability of Necessity and Sufficiency Estimator: Once we have removed the noise that is neither sufficient nor necessary, we can further delve into the quantification of the sufficiency and necessity of different modalities. We design the Probability of Necessity and Sufficiency (PNS) estimator module to quantify and adaptively balance the sufficient and necessary information of different modalities. The architecture of the PNS estimator is shown in Fig. 4.

Inspired by Lemma 3.1, we can estimate PNS through the sum of $P(y'_{x'}, x, y)$ and $P(y_x, x', y')$. We define event x as the feature branch F_{tm} being involved in the fusion, and event y as the model obtaining the correct result. Consequently, x' represents the feature F_{tm} not being involved in the fusion, and y' represents the model not obtaining the correct output. However, the calculation of $P(y'_{x'}, x, y)$ and $P(y_x, x', y')$ intertwine the factual

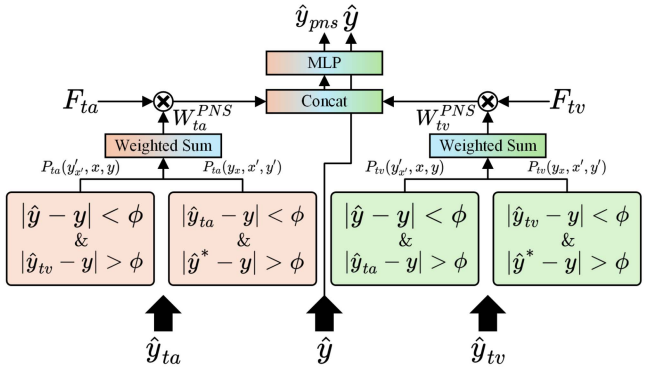


Fig. 4. Pipeline of PNS Estimator. The PNS estimator consists of four submodules to estimate the PNS of F_{ta} and F_{tv}, respectively. By creating factual scenarios (denoted as y) and counterfactual scenarios (denoted as $\hat{y}_{ta}$, $\hat{y}_{tv}$, $\hat{y}$), and restricting the absolute value threshold upper and lower bounds of the output of these scenarios relative to the label y, how much each modality contributes causally to the results are quantified. The final output is obtained through weighted summation guided by PNS. If different branches share high mutual information, PNS can accurately assess their necessity and sufficiency, enabling balanced fusion.

and the counterfactual, in the real world, we can only observe one potential outcome (i.e. Whether a modality participates in fusion is a definitive binary choice in the neural network.). This indicates the need to design parallel and independent modules to simulate factual and counterfactual scenarios.

Taking the estimation of $P_{tm}(y'_{x'}, x, y)$ as an example, its meaning is the probability that the involvement of F_{tm} leads to a correct output, and by forcibly setting F_{tm} not to participate, the correct output cannot be achieved. Its connotation is closely related to the definition of necessity. To determine whether the correct output is achieved, for regression tasks, we set a threshold ϕ between the predicted value and the ground truth. If the difference is less than ϕ, it indicates that the prediction is correct; otherwise, the prediction is incorrect. For classification tasks, the correctness can be directly judged by whether the predicted category matches the true category.

We denote the output of the full-modality fusion as $\hat{y}$, the output using only the F_{tm} modality as $\hat{y}_{tm}$, the output without using the F_{tm} modality as $F_{tm'}$ where m' denotes the other modality, and the ground truth as y. These outputs can be obtained by applying MLPs to different views:

$$\begin{aligned} \hat{y}_{ta} &= MLP(F_{ta}), \\ \hat{y}_{tv} &= MLP(F_{tv}), \\ \hat{y} &= MLP([F_{ta}; F_{tv}]), \end{aligned} \tag{13}$$

where $[\cdots;\cdots]$ denotes the concatenation operation. According to the meaning of $P_{tm}(y'_{x'}, x, y)$, it can be directly obtained through the conjunction of $|\hat{y} - y| < \phi$ and $|\hat{y}_{tm'} - y| > \phi$, where $||$ denotes the absolute value calculation operation. Similarly, in accordance with the semantics of sufficiency, $P_{tm}(y_x, x', y')$ represents the probability that the model outputs incorrectly when all modalities, including F_{tm}, are absent, but by only forcibly including F_{tm}, the prediction $\hat{y}$ matches the ground truth y. It can be obtained through the conjunction of $|\hat{y}_{tm} - y| < \phi$ and $|\hat{y}^* - y| > \phi$, where $\hat{y}^*$ indicating that none of modality is used for prediction, and we use a random guessing scale $\hat{y}^*$ to simulate the model output.

To achieve a balanced model, it is crucial to avoid overemphasizing either sufficiency or necessity. Without balance, the model might: *1) Overemphasize sufficiency:* This can lead to ignoring modalities that are less predictive in isolation but are critical when combined with others. For example, audio features that convey tone might be overlooked if the model only focuses on highly predictive visual features. This imbalance can result in incomplete or biased predictions. *2) Overemphasize necessity:* Focusing too much on noisy or unreliable modalities can degrade the model's performance. For instance, if the model relies heavily on a modality that is prone to errors or inconsistencies, it may fail to capture the true underlying patterns effectively. To address these issues, the fixed hyperparameter λ in the W_{tm}^{PNS} framework explicitly controls the trade-off between necessity and sufficiency. This ensures that the model leverages both aspects to produce robust and accurate predictions. For the features F_{tm}, we use the hyperparameter λ to balance the necessity and sufficiency. The weight W_{tm}^{PNS} is calculated as $W_{tm}^{PNS} = P_{tm}(y'_{x'}, x, y) + \lambda P_{tm}(y_x, x', y')$. The features are weighted and summed based on PNS to serve as the module output $\hat{y}_{pns}$:

$$\hat{y}_{pns} = MLP([W_{ta}^{PNS} F_{ta}; W_{tv}^{PNS} F_{tv}]). \tag{14}$$

This module involves two primary loss functions. To ensure that different feature branches accurately estimate the model output, task loss is employed to optimize each feature branch. It is worth mentioning that the estimation of $\hat{y}$ is also very important. However, since it is essentially the task loss, we treat it as a separate optimization objective for the entire network. To guide the entire model training with PNS, the distance between PNS output and the label should also be minimized, the overall loss is:

$$\mathcal{L}_{PNS} = \sum_{m \in \{a,v\}} \mathcal{L}_{task}(\hat{y}_{tm}, y) + \mathcal{L}_{task}(\hat{y}_{pns}, y), \tag{15}$$

where $\mathcal{L}_{task}$ represents the task loss. For regression tasks such as multimodal sentiment analysis, it is the Mean Absolute Error (MAE). For classification tasks such as multimodal humor detection, it is the cross-entropy loss.

5) Training: The loss function of UCMIB-PNS consists of three parts: task loss, UCMIB module loss, and PNS estimator loss. We use two hyperparameters η_1 and η_2 to balance the different loss functions:

$$\mathcal{L} = \mathcal{L}_{task}(\hat{y}, y) + \eta_1 \mathcal{L}_{UCMIB} + \eta_2 \mathcal{L}_{PNS}. \tag{16}$$

The overall algorithmic procedure of UCMIB-PNS is detailed in Appendix C available online.

IV. Experiments

A. Dataset

We conduct experiments on four publicly available datasets (CMU-MOSI [29], CMU-MOSEI [30], UR-FUNNY [31], CH-SIMS [32]) to validate the effectiveness of our proposed method.

CMU-MOSI. The CMU-MOSI is a multimodal sentiment analysis dataset containing 2,199 short video clips extracted from YouTube movie reviews. It includes 1,284 samples for training, 229 for validation, and 686 for testing. The sentiment labels are annotated on a scale, ranging from -3 (strongly negative) to +3 (strongly positive).

CMU-MOSEI. The CMU-MOSEI is an extension of CMU-MOSI. This dataset includes 23,453 video clips from YouTube, covering a broader range of topics. The dataset is divided into 16,326 samples for training, 1,871 samples for validation, and 4,658 samples for testing. Its annotation framework is completely consistent with the CMU-MOSI dataset.

UR-FUNNY. The UR-FUNNY dataset is designed for humor detection, consisting of 9,588 video clips extracted from TED Talks. The dataset is divided into 7,614 samples for training, 980 samples for validation, and 994 samples for testing. The binary labels are provided for each clip, indicating whether the clip is humorous or not.

CH-SIMS. The CH-SIMS is a Chinese multimodal sentiment analysis dataset containing 2,281 video clips from movies, TV shows, and variety programs. The dataset is divided into 1,366 samples for training, 457 samples for validation, and 458 samples for testing. Sentiment labels are annotated on a scale, ranging from -1 (strongly negative) to +1 (strongly positive).

B. Evaluation Metrics

For CMU-MOSI and CMU-MOSEI, regression performance is assessed using Mean Absolute Error (MAE) and Correlation Coefficient (Corr). For binary sentiment classification, we evaluate results under two configurations: positive versus negative (P/N) and non-negative versus negative (NN/N), utilizing ACC-2 and F1-score. Furthermore, ACC-7 is employed to gauge the precision of fine-grained sentiment classification across the full spectrum from -3 to 3.

For UR-FUNNY, the evaluation metric is ACC-2 in this binary task.

For CH-SIMS, regression accuracy is measured using Mean Absolute Error (MAE). For binary classification, focusing on distinguishing negative from non-negative (N/NN) sentiments, we rely on ACC-2 and F1-score. Additionally, fine-grained sentiment classification across the range of -1 to 1 is evaluated using ACC-5.

C. Implement Details

Feature Extraction: To ensure a fair comparison, we adopt the same feature extraction approach as HyCon [78], which is also widely used by most baseline methods. Specifically, BERT [79] serves as the text encoder. For the audio modality, spectral features, speech polarity, and harmonic parameters are extracted using COVERAP [70]. For the visual modality, Facet [71] is employed to obtain action units, facial landmarks, and other expression-related data.

Hyperparameter Settings: We utilize Adam as the optimizer with a fixed learning rate of 1e-4, a batch size of 64, and hyperparameters η_1 and η_2 set to 0.1 and 0.2, respectively. For CMU-MOSI and CH-SIMS, the ratio α between the codebook size and batch size is set to 2, i.e. the size of the codebook N is 128. For UR-FUNNY and CMU-MOSEI, α is set to 3, corresponding to a codebook size N of 192. The balance parameter λ for necessity and sufficiency is fixed at 1.2. The threshold in PNS estimator ϕ is set to 1.0 for CMU-MOSI and CMU-MOSEI, and 0.3 for CH-SIMS. All experiments in this section are conducted on a single Tesla V100 GPU with 32GB of memory. The results of the hyperparameter sensitivity analysis are presented in Section IV-G.

D. Baselines

To demonstrate the effectiveness of our method for modality fusion under complex conditions, we compare it with existing MSA models including TFN [8], LMF [9], MulT [34], MISA [25], Self-MM [26], MMIM [24], BBFN [14], CENET [17], HyCon-BERT [78], FDMER [80], TETFN [12], TGMN [74], CRNet [81], CET-M [82], TF-BERT [83], FMFN [84], CMLG [85] on standard datasets. For some widely used recently methods (Self-MM [26], MMIM [24], CENET [17], TETFN [12]), we also evaluate their performance under noisy conditions for a comprehensive comparison. The complexity analysis of the model and baselines is provided in Appendix I.

E. Performance Results

To demonstrate the superiority of UCMIB-PNS under different datasets and varying noise conditions, we conducted comparative experiments with baselines on CMU-MOSI and CMU-MOSEI. The results are shown in Table I. Besides, the experimental results on CH-SIMS and UR-FUNNY are presented in Table II. Finally, to showcase the model's capability under visual noise conditions, we added noise to the visual modality of CMU-MOSI and CMU-MOSEI, following the noise injection method in [36]. The results are shown in Table III. In Appendix D, we provide more comprehensive experiments under more noise conditions.

On the CMU-MOSI dataset, we find that, except for ACC-2 and the F1 score under the non-negative/negative (NN/N) setting, UCMIB-PNS achieves the best performance across all other metrics. Specifically, compared to the previous state-of-the-art values, the UCMIB-PNS model has achieved a reduction of 0.16 in MAE, an increase of 1.42% in ACC-7, and improvements of approximately 1.1% in ACC-2 and F1 score under the positive/negative (P/N) setting. These results confirm the effectiveness of our proposed UCMIB-PNS approach, which integrates modality sufficiency and necessity.

On the CMU-MOSEI, UCMIB-PNS can draw conclusions similar to those on the CMU-MOSI dataset, achieving the best performance across all metrics. Specifically, it increases the correlation by 0.13, reduces the MAE by 0.10, and enhances ACC-7 by 0.93%. For ACC-2 and F1, it achieves approximately a 1.1% improvement under the non-negative/negative (NN/N) setting and about a 0.8% improvement under the positive/negative (P/N) setting. These results demonstrate that the UCMIB-PNS method

TABLE I
THE COMPARATIVE RESULTS OF UCMIB-PNS VERSUS OTHER BASELINES ON THE CMU-MOSI AND CMU-MOSEI DATASETS ARE AS FOLLOWS: FOR METRICS ACC-2 AND F1, VALUES TO THE LEFT OF THE "/" SYMBOL REPRESENT PERFORMANCE IN THE NON-NEGATIVE/NEGATIVE (NN/N) SETTING, WHILE THOSE ON THE RIGHT CORRESPOND TO THE POSITIVE/NEGATIVE (P/N) SETTING. METRICS MARKED WITH THE ↑ INDICATE THAT HIGHER VALUES IMPLY BETTER PERFORMANCE, WHEREAS METRICS MARKED WITH THE ↓ SUGGEST THAT LOWER VALUES ARE PREFERABLE. PREVIOUS STATE-OF-THE-ART (SOTA) VALUES ARE UNDERLINED, AND THE CURRENT STATE-OF-THE-ART (SOTA) VALUES ARE HIGHLIGHTED IN BOLD

Models	CMU-MOSI					CMU-MOSEI				
	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑
TFN[a] [8]	0.901	0.698	34.9	-/80.8	-/80.7	0.593	0.700	50.2	-/82.5	-/82.1
LMF[a] [9]	0.917	0.695	33.2	-/82.5	-/82.4	0.623	0.677	48.0	-/82.0	-/82.1
MulT[a] [34]	0.861	0.711	-	81.5/84.1	80.6/83.9	0.580	0.703	-	-/82.5	-/82.3
MISA[a] [25]	0.804	0.764	42.3	80.79/82.10	80.77/82.03	0.568	0.724	-	82.59/84.23	82.67/83.97
Self-MM[a] [26]	0.712	0.795	45.79	82.54/84.77	82.68/84.91	0.529	0.767	53.46	82.68/84.96	82.95/84.93
MMIM[a] [24]	0.700	0.800	46.65	84.14/86.06	84.00/85.98	0.526	0.772	54.24	82.24/85.97	82.66/85.94
BBFN[b] [14]	0.755	0.776	45.00	-/84.30	-/84.30	0.529	0.767	54.8	-/86.20	-/86.10
CENET[b] [17]	0.725	0.795	44.90	83.53/85.21	83.49/85.22	0.526	0.778	54.26	83.52/86.38	83.85/86.32
HyCon-BERT[b] [78]	0.713	0.790	46.6	-/85.2	-/85.1	0.601	0.776	52.8	-/85.4	-/85.6
FDMER[b] [80]	0.724	0.788	44.10	-/84.60	-/84.70	0.536	0.773	54.10	-/86.10	-/85.80
TETFN[b] [12]	0.717	0.800	-	84.05/86.10	83.83/86.07	0.551	0.748	-	84.25/85.18	84.18/85.27
TGMN[b] [74]	0.707	0.786	-	-/86.94	-/87.01	0.529	0.775	-	-/86.22	-/86.29
CRNet[b] [81]	0.712	0.797	47.4	-/86.4	-/86.4	0.541	0.771	53.8	-/86.2	-/86.1
CET-M[b] [82]	0.696	0.805	47.7	84.0/86.0	83.8/85.9	0.523	0.773	54.9	83.4/86.2	83.6/86.1
TF-BERT[b] [83]	0.736	0.776	45.63	**85.57**/86.59	**85.53**/86.66	0.546	0.765	52.52	83.64/85.99	83.81/86.05
FMFN[b] [84]	0.728	0.794	-	84.8/87.0	85.0/87.1	0.533	0.774	-	83.2/86.3	82.8/86.4
CMLG[b] [85]	0.706	0.798	-	84.26/86.43	84.19/86.42	0.547	0.758	-	84.46/85.75	84.54/85.54
UCMIB-PNS	**0.680**	**0.811**	**49.12**	85.28/**87.65**	85.15/**87.59**	**0.513**	**0.791**	**55.83**	**85.55/87.23**	**85.72/87.13**

[a]: results are from [24]; [b]: results are from corresponding original papers.

TABLE II
THE COMPARISON RESULTS BETWEEN UCMIB-PNS AND OTHER BASELINE MODELS ON CH-SIMS AND UR-FUNNY DATASETS ARE PRESENTED BELOW. THE MARKS IN THIS TABLE CORRESPOND TO THOSE IN TABLE I

Model	CH-SIMS				UR-FUNNY
	MAE↓	ACC-2↑	F1↑	ACC-5↑	ACC-2↑
TFN[a] [8]	0.432	78.38	78.62	39.30	68.57
LMF[a] [9]	0.441	77.77	77.88	40.53	67.53
MulT[a] [34]	0.453	78.56	79.66	37.94	70.55
MISA[a] [25]	-	-	-	-	70.61
Self-MM[a] [26]	0.425	80.04	80.44	41.53	-
BBFN[b] [14]	-	-	-	-	71.68
TETFN[a] [12]	0.420	81.18	80.24	41.79	-
FDMER[b] [80]	-	-	-	-	71.87
TGMN[b] [74]	-	81.18	81.43	-	-
CRNet[b] [81]	0.416	80.7	80.7	-	-
FMFN[b] [84]	0.416	80.7	80.7	44.2	-
CMLG[b] [85]	0.408	81.40	81.23	43.54	-
UCMIB-PNS	**0.396**	**85.05**	**84.86**	**44.42**	**72.13**

[a]: results are from [26] and its corresponding github page; [b]: results are from corresponding original papers.

TABLE III
THE COMPARISON RESULTS (F1-SCORE) BETWEEN UCMIB-PNS AND OTHER BASELINE MODELS ON VISUAL NOISY CMU-MOSI AND CMU-MOSEI UNDER POSITIVE/NEGATIVE (P/N) SETTING. THE MARKS ARE CONSISTENT WITH THOSE USED IN TABLE I

Noisy CMU-MOSI					
Model	Clean	Salt-Pepper Noise		Gaussian Noise	
	ϵ=0	ϵ=5	ϵ=10	ϵ=5	ϵ=10
Self-MM [26]	84.91	83.88	83.29	83.76	82.53
MMIM [24]	85.98	85.13	83.30	84.83	84.05
CENET [17]	85.22	84.17	84.09	84.41	83.98
TETFN [12]	86.07	84.02	82.66	84.32	83.76
UCMIB-PNS	**87.59**	**85.89**	**85.65**	**86.24**	**85.98**

Noisy CMU-MOSEI					
Model	Clean	Salt-Pepper Noise		Gaussian Noise	
	ϵ=0	ϵ=5	ϵ=10	ϵ=5	ϵ=10
Self-MM [26]	84.93	83.61	83.01	84.56	84.13
MMIM [24]	85.94	84.11	83.30	85.38	83.64
CENET [17]	86.32	84.57	84.29	85.39	85.19
TETFN [12]	85.27	84.84	83.99	85.13	83.62
UCMIB-PNS	**87.13**	**86.18**	**85.61**	**86.03**	**85.97**

can adapt to more complex datasets and still improve the model's fusion performance through the assessment of sufficiency and necessity.

On the CH-SIMS, UCMIB-PNS has achieved remarkably significant improvements in accuracy. Specifically, the MAE has decreased by 0.12, ACC-2 has increased by 3.65%, F1 has improved by 3.43%, and ACC-5 has increased by 0.22%. Even in different language scenarios, UCMIB-PNS remains effective and achieves highly noticeable improvements across various granularity levels of sentiment classification tasks. On the UR-FUNNY dataset, it achieved a precision improvement of 0.26%. These results confirm the broad applicability of UCMIB-PNS across diverse multimodal tasks.

The results from the comparison of UCMIB-PNS with other baseline models on the noisy CMU-MOSI and CMU-MOSEI datasets demonstrate the robustness and superior performance of UCMIB-PNS across various noise conditions. As the noise level increases, the trend of decreasing model accuracy is observed. UCMIB-PNS consistently achieves the highest F1-scores in both clean and noisy settings. This highlights the effectiveness of its adaptive fusion mechanism based on modality sufficiency and necessity, which allows it to maintain high performance even under increasing levels of salt-pepper and Gaussian noise. Overall, UCMIB-PNS proves to be a highly versatile and reliable model for multimodal sentiment analysis.

TABLE IV
THE ABLATION RESULTS ON THE CMU-MOSI, CMU-MOSEI DATASETS

Models	CMU-MOSI					CMU-MOSEI				
	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑
UCMIB-PNS	**0.680**	**0.811**	**49.12**	**85.28/87.65**	85.15/**87.59**	**0.513**	0.791	**55.83**	**85.55/87.23**	**85.72/87.13**
w/o Codebook	0.699	0.806	46.50	84.69/86.28	84.67/86.30	0.522	0.785	55.12	85.10/86.76	85.28/86.66
w/o UACL	0.707	0.800	46.64	84.40/85.82	84.33/85.79	0.520	**0.792**	54.71	83.56/86.13	83.91/86.10
w/o $\mathcal{L}_{CL}$	0.693	0.807	44.61	84.26/86.59	84.06/86.47	0.526	0.768	53.85	84.63/86.82	84.94/86.80
w/o $\mathcal{L}_{D_{SKL}}$	0.704	0.806	45.18	**85.28**/86.59	**85.24**/86.59	0.515	0.787	54.82	85.43/86.63	85.54/86.49
w/o $\mathcal{L}_{UCMIB}$	0.717	0.798	44.90	83.67/85.21	83.58/85.17	0.534	0.751	54.13	83.17/85.28	83.47/85.20
w/o $\mathcal{L}_{PNS_{tv}}$	0.703	0.801	46.06	84.69/86.74	84.54/86.65	0.522	0.778	54.54	85.08/86.60	85.09/86.34
w/o $\mathcal{L}_{PNS_{ta}}$	0.685	0.804	46.65	84.11/86.13	84.01/86.09	0.521	0.783	54.67	84.95/86.85	85.10/86.70
w/o $\mathcal{L}_{PNS}$	0.718	0.799	44.31	83.97/85.37	83.94/85.39	0.532	0.777	53.87	84.20/85.33	84.16/85.01
T	0.777	0.783	43.59	82.22/83.99	82.17/84.01	0.537	0.764	54.11	81.46/85.36	82.00/85.38
V	1.432	0.254	15.60	48.69/46.34	37.15/34.92	0.824	0.157	42.34	62.14/61.89	62.60/60.90
A	1.426	0.141	15.01	51.17/49.85	47.83/46.68	0.833	0.104	41.36	69.39/62.88	62.14/53.77
T+A	0.758	0.788	42.42	82.65/84.30	82.62/84.32	0.532	0.779	54.75	84.25/86.27	84.33/86.03
T+V	0.715	0.792	46.50	82.94/84.45	82.90/84.46	0.530	0.775	53.46	85.32/86.08	85.54/85.81
A+V	1.459	0.090	20.11	55.98/55.48	56.09/55.76	0.837	0.108	42.56	71.02/62.85	60.97/58.51

TABLE V
THE ABLATION RESULTS (F1-SCORE) ON THE CMU-MOSI DATASET WITH VISUAL NOISE UNDER POSITIVE/NEGATIVE (P/N) SETTING

Noisy CMU-MOSI					
Models	Clean	Salt-Pepper Noise		Gaussian Noise	
	ϵ=0	ϵ=5	ϵ=10	ϵ=5	ϵ=10
UCMIB-PNS	**87.59**	85.89	**85.65**	**86.24**	**85.98**
w/o Codebook	86.30	84.97	84.76	85.66	85.39
w/o UACL	85.79	85.06	84.53	84.81	84.04
w/o $\mathcal{L}_{CL}$	86.47	85.44	85.31	85.61	85.18
w/o $\mathcal{L}_{D_{SKL}}$	86.59	85.40	84.92	85.68	85.04
w/o $\mathcal{L}_{UCMIB}$	85.17	84.70	84.40	84.53	83.85
w/o $\mathcal{L}_{PNS_{tv}}$	86.65	**86.05**	85.54	85.58	85.44
w/o $\mathcal{L}_{PNS_{ta}}$	86.09	85.91	85.05	85.44	85.08
w/o $\mathcal{L}_{PNS}$	85.39	84.47	83.78	84.03	83.45

In summary, these experiments have demonstrated that UCMIB-PNS can adapt to a wide range of multimodal scenarios, including datasets of varying sizes, different languages, diverse types of multimodal tasks, and various noise conditions, achieving optimal performance across all these settings. Unlike traditional modality fusion methods based on maximum likelihood, the adaptive fusion approach of UCMIB-PNS, which is grounded in modality sufficiency and necessity, effectively enhances multimodal integration and significantly improves model accuracy.

F. Ablation Study

We conducted ablation experiments on CMU-MOSI and CMU-MOSEI, and the results are shown in Table IV. Besides, we performed ablation experiments on CMU-MOSI with visual noise, and the results are shown in Table V to verify the effectiveness of each module. Additional ablation experiments on noisy datasets are provided in Appendix E.

The effect of the Cross-Model Information Bottleneck module (the top part of Tables IV and V): To verify the effectiveness of the UCMIB module, we conduct experiments by removing the codebook (denoted as w/o Codebook), removing the uncertainty in contrastive learning (denoted as w/o UACL), removing the entire contrastive learning (denoted as w/o $\mathcal{L}_{CL}$), removing the symmetric KL loss (denoted as w/o $\mathcal{L}_{D_{SKL}}$), and removing the entire UCMIB loss (denoted as w/o $\mathcal{L}_{UCMIB}$). After that, we observe their impacts on the performance. We found that the complete model achieved the best results on the CMU-MOSI, CMU-MOSEI, and noisy CMU-MOSI datasets, while the model with the complete removal of the module loss achieved the worst results. Specifically, the conclusions drawn from the CMU-MOSI, CMU-MOSEI, and noisy CMU-MOSI datasets are consistent. The codebook, by maintaining a larger number of negative samples, can be effectively combined with uncertain contrastive learning to help the model overcome the recognition of some challenging samples. An interesting finding is that the uncertainty in contrastive learning is even more critical than the contrastive learning itself. Aligning representations across different modalities without considering the inherent uncertainty within each modality may potentially impair model accuracy. The symmetric KL divergence also contributes to model improvement, as it encourages the learning of unique features from different views and promotes more consistent model outputs. These observations collectively demonstrate the effectiveness of UCMIB and its sub-modules in handling both clean and noisy multimodal fusion.

The effect of the PNS estimator module (the middle part of Tables IV and V): To verify the effectiveness of the PNS estimator and its sub-modules, we successively removed the estimation of sufficiency and necessity for the text-audio branch (denoted as w/o $\mathcal{L}_{PNS_{ta}}$), the estimation of sufficiency and necessity for the text-vision branch (denoted as w/o $\mathcal{L}_{PNS_{tv}}$), and the entire PNS estimator module (denoted as w/o $\mathcal{L}_{PNS}$). We found that on CMU-MOSI, CMU-MOSEI, and the noisy CMU-MOSI, removing any single branch does not lead to better results, while removing all branches results in the worst model performance. Moreover, on the CMU-MOSI and noisy CMU-MOSI datasets, the estimation of the sufficiency and necessity of the text-audio branch is more critical, while on the CMU-MOSEI dataset, the text-vision branch is slightly more important. This may be related to the different modal characteristics present in each dataset. The findings confirm the importance of the PNS

TABLE VI
THE SENSITIVITY ANALYSIS OF η_1

η_1	CMU-MOSI	CMU-MOSEI	CH-SIMS	UR-FUNNY
0	44.90	54.13	43.10	69.52
0.1	**49.12**	**55.83**	**44.42**	**72.13**
0.2	47.95	54.24	41.79	71.33
0.4	45.77	52.29	41.53	70.52

TABLE VII
THE SENSITIVITY ANALYSIS OF η_2

η_2	CMU-MOSI	CMU-MOSEI	CH-SIMS	UR-FUNNY
0	44.31	53.87	39.16	69.72
0.1	47.38	54.58	42.89	71.23
0.2	**49.12**	**55.83**	**44.42**	**72.13**
0.4	46.79	54.48	40.53	71.03

TABLE VIII
THE SENSITIVITY ANALYSIS OF λ

λ	CMU-MOSI	CMU-MOSEI	CH-SIMS	UR-FUNNY
0	44.75	52.77	38.29	70.52
0.4	46.36	53.49	39.16	71.03
0.8	47.08	54.04	41.35	71.33
1.2	**49.12**	**55.83**	**44.42**	**72.13**
1.6	47.38	54.21	43.10	71.83

estimator and its sub-modules. Even when the sufficiency and necessity of one perspective are given, the model still struggles to balance the fusion weights of different views through end-to-end learning. Therefore, the estimation of sufficiency and necessity for all views is of great importance.

The effect of the modality combination (the bottom part of Table IV) clearly demonstrates the necessity of leveraging complementary information across modalities. Among unimodal settings, the textual modality (T) performs significantly better than the visual (V) and acoustic (A) counterparts across all metrics on both CMU-MOSI and CMU-MOSEI, confirming its dominant role in sentiment analysis. However, fusing textual information with either visual (T+V) or acoustic (T+A) inputs yields noticeable performance improvements, suggesting that even though V and A modalities are weak individually, they provide valuable complementary cues when integrated with text. Notably, the T+V combination achieves performance close to the full model, especially on CMU-MOSEI, highlighting the informative role of visual features in larger datasets. In contrast, the A+V setting performs poorly, reinforcing that the absence of textual signals severely degrades sentiment prediction accuracy. The full multimodal setting achieves the best overall results. These results confirm the effectiveness of multimodal fusion and the importance of balanced cross-modal learning.

G. Sensitivity Analysis

In this section, we perform a sensitivity analysis on the three most critical hyperparameters involved in the proposed UCMIB-PNS framework: η_1, η_2, and λ. We adopt ACC-7 as the evaluation metric for CMU-MOSI and CMU-MOSEI, ACC-5 for CH-SIMS, and ACC-2 for UR-FUNNY. The sensitivity analysis results for the remaining two hyperparameters α and ϕ on the clean datasets are provided in Appendix F available online. The results of the sensitivity analysis under noisy conditions are provided in Appendix G available online. In addition, we provide the counterfactual evaluation results of UCMIB-PNS in Appendix J available online.

The results in Table VI demonstrate that the hyperparameter η_1, which controls the strength of the Uncertain Cross-modal Information Bottleneck (UCMIB) in the loss function, plays a critical role in model performance. As η_1 increases from 0 to 0.1, performance improves consistently across all datasets, indicating that moderate information compression between modalities effectively reduces redundant or noisy cross-modal signals and enhances generalization. However, further increasing η_1 beyond 0.1 leads to a performance drop, suggesting that excessive compression may hinder the optimization of other loss components, ultimately weakening the overall training process. Therefore, setting $\eta_1 = 0.1$ corresponds to an appropriate strength of information bottleneck compression.

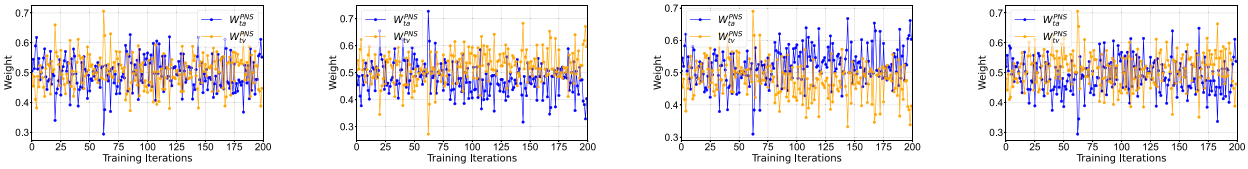
The sensitivity analysis results in Table VII demonstrate the influence of the hyperparameter η_2, which controls the loss weight of the Probability of Necessity and Sufficiency (PNS) Estimator. As η_2 increases from 0 to 0.2, performance steadily improves across all datasets, suggesting that better estimation of PNS enhances overall model effectiveness. However, further increasing η_2 to 0.4 results in performance degradation, likely due to interference with other optimization objectives. This trend is consistent with the pattern observed for η_1, where moderate strength improves performance, but excessive weighting disrupts the overall learning process. Therefore, $\eta_2 = 0.2$ strikes the best balance between estimation accuracy and stable joint optimization.

Table VIII presents the sensitivity analysis of the hyperparameter λ, which controls the balance between sufficiency and necessity. We observe a consistent performance improvement as λ increases, with the best results achieved at $\lambda = 1.2$ across all datasets. This indicates that assigning a higher weight to necessity—often underemphasized in existing methods—can lead to better performance. When λ is too small, the model underperforms due to insufficient emphasis on necessity. However, when λ becomes too large, performance slightly declines, indicating that overemphasizing necessity may also be suboptimal. These findings suggest that properly emphasizing necessity strikes an effective balance and leads to enhanced model performance.

H. Visualization and Case Study

To rigorously demonstrate the robust superiority of UCMIB-PNS under noisy conditions, we present a comprehensive visualization and case-study analysis. More visualization results of noise robustness are provided in Appendix H.

1) Visualization of Dynamic Weight Assignment: To reveal how the PNS estimator dynamically allocates weights under different noise conditions, we visualized the weight curves for text-audio and text-vision branches of different modal noises during the training iteration process, as shown in Fig. 5. We find that when there is no noise or when the text contains



(a) The weight curve of W_{tm}^{PNS} during training when no noise is injected. (b) The weight curve of W_{tm}^{PNS} during training under the condition of audio salt-pepper noise $\epsilon = 5$. (c) The weight curve of W_{tm}^{PNS} during training under the condition of visual salt-pepper noise $\epsilon = 5$. (d) The weight curve of W_{tm}^{PNS} during training under the condition of text blank noise $\epsilon = 50$.

Fig. 5. Visualization of the weight curve of W_{tm}^{PNS} under different modal noise conditions, with each iteration representing a sampling point. As training progresses, the noisy modality branches are assigned smaller weights.

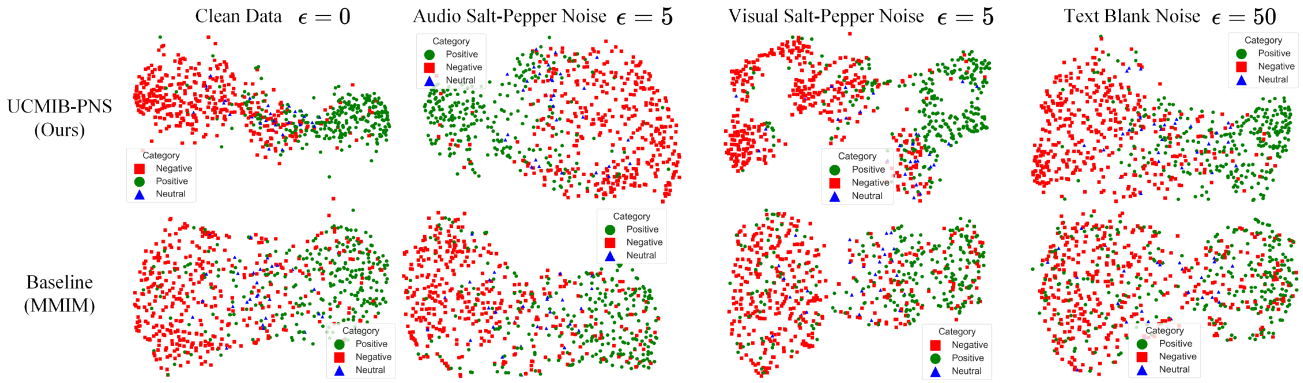


Fig. 6. The visualization results of the final representations using t-SNE [86]. The samples are derived from the test set of CMU-MOSI. Under both noise-free and different types of noise conditions, the sentimental decision boundary of UCMIB-PNS is clearer than the baseline, confirming its strong adaptability to complex scenarios.

noise, the weights of the text-audio and text-vision branches are relatively balanced, with no particularly significant differences. This indicates that during the training process, the information contained in the text-audio and text-vision branches is complementary, with both branches possessing considerable sufficiency and necessity. The differences only lie in the slight variations in their distribution across different batches. This is because the information they obtain from the text is similar. Essentially, they rely on the relevance of audio and vision, as well as their integration with the text, to gain weight benefits. When noise appears in the visual and audio modalities, in the initial stage of training, the model adapts to these noises through the UCMIB module and learnable parameters, without showing any significant differences in weight allocation. As the training progresses to the later stages, the PNS estimator successfully captures the noise variations and reduces the weight allocation for the branches where the noisy modalities exist. The visualization results of the training process are consistent with rational understanding and are also the expected outcomes of this paper. If one modality contains noise, the model needs to seek more sentimental clues from other modalities. This demonstrates that the PNS estimator can effectively reallocate weights across modalities in response to noise, enhancing the model's robustness and adaptability.

2) Visualization of Joint Representation: On the test samples of the CMU-MOSI dataset, we visualize the baseline and UCMIB-PNS under clean data and different modal types of noise using t-SNE, as shown in Fig. 6. We find that noise directly interferes with the final representations of modalities, making the decision boundaries for sentiment unclear. Moreover, this interference has an extremely severe impact on the traditional baseline, especially when it comes to text blank noise, which can cause a more significant impact compared to the salt-and-pepper noise in the audio and video modalities. However, we found that the proposed UCMIB-PNS not only achieves better representations than the baseline in the noise-free scenario, but also, thanks to the noise filtering mechanism of UCMIB and the modality re-weighting mechanism of PNS based on the sufficiency and necessity of modalities, UCMIB-PNS is able to better resist noise from different modalities and obtain clearer decision boundaries compared to the baseline under various types of noise. In summary, UCMIB-PNS can effectively resist noise interference and promote the fusion of multimodal representations in complex scenarios.

3) Case Analysis: Moreover, we randomly selected five samples and compared the predictions with the baseline (MMIM [24]) under different noise conditions. The results of the case analysis are shown in Fig. 7. These noise conditions are truly present in real-world scenarios. We find that even though all the samples have clear sentimental semantics, the baseline struggles to produce the correct output when disturbed by noise. Our proposed UCMIB-PNS can stably produce the correct output, which demonstrates the advantages of the UCMIB module and the PNS estimation module in dealing with noisy samples. Our proposed UCMIB-PNS model proves to be highly effective and robust, confirming its potential for real-world applications. In addition, we provide a case analysis of the PNS working mechanism based on causal paths in Appendix K.

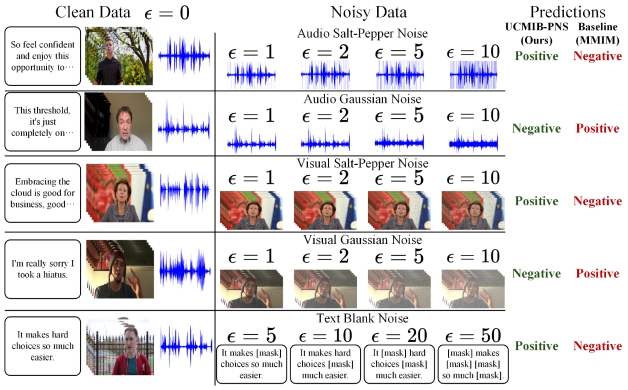


Fig. 7. A case analysis was conducted on five test samples from CMU-MOSEI. Compared with the baseline (MMIM [24]), our model can adapt to different types and levels of noise and produce correct outputs.

V. Conclusion

In this paper, we propose a novel modality fusion method, UCMIB-PNS, which is designed to balance the sufficiency and necessity of information across modalities for robust multimodal sentiment analysis. Grounded in information bottleneck theory and guided by probabilistic causality, UCMIB-PNS employs the UCMIB module to reduce redundant information and maximize discriminative cues within modalities. The PNS estimator further enhances this balance by dynamically estimating and reweighting the sufficiency and necessity of information. This approach effectively filters out noise and highlights the most informative features, thereby significantly improving the robustness and accuracy of multimodal sentiment analysis. UCMIB-PNS achieves state-of-the-art performance on four public datasets under both clean and noisy settings. Comprehensive experiments and extended analyses confirm that our method not only optimizes the fusion of multimodal representations but also remains resilient to various types of noise, leading to more reliable sentiment predictions. Future work will focus on further exploring the generalizability of UCMIB-PNS to other multimodal tasks and datasets. Furthermore, efforts should also be directed towards enhancing the model through large-scale pretraining, with the goal of incorporating sufficient and necessary knowledge to improve its performance in complex, real-world scenarios.

References

[1] M. D. A. Rahman, M. S. Hossain, N. A. Alrajeh, and B. B. Gupta, "A multimodal, multimedia point-of-care deep learning framework for COVID-19 diagnosis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 17, no. 1s, pp. 1–24, 2021.

[2] H. Yin, S. Yang, X. Song, W. Liu, and J. Li, "Deep fusion of multimodal features for social media retweet time prediction," *World Wide Web*, vol. 24, no. 4, pp. 1027–1044, 2021.

[3] C. Zhang et al., "M3Care: Learning with missing modalities in multimodal healthcare data," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 2418–2428.

[4] M. Alibeigi et al., "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 20178–20188.

[5] S. Qiu et al., "Multimodal deep learning for Alzheimer's disease dementia assessment," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 3404.

[6] U. Singh, K. Abhishek, and H. K. Azad, "A survey of cutting-edge multimodal sentiment analysis," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–38, 2024.

[7] Z. Qu, Y. Meng, G. Muhammad, and P. Tiwari, "QMFND: A quantum multimodal fusion-based fake news detection model for social media," *Inf. Fusion*, vol. 104, 2024, Art. no. 102172.

[8] M. Yang et al., "Invariant learning via probability of sufficient and necessary causes," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 79832–79857, 2023.

[9] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[10] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[11] C. Huang, J. Zhang, X. Wu, Y. Wang, M. Li, and X. Huang, "TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis," *Knowl.-Based Syst.*, vol. 269, 2023, Art. no. 110502.

[12] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognit.*, vol. 136, 2023, Art. no. 109259.

[13] Q. Huang, J. Chen, C. Huang, X. Huang, and Y. Wang, "Text-centered cross-sample fusion network for multimodal sentiment analysis," *Multimedia Syst.*, vol. 30, no. 4, 2024, Art. no. 228.

[14] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interaction*, 2021, pp. 6–15.

[15] F. Wang et al., "TEDT: Transformer-based encoding–decoding translation network for multimodal sentiment analysis," *Cogn. Computation*, vol. 15, no. 1, pp. 289–303, 2023.

[16] J. Huang, J. Zhou, Z. Tang, J. Lin, and C. Y.-C. Chen, "TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis," *Knowl.-Based Syst.*, vol. 285, 2024, Art. no. 111346.

[17] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, and X. Gao, "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Trans. Multimedia*, vol. 25, pp. 4909–4921, 2023.

[18] J. Chen, Q. Huang, C. Huang, and X. Huang, "Actual cause guided adaptive gradient scaling for balanced multimodal sentiment analysis," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 21, 2025, Art. no. 179.

[19] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 7837–7851.

[20] K. Kim and S. Park, "AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis," *Inf. Fusion*, vol. 92, pp. 37–45, 2023.

[21] F. Liu, Z. Fu, Y. Wang, and Q. Zheng, "TACFN: Transformer-based adaptive cross-modal fusion network for multimodal emotion recognition," *CAAI Artif. Intell. Res.*, vol. 2, 2023, Art. no. 9150019.

[22] G. Yi et al., "VLP2MSA: Expanding vision-language pre-training to multimodal sentiment analysis," *Knowl.-Based Syst.*, vol. 283, 2024, Art. no. 111136.

[23] Z. Fu, F. Liu, Q. Xu, X. Fu, and J. Qi, "LMR-CBT: Learning modality-fused representations with CB-transformer for multimodal emotion recognition from unaligned multimodal sequences," *Front. Comput. Sci.*, vol. 18, no. 4, 2024, Art. no. 184314.

[24] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9180–9192.

[25] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[26] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10790–10797.

[27] R. Lin and H. Hu, "Multi-task momentum distillation for multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 15, no. 2, pp. 549–565, Apr.–Jun. 2024.

[28] Q. Lu, X. Sun, Y. Long, Z. Gao, J. Feng, and T. Sun, "Sentiment analysis: Comprehensive reviews, recent advances, and open challenges," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15092–15112, Nov. 2024.

[29] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.

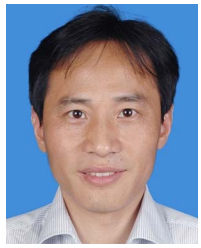
[30] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[31] M. K. Hasan et al., "UR-FUNNY: A multimodal language dataset for understanding humor," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 2046–2056.
[32] W. Yu et al., "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.
[33] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[34] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2019, pp. 6558–6569.
[35] J. Qin, F. Liu, and L. Zong, "BC-PMJRS: A brain computing-inspired predefined multimodal joint representation spaces for enhanced cross-modal learning," *Neural Netw.*, vol. 188, 2025, Art. no. 107449.
[36] Q. Zhang et al., "Provable dynamic fusion for low-quality multimodal data," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 41753–41769.
[37] Y. Fang et al., "Dynamic multimodal information bottleneck for multimodality classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 7696–7706.
[38] Q. Zhang et al., "Multimodal fusion on low-quality data: A comprehensive survey," 2024, *arXiv:2404.18947*.
[39] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop*, 2015, pp. 1–5.
[40] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learn. Representations*, 2017.
[41] S. Bang, P. Xie, H. Lee, W. Wu, and E. Xing, "Explaining a black-box by using a deep variational information bottleneck approach," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11396–11404.
[42] H. Zhu, W. Liu, Z. Gao, and H. Zhang, "Explainable classification of benign-malignant pulmonary nodules with neural networks and information bottleneck," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 2028–2039, Feb. 2025.
[43] S. Seo, S. Kim, J. Jung, Y. Lee, and C. Park, "Self-explainable temporal graph networks based on graph information bottleneck," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2024, pp. 2572–2583.
[44] S. Seo, S. Kim, and C. Park, "Interpretable prototype-based graph information bottleneck," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 76737–76748.
[45] X. Lu, K. Lee, P. Abbeel, and S. Tiomkin, "Dynamics generalization via information bottleneck in deep reinforcement learning," 2020, *arXiv:2008.00614*.
[46] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.
[47] K. Ahuja et al., "Invariance principle meets information bottleneck for out-of-distribution generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 3438–3450.
[48] B. Dai, C. Zhu, B. Guo, and D. Wipf, "Compressing neural networks using the variational information bottleneck," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1135–1144.
[49] A. Srivastava, O. Dutta, J. Gupta, S. Agarwal, and P. AP, "A variational information bottleneck based method to compress sequential networks for human action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2745–2754.
[50] B. Razeghi, S. Rezaeifar, S. Ferdowsi, T. Holotyak, and S. Voloshynovskiy, "Compressed data sharing based on information bottleneck model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2022, pp. 3009–3013.
[51] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 14200–14213.
[52] S. Mai, Y. Zeng, and H. Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE Trans. Multimedia*, vol. 25, pp. 4121–4134, 2023.
[53] Z. Gao, X. Jiang, X. Xu, F. Shen, Y. Li, and H. T. Shen, "Embracing unimodal aleatoric uncertainty for robust multimodal fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26876–26885.
[54] S. Cui et al., "Enhancing multimodal entity and relation extraction with variational information bottleneck," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1274–1285, 2024.
[55] T. Wang, J. Huang, H. Zhang, and Q. Sun, "Visual commonsense R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10760–10770.
[56] X. Yang, H. Zhang, G. Qi, and J. Cai, "Causal attention for vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9847–9857.
[57] C. Huang, H. Wei, Q. Huang, F. Jiang, Z. Han, and X. Huang, "Learning consistent representations with temporal and causal enhancement for knowledge tracing," *Expert Syst. Appl.*, vol. 245, 2024, Art. no. 123128.
[58] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11624–11641, Oct. 2023.
[59] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual VQA: A cause-effect look at language bias," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12700–12710.
[60] T. Sun, W. Wang, L. Jing, Y. Cui, X. Song, and L. Nie, "Counterfactual reasoning for out-of-distribution multimodal sentiment analysis," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 15–23.
[61] P.-J. Huang, H. Xie, H.-C. Huang, H.-H. Shuai, and W.-H. Cheng, "CA-FER: Mitigating spurious correlation with counterfactual attention in facial expression recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 977–989, Jul.–Sep. 2024.
[62] C. Huang, J. Chen, Q. Huang, S. Wang, Y. Tu, and X. Huang, "AtCAF: Attention-based causality-aware fusion network for multimodal sentiment analysis," *Inf. Fusion*, vol. 114, 2025, Art. no. 102725.
[63] K. Zhang, M. Gong, and B. Schölkopf, "Multi-source domain adaptation: A causal view," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 3150–3157.
[64] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 3020–3029.
[65] S. Magliacane, T. Van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, "Domain adaptation by using causal inference to predict invariant conditional distributions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10869–10879.
[66] C. Liu et al., "Learning causal semantic representation for out-of-distribution prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 6155–6170.
[67] M. Yang et al., "Invariant learning via probability of sufficient and necessary causes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 79832–79857.
[68] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[69] J. M. Robins, "Discussion of causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 695–698, 1995.
[70] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.
[71] T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, *Visual Analysis of Humans: Looking at People*. London, U.K.: Springer, 2011.
[72] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[73] B. Yang, B. Shao, L. Wu, and X. Lin, "Multimodal sentiment analysis with unidirectional modality translation," *Neurocomputing*, vol. 467, pp. 130–137, 2022.
[74] Y. Luo, R. Wu, J. Liu, and X. Tang, "A text guided multi-task learning network for multimodal sentiment analysis," *Neurocomputing*, vol. 560, 2023, Art. no. 126836.
[75] C. Zhu et al., "Skeafn: Sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis," *Inf. Fusion*, vol. 100, 2023, Art. no. 101958.
[76] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[77] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.
[78] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2276–2289, Jul.–Sep. 2023.
[79] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2019, pp. 4171–4186.
[80] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1642–1651.

[81] H. Shi et al., "Co-space representation interaction network for multimodal sentiment analysis," *Knowl.-Based Syst.*, vol. 283, 2024, Art. no. 111149.
[82] B. Sun, L. Jia, Y. Cui, N. Wang, and T. Jiang, "Conv-enhanced transformer and robust optimization network for robust multimodal sentiment analysis," *Neurocomputing*, vol. 634, 2025, Art. no. 129842.
[83] J. Hou, N. Omar, S. Tiun, S. Saad, and Q. He, "TF-BERT: Tensor-based fusion BERT for multimodal sentiment analysis," *Neural Netw.*, vol. 185, 2025, Art. no. 107222.
[84] X. Li, H. Zhang, Z. Dong, X. Cheng, Y. Liu, and X. Zhang, "Learning fine-grained representation with token-level alignment for multimodal sentiment analysis," *Expert Syst. Appl.*, vol. 269, 2025, Art. no. 126274.
[85] R. Wang, Q. Yang, S. Tian, L. Yu, X. He, and B. Wang, "Transformer-based correlation mining network with self-supervised label generation for multimodal sentiment analysis," *Neurocomputing*, vol. 618, 2025, Art. no. 129163.
[86] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[87] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition," in *Proc. 17th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process.*, 2020, pp. 117–121.
[88] Z. Chen, L. Hu, W. Li, Y. Shao, and L. Nie, "Causal intervention and counterfactual reasoning for multi-modal fake news detection," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 627–638.
[89] Z. Zhu et al., "TFCD: Towards multi-modal sarcasm detection via training-free counterfactual debiasing," in *Proc. Int. Joint Conf. Artif. Intell.*, 2024, pp. 6687–6695.
[90] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1025–1034.

Jili Chen received the MS degree in educational technology in 2025 from Zhejiang Normal University, Jinhua, China, where he is currently working toward the doctor's degree in computer science and technology. His research interests include causal inference, intelligent education, large language models and applications, and computer vision.

Yihua Zhong received the MS degree in educational technology from Zhejiang Normal University, Jinhua, China, in 2024. He is currently working toward the doctor's degree in intelligent education from East China Normal University, Shanghai, China. His research interests include AI for education, large language models and applications, and educational agent.

Qionghao Huang received the master's degree in software engineering from South China Normal University in 2018, and the PhD degree in intelligence education from South China Normal University. He is currently an associate professor with Zhejiang Normal University, Jinhua, China. His research interests include deep learning methods, affective computing, Big Data mining and intelligent education application. Dr. Huang is the academic editor and an editor board member of *PLOS ONE*.



Changqin Huang (Member, IEEE) received the PhD degree in computer science from Zhejiang University, China, in 2005. He completed his visiting research with the University of California, Irvine, CA, USA, in 2011, and La Trobe University, Melbourne, VIC, Australia, in 2018. He is currently a tenured professor with Zhejiang University and distinguished professor with Zhejiang Normal University, and the director of the Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, China. He has authored or coauthored several papers in leading journals such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Knowledge and Data Engineering*, C&E, CHB, and BJET. His research interests include machine learning to Big Data in education and intelligent education. Dr. Huang is a Guangdong specially appointed professor (pearl river scholar) and an associate editor for the IEEE TLT.

Fan Jiang received the PhD degree in computer science and technology from Zhejiang Normal University, Jinhua, China. He is currently a lecturer with Guangdong Polytechnic Normal University, Guangzhou, China. His research interests include affective computing, computer vision, and AI in education.

Xiaodi Huang (Senior Member, IEEE) received the PhD degree in computer science. He is currently an associate professor with the School of Computing and Mathematics. His research interests include applied machine learning, visual data analysis, and computer applications. Dr. Huang has authored or coauthored more than 170 peer-reviewed papers in high-impact journals and premier conferences, including IEEE Transactions, ACM Transactions, and other leading venues. He is a member of ACM. He actively contributes to the academic community as an editorial board member for esteemed journals, the chair of international conferences, and a committee member and organizer for numerous global events. Dr. Huang is also the leader in innovative curriculum development in computing and promotes interdisciplinary collaboration through national and international research initiatives. He plays a role in shaping global academic and professional standards.

Xun Wang (Member, IEEE) is currently a professor, PhD supervisor, and the Dean of the School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou, China. He is also the director of Zhejiang Engineering Laboratory for Visual Media Big Data Technology. He has authored more than 80 papers in journals and conferences. His research interests include multimedia processing, computer vision, machine learning, computer graphics, and intelligent systems. He is a member of ACM and distinguished member of CCF.