Conditional Information Bottleneck for Multimodal Fusion: Overcoming Shortcut Learning in Sarcasm Detection

Yihua Wang[1, 2 *], **Qi Jia**[2], **Cong Xu**[2†], **Feiyu Chen**[1], **Yuhan Liu**[1], **Haotian Zhang**[1], **Liang Jin**[2], **Lu Liu**[2], **Zhichun Wang**[1‡]

[1]School of Artificial Intelligence, Beijing Normal University, Beijing, China
[2]IEIT SYSTEMS Co., Ltd., Beijing, China
wangyihua@mail.bnu.edu.cn, xucong@ieisystem.com, zcwang@bnu.edu.cn

Abstract

Multimodal sarcasm detection is a complex task that requires distinguishing subtle complementary signals across modalities while filtering out irrelevant information. Many advanced methods rely on learning shortcuts from datasets rather than extracting intended sarcasm-related features. However, our experiments show that shortcut learning impairs the model's generalization in real-world scenarios. Furthermore, we reveal the weaknesses of current modality fusion strategies for multimodal sarcasm detection through systematic experiments, highlighting the necessity of focusing on effective modality fusion for complex emotion recognition. To address these challenges, we construct MUStARD++^{R} by removing shortcut signals from MUStARD++. Then, a Multimodal Conditional Information Bottleneck (MCIB) model is introduced to enable efficient multimodal fusion for sarcasm detection. Experimental results show that the MCIB achieves the best performance without relying on shortcut learning.

Code — https://github.com/sljgkjhwe/MCIB.git

Datasets —
https://pan.quark.cn/s/975e0d976744#/list/share (3BFf)

Extended version —
https://doi.org/10.48550/arXiv.2508.10644

Introduction

Multimodal sentiment analysis (MSA) (Zadeh et al. 2017) integrates various modalities to interpret human emotions, sentiments, and opinions. In MSA tasks, multimodal sarcasm detection (Castro et al. 2019) is particularly challenging due to the subtle contrast between surface meaning and underlying intent, which may convey humor, ridicule, or contempt (Joshi, Bhattacharyya, and Carman 2017). See Figure 1, Sheldon's words express surprise, but his facial expression shows disgust, and his tone is mocking, conveying sarcasm. Despite advances in large language models, their performance remains capped by supervised baselines, falling short of human-level understanding (Hu et al. 2025).

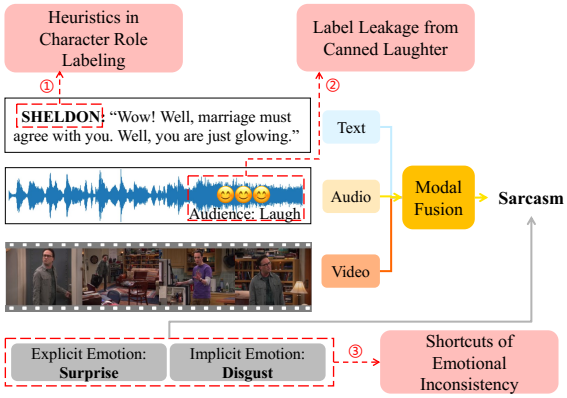

Figure 1: Multimodal sarcasm analysis often relies on sitcoms like Friends, which primarily feature character dialogues. The illustration presents the multimodal sarcasm detection task and highlights several shortcut learning issues.

Numerous studies have explored multimodal sarcasm detection (MSD) from various perspectives. Many focus on the inherent characteristics of sarcasm or are specifically designed for prior information: such as contrasting emotional polarity (Lu et al. 2024; Wang et al. 2024), or incorporate external knowledge (Yue et al. 2023), character profiles (Li et al. 2024), emotional cues (Gao, Nayak, and Coler 2024), and fine-grained visual features (Tiwari et al. 2023).

As Figure 1 shows, models tend to learn shortcuts (Geirhos et al. 2020) that lack generalization to diverse real-world scenarios. We identify three major shortcut issues in MSD research with the representative dataset MUStARD++ (Ray et al. 2022). (1) Heuristics in Character Role Labeling: In sarcasm datasets, some characters exhibit a speaking style that favors sarcasm, such as Sheldon in the TV show THE BIG BANG THEORY. Introducing these character labels during training causes the model to form a bias toward sarcasm associated with a specific person. (2) Label Leakage from Canned Laughter: Canned laughter (artificial background laughter) often causes label leakage. A review of MUStARD++ reveals that canned laughter frequently follows sarcastic utterances, but rarely non-sarcastic ones. Thus, the model tends to classify samples with canned

*Work completed while interning at IEIT SYSTEMS Co., Ltd.
†Corresponding Authors.
‡Corresponding Authors.

laughter as sarcasm. (3) Shortcuts of Emotional Inconsistency: The dataset provides implicit/explicit emotion labels, and their contrast serves as a strong indicator of sarcasm. Some models are designed to learn implicit/explicit emotions rather than true sarcasm, limiting their ability to detect sarcasm due to over-reliance on inconsistency.

The key to understanding the intended complex information and genuinely improving performance is addressing the core multimodal fusion problem, rather than designing modules based on shortcut learning. Various fusion methods (Ray et al. 2022; Zhang et al. 2024, 2023) do not result in significant information gains, some even reducing task performance, highlighting their poor generalization and adaptability to sarcasm tasks. Higher modal complementarity leads to poorer robustness when certain modes are absent or misaligned, and noise has a greater impact on modes with high complementarity (Li et al. 2023). MSD involves heterogeneous information with complementary but also irrelevant or misleading redundancy. Therefore, this underscores the need for an efficient fusion strategy that integrates complementary information while mitigating redundancy.

To this end, we construct MUStARD++^{R} by removing canned laughter, character labels, and emotional polarity from the dataset MUStARD++. Meanwhile, we propose a robust **M**ultimodal fusion framework with a **C**onditional **I**nformation **B**ottleneck for sarcasm detection, named **MCIB**. The MCIB achieves the state-of-the-art performance solely through effective multimodal fusion, without relying on additional shortcut cues. The MCIB offers two key functionalities. First, it addresses the limitations of traditional information bottlenecks by enabling the fusion of more than two information sources, overcoming the challenge of processing multiple modalities simultaneously. Second, our multimodal fusion strategy is designed to extract and separate inconsistent yet relevant information from each modality, effectively reducing inter-modal redundancy and preserving critical features for more accurate multimodal sarcasm recognition.

Our main contributions: (1) We conduct an in-depth analysis and provide a comprehensive summary of the critical issues in multimodal sarcasm detection, addressing the issue of learning shortcuts and refactoring the benchmark. (2) We rethink multimodal fusion from a novel perspective and propose a flexible and robust multimodal conditional information bottleneck fusion method. (3) We perform extensive experiments to demonstrate the performance of our fusion method, showing that the model significantly improves the efficiency of extracting meaningful information that aids in sarcasm detection across multiple modalities.

Related Work

Multimodal Sarcasm Detection has evolved from text and image-text (Wen, Jia, and Yang 2023) to video clips, which better reflect real-life scenarios. Many methods leverage sarcasm traits to aid recognition, such as incongruity (Lu et al. 2024). Song et al. (Song et al. 2024) propose an utterance-level attention and incongruity learning network to capture incongruity representations in sarcastic expressions. Since sarcasm is closely linked to emotions and sentiments, an adaptive representation learning model (Zhang et al. 2024) is proposed based on an expert network to analyze the emotion and sarcasm tasks jointly. Chauhan et al. (Chauhan et al. 2020) design a multitasking collaborative framework for cross-training and sharing attention between sarcasm, implicit/explicit emotions, and implicit/explicit sentiments. Additionally, Zhang et al. (Zhang et al. 2023) employ a dual-gating network and three separate layers to achieve cross-modal interactive learning. Other research attempts to leverage various external information. Studies such as (Liang et al. 2022) and (Yue et al. 2023) have utilized tools like KnowleNet to implement knowledge fusion networks for MSD. Moreover, works like (Tomar et al. 2024), (Tiwari et al. 2023), and (Pandey and Vishwakarma 2025) enhance sarcasm recognition by incorporating conversational behavioral cues, fine-grained eye gaze information, or deep visual features acquired through self-conditioning ConvNet, respectively. While (Gao, Nayak, and Coler 2024; Helal et al. 2024) rely on pre-training with explicit and implicit emotions, the best results from (Li et al. 2024) similarly through character labeling. Besides, some novel approaches (Liu, Zhang, and Song 2023; Tiwari et al. 2024) suggest that quantum mechanics are well-suited for capturing the complexity and uncertainty in sarcasm through a quantum probability-driven model or quantum fuzzy neural networks.

Information Bottleneck (IB) based approaches (Slonim and Tishby 1999; Alemi et al. 2022) have demonstrated their effectiveness in various tasks such as cross-modal clustering (Yan et al. 2023) and representation learning (Zhang et al. 2022a; Ding et al. 2023), and have also been widely applied in multimodal fusion. For example, Zhang et al. (Zhang et al. 2022b) improve MSA performance by applying IB constraints to pairs of modalities. Mai et al. (Mai, Zeng, and Hu 2022) ensure that latent modal representations can effectively handle the target task by introducing IB constraints between each modality and the predicted target. Chen et al. (Chen et al. 2023) adopt a similar approach, enhancing multimodal fusion by adding cross-attention. Furthermore, Liu et al. (Liu, Cao, and Zhang 2024) apply IB during the representation fusion stage to discard irrelevant information from individual modalities. Xiao et al. (Xiao et al. 2024) propose a two-layer IB structure that minimizes mutual information between input and latent states while maximizing it between latent and residual states. Moreover, conditional mutual information (Wyner 1978) has demonstrated strong generalization performance in various applications, such as feature selection (Fleuret 2004), modality enhancement (Ji et al. 2022), and multimodal selection (He et al. 2024). Gondek et al. (Gondek and Hofmann 2003) propose maximizing conditional mutual information to obtain relevant but novel clustering solutions, avoiding redundant information already captured by known structures or categories. Li et al. (Li et al. 2023) demonstrate through CIB calculations that higher modal complementarity reduces robustness while noise impacts highly complementary modes more severely. Thus, recognizing the potential of CIB (Molavipour, Bassi, and Skoglund 2021) for extracting relevant information in multimodal scenarios, we design a CIB-based method for multimodal fusion.

Task Analysis and Refactoring

In the MSD task, some shortcuts arise from task-specific traits, while others capture superficial cues or spurious correlations rather than the true intent of sarcasm. These easy-to-learn shortcut signals hinder the model's generalization to broader testing environments.

Heuristics in Character Role Labeling. Capturing character-specific traits can improve sarcasm detection, but methods relying on character association lack generalizability, as such features are often unavailable in real-world settings. We analyzed the proportion of sarcasm and non-sarcasm in the dialogues of different characters in the MUStARD++. The chi-square test is used to examine the relationship between character and sarcasm, with the null hypothesis stating no association. The test statistic of 166.7 and p-value of 3.89×10^{-27} provide strong evidence against the null hypothesis. The actors exhibited a distinct tendency towards either sarcastic or non-sarcastic behavior, suggesting that character traits strongly influence the judgment of sarcasm. The experimental findings (Castro et al. 2019; Li et al. 2024) reveal that the majority of models can enhance sarcasm recognition merely by incorporating character embeddings. However, the performance of such models trained with character embeddings will drop drastically when tested on other sentiment tasks. Thus, it is essential to design a robust model that captures the character's intended features rather than relying on character labels.

Label Leakage from Canned Laughter. Since most sarcasm datasets are derived from sitcoms, they typically include canned laughter. Canned laughter, a pre-recorded sound used to enhance humor or guide audience reactions, is also prevalent in MUStARD++. It often follows ironic or humorous utterances, leading to label leakage. Most studies (Arora et al. 2023; Tomar et al. 2023; Li et al. 2024) ignore the canned laughter while processing the data. Using the pre-trained speech recognition model SpeechPrompt v2(Chang et al. 2023), we validated on MUStARD++ and MUStARD++^{R} (with and without canned laughter). Compared to MUStARD++, performance on MUStARD++^{R} shows a significant drop, with F1-score decreasing from 73.47 to 43.59 (a drop of 29.88) and accuracy from 78.33 to 63.03 (a drop of 15.3). This suggests that the model learned the features associated with canned laughter, indirectly confirming the label leakage issue.

Shortcuts of Emotional Inconsistency. Sarcasm is often a sugar-coated bomb, expressing the opposite of its literal meaning by masking implicit emotions with explicit ones. The statistical results from the MUStARD++ dataset reveal that sarcasm is predominantly associated with different emotions, accounting for 99% of cases, while only 1% is linked to the same emotions. In contrast, non-sarcasm is mainly associated with the same emotions (94.7%), with only 5.3% corresponding to different emotions. This indicates that sarcastic sentences often exhibit discrepancies between explicit and implicit sentiment labels. The Phi coefficient analysis supports this observation, revealing a very strong association between emotional consistency and sarcasm ($\phi = 0.94$). Many studies (Chauhan et al. 2020; Shah, Reddy, and Bhattacharyya 2022; Helal et al. 2024) use this pattern to detect sarcasm. But in real-world scenarios, the explicit/implicit emotion labels do not exist. Moreover, inferring explicit/implicit emotion labels may be more challenging than directly detecting sarcasm. These limitations suggest that relying on explicit/implicit emotion labels for MSD is unfeasible.

To avoid relying on shortcuts and improve the fairness, robustness, trustworthiness, and deployability of the evaluation benchmark, based on the above analysis of shortcut learning in multimodal sarcasm tasks, we reconstructed MUStARD++^{R} from the publicly available dataset MUStARD++. Each data sample is annotated with sarcasm labels, implicit and explicit emotions, arousal, and valence tags. The first step in modifying the dataset is to remove all shortcut labels, leaving only those related to sarcasm. Next, we eliminated video segments that contain canned laughter. Specifically, we used the timestamp of the first word in each utterance as the start of the video segment and the timestamp of the last word as the endpoint. By removing labels linked to shortcut learning, MUStARD++^{R} allows for a more accurate evaluation of multimodal fusion models in detecting sarcasm through integrated cross-modal information.

Methodology

This section defines the formal optimization objective for multimodal fusion, rethinks existing fusion methods, and introduces our framework, including the MCIB algorithm and overall model architecture.

Notations. We define the modalities audio, video, and text as $\{x_0, x_1, x_2\}$, which are interdependent random variables. The redundant information causing interference among modalities is denoted by R, while the information gain from modality complementarity is represented by C. y denotes the ground truth of the target task.

Problem formulation. Our objective is to minimize the loss between the fused multimodal information and the target y. As shown in eq. (1), we aim to reduce redundancy R and maximize the utilization of complementarity C between modalities:

$$\begin{aligned} \min_{Fusion} \quad & Loss\left(y, f(x_0, x_1, x_2)\right), \\ \text{subject to} \quad & R(x_0, x_1, x_2; y) \leq \delta \\ & C(x_0, x_1, x_2; y) \geq \epsilon \end{aligned} \tag{1}$$

where δ sets the upper limit for allowable redundancy, and ϵ defines the minimum threshold for the required level of complementarity utilization among modalities.

Modal Fusion Rethink. We rethink the limited effectiveness of previous fusion methods in light of the complementary and redundant properties of multimodal data. Ablation tests on various combinations of modality reveal limited gains from modality fusion. In some cases, adding modalities reduces performance: audio may decrease overall effectiveness. Figure 3 illustrates the reasons for ineffective modality fusion: the Redundancy region (which refers not only to "repeated information" that may be useless for the object but also to negative interference in modality fusion) is larger than the Complementarity region, and the overlapping middle region is too small. This suggests that misleading redundant information outweighs the informational gain

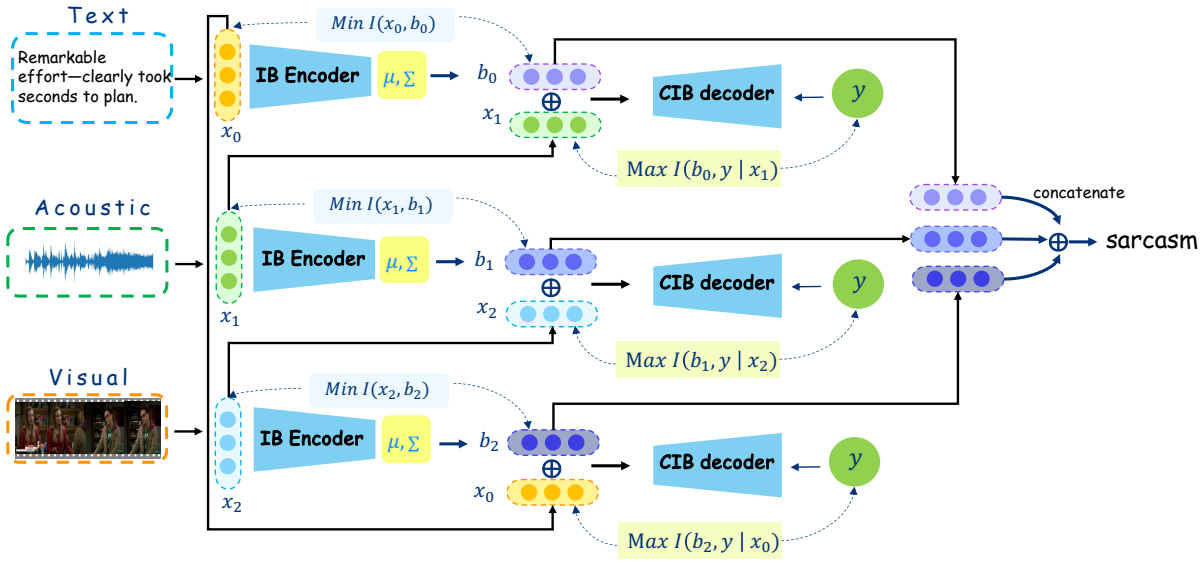

Figure 2: The diagram illustrates the overall architecture of the MCIB model. The multimodal fusion component employs three parallel conditional information bottleneck structures to filter out irrelevant information and extract relevant information between each pair of modalities. For each pair of modalities, we first minimize the mutual information between the primary modality and the latent state to achieve filtering and compression through the IB. We then maximize the conditional mutual information among the auxiliary modality, latent state, and prediction target. Finally, the bidirectional optimization within CIB produces an intermediate representation b that encapsulates the essential information required for our prediction target.

from the added modality. For effective modality addition, irrelevant information that may mislead predictions should be reduced, while enhancing the benefits of complementary useful information.

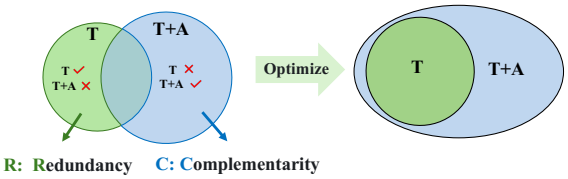

Figure 3: The diagram illustrates the optimization direction for multimodal fusion. The two circles represent correctly predicted samples by the Text modality alone and by the Text + Acoustic combination. The overlapping region shows samples correctly predicted by both T and T + A; region R represents samples correctly predicted by T but misclassified when A is added; and region C represents samples only correctly predicted with T + A, while T alone fails.

Our Framework

To extract complementary information and remove redundancy across modalities, we apply the CIB principle to build a multimodal fusion model (see Fig. 2). We first extract detailed modality features, then design a fusion algorithm based on multimodal pairwise strategy. Compact cross-modal representations are optimized to enhance target understanding, and a balanced, robust loss function with Lagrange constraints ensures effective model training.

MCIB Algorithm To achieve the multimodal fusion objective in eq. (1), we distinguish between the primary modality x_p and the auxiliary modality x_a, where $x_p, x_a \in \{x_0, x_1, x_2\}$. Only the information in the primary modality x_p beneficial to the target is preserved, while the auxiliary modality x_a provides complementary information to x_p. We aim to encode as much task-related (y) information as possible in the latent state b, where $b \in \{b_0, b_1, b_2\}$ represents generation through optimization of the conditional information bottleneck. As shown in Figure 4, the trade-off between compressing redundant information from the primary modality and retaining complementary information provided by the auxiliary modality is optimized by minimizing mutual information while adhering to specific constraints. We reformulate this constrained optimization problem into an unconstrained Lagrangian form eq. (2), known as the conditional information bottleneck:

$$\min_{p(b|x_p,x_a)} \underbrace{I(x_p;b)}_{\text{Compress redundancy}} - \lambda \underbrace{I(b;y \mid x_a)}_{\text{Retain complementarity}}, \quad (2)$$

where $I(x_p; b)$ encourages compression of redundancy from x_p. $I(b; y \mid x_a)$ ensures b retains complementary information useful for predicting y given x_a. λ balances the trade-off between compression and retention.

Compress redundancy. This term penalizes the mutual information between the primary modality x_p and the latent state b, encouraging b to be a compressed representation of

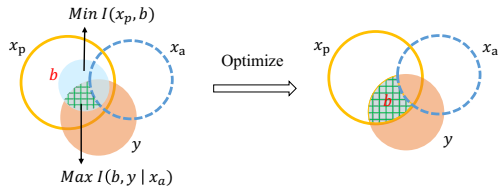

Figure 4: The optimization process of the multimodal conditional information bottleneck is shown. The left is the initial state, where the blue section represents the latent state b generated from the primary modality, and the green cells denote the conditional mutual information between the auxiliary modality, latent state, and target. The right is the ideal state: b contains all information relevant to the target y, free from redundancy, and integrates complementary information from the primary modality concerning the auxiliary modality.

x_p. The mutual information between x_p and b is:

$$I(x_p; b) = \int p(x_p, b) \log \frac{p(b \mid x_p)}{p(b)} \mathrm{d}x_p \mathrm{d}b. \tag{3}$$

Computing the marginal distribution $p(b)$ is intractable, so we introduce a variational prior $r(b)$ to approximate $p(b)$. Assuming $p(b \mid x_p)$ is approximated by a variational distribution $q(b \mid x_p)$, the upper bound becomes:

$$I(x_p; b) \leq \mathbb{E}_{p(x_p)} \left[D_{\mathrm{KL}}(q(b \mid x_p) \| r(b)) \right]. \tag{4}$$

We define the loss term for compression as the KL divergence between $q(b \mid x_p)$ and $r(b)$. Specifically, in designing the optimizer: choose a standard Gaussian $\mathcal{N}(0, I)$ as Prior $r(b)$, while utilizing a transformer architecture to capture detailed and relevant information from x_p for simulating $q(b \mid x_p)$. The encoder models $q(b \mid x_p)$ as a Gaussian distribution $\mathcal{N}(\mu(x_p), \sigma^2(x_p))$. The latent variable b is obtained by reparameterizing and sampling from $q(b \mid x_p)$. The KL divergence is calculated as:

$$D_{\mathrm{KL}}(q(b \mid x_p) \| r(b)) = \frac{1}{2} \sum_{i=1}^{d} (\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2). \tag{5}$$

eq. (5) encourages b to gather the most essential information from x_p, effectively compressing the redundancy in x_p.

Retain complementarity. This term encourages b to retain information about y that is complementary to x_a. The conditional mutual information is defined as:

$$I(b; y \mid x_a) = \int p(x_a, b, y) \log \frac{p(y \mid b, x_a)}{p(y \mid x_a)} \mathrm{d}x_a \mathrm{d}b \mathrm{d}y. \tag{6}$$

Direct computation is intractable. Introducing a variational distribution $q(y \mid b, x_a)$, we obtain:

$$\begin{aligned} I(b; y \mid x_a) &= \mathbb{E}_{p(x_a, b, y)} \left[\log \frac{p(y \mid b, x_a)}{p(y \mid x_a)} \right] \\ &\geq \mathbb{E}_{p(x_a, b, y)} [\log q(y \mid b, x_a) - \log p(y \mid x_a)]. \end{aligned} \tag{7}$$

Ignoring the constant term $\log p(y \mid x_a)$ (since it does not depend on b), we get the lower bound:

$$I(b; y \mid x_a) \geq \mathbb{E}_{p(x_a, b, y)} \left[\log q(y \mid b, x_a) \right]. \tag{8}$$

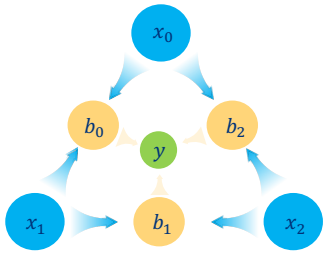

Figure 5: By constructing three latent state b_0, b_1 and b_2, pertinent information transfer between the three modal x_0, x_1 and x_2 is facilitated. Finally, integrated data leads to the prediction of y.

We define the loss term for retention as the negative expected log-likelihood. Then the Evidence Lower Bound Objective (ELBO) method is adopted for the variational approximation of $q(y \mid b, x_a)$. Combining b and x_a, we construct a transformer-based neural network estimator to model $q(y \mid b, x_a)$. The expected log-likelihood is approximated using samples $b^{(l)}$ drawn from $q(b \mid x_p)$:

$$\mathbb{E}_{p(x_a, b, y)} [\log q(y \mid b, x_a)] \approx \frac{1}{L} \sum_{l=1}^{L} \log q(y \mid b^{(l)}, x_a), \tag{9}$$

where L is the number of samples.

This loss encourages the model to reconstruct y from the combined b and x_a, maximizing the conditional mutual information $I(b; y \mid x_a)$. By precisely formulating each step and keeping the derivation concise, we establish a tractable lower bound for $I(b; y \mid x_a)$ using variational inference and ELBO methods, which effectively integrates the auxiliary modality x_a with the latent representation b to enhance predictive performance. The overall loss function is:

$$\begin{aligned} \mathcal{L} = &\frac{1}{2} \sum_{i=1}^{d} (\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2) \\ &- \lambda \frac{1}{L} \sum_{l=1}^{L} \log q(y \mid b^{(l)}, x_a). \end{aligned} \tag{10}$$

Overall Architecture and Optimization To uncover latent information within each modality, we design a module for fine-grained feature extraction. Using GENTLE, we align audio and segment audio at the word level, which then serves as a reference for aligning fine-grained features in the visual modality. Next is the multimodal feature fusion module MCIB, which integrates complementary information across modalities while filtering out redundancy. Specifically, the MCIB minimizes the mutual information between the primary modality x_p and the latent state b, condensing x_p and filtering out redundant information. Simultaneously, it maximizes the conditional mutual information between the auxiliary modality x_a, the latent state b, and the target y, aiming to incorporate additional relevant information from x_a so that b holds useful information for predicting y. The model concatenates the trained latent state b for prediction, which contains the "redundancy-removed, effective

complementarity" information distilled through the conditional information bottleneck.

As shown in Figure 5, in the context of multimodal learning, we alternately designate each of the three modalities as the primary and auxiliary modalities to train the fusion framework based on the MCIB algorithm jointly. To better control the degree of information compression, we introduce modality-specific hyperparameters λ_0, λ_1, and λ_2 to balance the conditional information bottleneck loss. Letting 0, 1, and 2 represent the three modalities, their respective loss functions are defined as eq. (11):

$$\begin{aligned} \mathcal{L}_0 &= \mathcal{L}_{\text{IB}_0} + \lambda_0 \mathcal{L}_{\text{CIB}_2}, \\ \mathcal{L}_1 &= \mathcal{L}_{\text{IB}_1} + \lambda_1 \mathcal{L}_{\text{CIB}_0}, \\ \mathcal{L}_2 &= \mathcal{L}_{\text{IB}_2} + \lambda_2 \mathcal{L}_{\text{CIB}_1}. \end{aligned} \tag{11}$$

Furthermore, to fully exploit the fused latent state b in the conditional information bottleneck, we introduce a prediction loss $\mathcal{L}_{\text{pred}}$ from b to y, where α_0, α_1, α_2 and β are the weighting coefficients. The model is trained with the following final objective: $\mathcal{L}_{\text{total}} = \alpha_0 \mathcal{L}_0 + \alpha_1 \mathcal{L}_1 + \alpha_2 \mathcal{L}_2 + \beta \mathcal{L}_{\text{pred}}$.

Experiments

Dataset and Evaluation Metrics. Experiments were conducted on MUStARD++ and MUStARD++^{R}, respectively. Given the balanced class distribution, we report results using weighted precision, weighted recall, and weighted F1-score metrics. In addition, we tested the MCIB method on MSA tasks (CMU-MOSI and CMU-MOSEI in the Extended version), achieving highly competitive results, which validate its generalization capability.

Implementation Details. To extract fine-grained features, we utilize several high-performance backbones. Text features are obtained using a pre-trained DeBERTa (He, Gao, and Chen 2021), yielding a representation size of $d_t = 768$. For audio features, Mel Frequency Cepstral Coefficients (MFCC) and Mel spectrograms are generated with Librosa (McFee et al. 2024), along with prosodic features from OpenSMILE 3 (Eyben, Wöllmer, and Schuller 2010), resulting in a combined representation of $d_a = 291$. The video features are extracted by processing utterance frames through the pool5 layer of a ResNet-152 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009), producing a visual representation of size $d_v = 2048$. All experiments were conducted on Nvidia A100 GPUs with 40GB of memory. Multiple trials demonstrate that combinations of random hyperparameters within $[1, 64]$ typically result in local optima after 10 searches. The hyperparameter analysis reveals that the most significant interactions occur between λ_2 & α_1, λ_1 & α_1, and α_0 & α_1. The stability analysis highlighted that α_1 demonstrated the best stability. Experimental results were averaged over five runs.

Comparison with baseline models

We compare our proposed MCIB with the following baselines on multimodal fusion and sentiment analysis. VyAnG-Net (Pandey and Vishwakarma 2025) enhances sarcasm detection by integrating visual-specific attention mechanisms with text captions, while PredGaze (Tiwari et al. 2023) improves sarcasm recognition by utilizing fine-grained visual information such as eye gaze. FIES (Gao, Nayak, and Coler 2024) proposes a multimodal approach that integrates audio, textual, sentiment, and emotion data to enhance sarcasm detection, while ABCA-IMI (Li et al. 2024) identifies sarcasm through multiple inconsistency detection mechanisms. Some methods did not release their code; for instance, the VyAnG-Net results were obtained by training on MUStARD and validating on MUStARD++. SIB (Mai, Zeng, and Hu 2022; Chen et al. 2023) using the IB between each single modality and the target, while DIB (Zhang et al. 2022b) applies the IB to pairs of modalities, enabling back-optimization. ITHP (Xiao et al. 2024) was designed with a two-layer IB guide to the modality information flow. TBJE (Delbrouck et al. 2020) is a cross-attention-based model with high generalization for multimodal fusion. MUStARD (Ray et al. 2022) leverages a collaborative gating mechanism for sarcasm and emotion recognition. Moreover, we report baseline results for the pre-trained speech model SpeechPrompt v2 (Chang et al. 2023), and leading

Method	Precision	Recall	F1-score
PredGaze$^{○}$	71.06	71.06	71.06
VyAnG-Net$^{○}$	72.41	72.05	72.23
FIES$^{○}$	71.48	71.43	71.45
FIES$^{○♡}$	74.31	74.29	74.29
ABCA-IMI$^{○}$	71.90	69.90	70.90
ABCA-IMI$^{○◇}$	76.20	74.20	75.20
SIB$^{○}$	69.67	69.54	69.39
SIB (w/o shortcuts)	68.30	68.26	68.21
DIB$^{○}$	70.96	71.12	70.98
DIB (w/o shortcuts)	70.44	70.43	70.42
ITHP $^{○◇}$	71.39	70.95	70.93
ITHP (w/o shortcuts)	68.27	68.29	68.27
TBJE$^{○♡}$	71.79	71.90	71.82
TBJE (w/o shortcuts)	69.50	68.59	68.75
MUStARD $^{○}$	71.02	70.80	70.98
MUStARD $^{○◇}$	71.38	71.38	71.38
MUStARD (w/o shortcuts)	71.07	70.03	70.03
SpeechPrompt v2$^{○}$	78.33	58.06	73.47
SpeechPrompt v2 (w/o shortcuts)	63.03	27.87	43.59
GPT-4o$^{◇}$	62.66	83.90	71.74
GPT-4o (w/o shortcuts)	67.11	85.47	75.19
Gemini 2.5$^{◇}$	62.89	84.75	72.20
Gemini 2.5 (w/o shortcuts)	69.12	78.99	73.73
MCIB$^{○}$	77.18	76.30	76.85
MCIB$^{○◇}$	76.34	75.84	75.75
MCIB (w/o shortcuts)	76.14	75.83	75.64

Table 1: Performance comparison of baseline methods on MUStARD++ and MUStARD++^{R}. Superscripts denote the use of specific shortcuts in MUStARD++: ○ for canned laughter, ◇ for character, and ♡ for emotion. Results labeled "(w/o shortcuts)" correspond to our clean dataset MUStARD++^{R}.

large language models: OpenAI's latest flagship model GPT-4o (OpenAI: Aaron Hurst and et al. 2024), the newest release from Google DeepMind Gemini 2.5 (Gheorghe Comanici and et al. 2025), under multimodal configurations.

As shown in Table 1, our approach achieved the highest F1 scores of 76.85% and 75.64% on the MUStARD++ and MUStARD++^{R} datasets, respectively, outperforming all baseline methods. The results indicate two aspects: the model's reliance on shortcuts (generalization ability) and its capacity to capture truly useful information (effectiveness of multimodal fusion).

First, our method MCIB achieves strong performance without relying on tricks or dataset-specific shortcuts, demonstrating robust generalization across different data conditions. Comparing results on MUStARD++ and MUStARD++^{R}, we observe that, after shortcut cues (such as character labels, canned laughter, or emotional inconsistency) are removed, the performance of most conventional methods drops to varying degrees. Interestingly, GPT-4o and Gemini 2.5 perform better in the absence of these cues, possibly because character information introduces noise that misleads LLMs. These findings indicate that many methods are sensitive to shortcuts, while MCIB remains robust.

Second, using MUStARD++^{R}, models cannot rely on shortcuts and must depend on architecture design and multimodal fusion strategies. The results demonstrate that MCIB neither relies on nor overfits to shortcut cues and achieves the most effective fusion strategy. In real-world scenarios with limited auxiliary information, our approach outperforms other multimodal fusion methods by more effectively integrating information across modalities.

Additionally, we conducted cross-training experiments between MUStARD++ and MUStARD++^{R} to further investigate the impact of shortcut learning (see Extended version).

Modality and Module Ablation Studies

This section examines modality fusion in MCIB, the impact of the transformer and fine-grained modules, and results from modality ablation and different modality combinations.

Method	Precision	Recall	F1-score
w/o Transformer	75.02	74.58	74.32
w/o Fine-Gained	71.65	71.21	71.19
x_v	69.98	70.00	69.99
x_a	70.43	69.16	68.97
x_t	70.60	70.83	70.98
x_{va}	72.06	72.08	72.04
x_{at}	73.75	73.75	73.69
x_{tv}	74.92	74.17	73.77
$x_{vt} + x_{av} + x_{ta}$	75.42	74.17	74.03
$x_{va} + x_{at} + x_{tv}$	76.14	75.83	75.64

Table 2: Ablation results: using an MLP (w/o Transformer), coarse-grained features (w/o Fine-Grained), modality removal, and varying modality pairs.

Transformer architecture. Our transformer-based encoder captures richer patterns than the MLP, yielding a 1.32% performance gain despite increased computational cost.

Fine-gained module ablation. We compare fine-grained and coarse-grained feature extraction to investigate their impact on multimodal tasks. The results show that aligning and representing the three modalities at the word level with fine-grained features enhances sarcasm detection.

Modal ablation. We conduct experiments with all three modalities. Among single modalities, text performs best, while vision performs worst. For dual modalities, performance is highest when text is primary and vision is auxiliary, and lowest when video is primary and audio is auxiliary.

Varying primary and auxiliary modality pairs. The three modalities yield six possible sequential pairs. To maximize complementary information, we prioritize combinations that include all modalities. This results in two configurations: $va + at + tv$ and $av + vt + ta$ (where the first modality is primary, and the second is auxiliary). Experimental results indicate that sarcasm detection performs best when visual assists the audio modality, audio assists text, and text assists the visual modality.

Visualizations

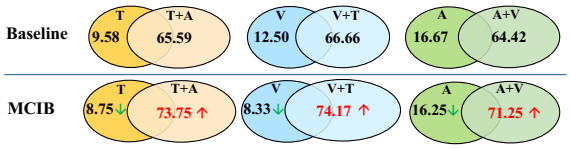

Figure 6: The diagrams illustrate the loss-benefit comparison for adding new modalities to visual, text, and audio modalities. In each Venn diagram, the numbers in each section represent the proportion of correctly predicted samples relative to the total test set.

Figure 6 provides an intuitive comparison of the effectiveness of the proposed MCIB method in modality fusion relative to baseline models. The directional trends of the arrows indicate that MCIB optimizes in the correct direction for modality fusion. Observing the overall improvement: misclassifications due to redundant or irrelevant information are significantly reduced across all modality combinations, while complementary information between modalities is effectively utilized. The accuracy of joint predictions from added modalities has also improved.

Conclusion

This paper first analyzes the shortcut learning problem existing in current multimodal sarcasm detection methods and restructures the task dataset to prevent the learning of shortcut features. Subsequently, we propose a multimodal conditional information bottleneck (MCIB) framework that effectively captures complementary inter-modal information while filtering out irrelevant and misleading redundancies. Extensive experimental results demonstrate that our model uses multimodal data more effectively, achieving state-of-the-art performance on the MSD task. In the future, we aim to refine MCIB into an easily integrable plug-in for various backbone models in multimodal sentiment analysis, thereby boosting their performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62276026).

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2022. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*.

Arora, S.; Futami, H.; Jung, J.-w.; Peng, Y.; Sharma, R.; Kashiwagi, Y.; Tsunoo, E.; and Watanabe, S. 2023. Universlu: Universal spoken language understanding for diverse classification and sequence generation tasks with a single network. *arXiv preprint arXiv:2310.02973*.

Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; and Poria, S. 2019. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4619–4629.

Chang, K.-W.; Wang, Y.-K.; Shen, H.; Kang, I.-t.; Tseng, W.-C.; Li, S.-W.; and Lee, H.-y. 2023. Speechprompt v2: Prompt tuning for speech classification tasks. *arXiv preprint arXiv:2303.00733*.

Chauhan, D. S.; Dhanush, S.; Ekbal, A.; and Bhattacharyya, P. 2020. Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 4351–4360.

Chen, X.; Wu, Z.; Tang, Y.; and Han, R. 2023. Multimodal Sentiment Analysis Based on Information Bottleneck and Attention Mechanisms. In *2023 2nd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, 150–156. IEEE.

Delbrouck, J.-B.; Tits, N.; Brousmiche, M.; and Dupont, S. 2020. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. *ACL 2020*, 1.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ding, S.; Du, W.; Ding, L.; Zhang, J.; Guo, L.; and An, B. 2023. Robust Multi-Agent Communication With Graph Information Bottleneck Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462.

Fleuret, F. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 5(9).

Gao, X.; Nayak, S.; and Coler, M. 2024. Improving sarcasm detection from speech and text through attention-based fusion exploiting the interplay of emotions and sentiments. In *Proceedings of Meetings on Acoustics*, volume 54, 060002. Acoustical Society of America.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Gheorghe Comanici, E. B.; and et al., M. S. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Gondek, D.; and Hofmann, T. 2003. Conditional information bottleneck clustering. In *3rd ieee international conference on data mining, workshop on clustering large data sets*, 36–42.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

He, Y.; Cheng, R.; Balasubramaniam, G.; Tsai, Y.-H. H.; and Zhao, H. 2024. Efficient Modality Selection in Multimodal Learning. *Journal of Machine Learning Research*, 25(47): 1–39.

Helal, N. A.; Hassan, A.; Badr, N. L.; and Afify, Y. M. 2024. A contextual-based approach for sarcasm detection. *Scientific Reports*, 14(1): 15415.

Hu, H.; Zhou, Y.; You, L.; Xu, H.; Wang, Q.; Lian, Z.; Yu, F. R.; Ma, F.; and Cui, L. 2025. EmoBench-M: Benchmarking Emotional Intelligence for Multimodal Large Language Models. arXiv:2502.04424.

Ji, B.; Zhang, T.; Zou, Y.; Hu, B.; and Shen, S. 2022. Increasing visual awareness in multimodal neural machine translation from an information theoretic perspective. *arXiv preprint arXiv:2210.08478*.

Joshi, A.; Bhattacharyya, P.; and Carman, M. J. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5): 1–22.

Li, S.; Du, C.; Huang, Y.; Huang, L.; and Zhao, H. 2023. Modality Complementariness: Towards Understanding Multi-modal Robustness.

Li, Y.; Li, Y.; Zhang, S.; Liu, G.; Chen, Y.; Shang, R.; and Jiao, L. 2024. An attention-based, context-aware multimodal fusion method for sarcasm detection using inter-modality inconsistency. *Knowledge-Based Systems*, 287: 111457.

Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; and Xu, R. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1767–1777. Association for Computational Linguistics.

Liu, W.; Cao, S.; and Zhang, S. 2024. Multimodal consistency-specificity fusion based on information bottleneck for sentiment analysis. *Journal of King Saud University-Computer and Information Sciences*, 36(2): 101943.

Liu, Y.; Zhang, Y.; and Song, D. 2023. A quantum probability driven framework for joint multi-modal sarcasm, sentiment and emotion analysis. *IEEE Transactions on Affective Computing*, 15(1): 326–341.

Lu, Q.; Long, Y.; Sun, X.; Feng, J.; and Zhang, H. 2024. Fact-sentiment incongruity combination network for multimodal sarcasm detection. *Information Fusion*, 104: 102203.

Mai, S.; Zeng, Y.; and Hu, H. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25: 4121–4134.

McFee, B.; McVicar, M.; Faronbi, D.; Roman, I.; Gover, M.; Balke, S.; Seyfarth, S.; Malek, A.; Raffel, C.; Lostanlen, V.; van Niekirk, B.; Lee, D.; Cwitkowitz, F.; Zalkow, F.; Nieto, O.; Ellis, D.; Mason, J.; Lee, K.; Steers, B.; and Südholt, D. 2024. librosa/librosa: 0.11.0. Open-source software library for audio and music analysis. Zenodo. Accessed: 2025-12-23.

Molavipour, S.; Bassi, G.; and Skoglund, M. 2021. Neural estimators for conditional mutual information using nearest neighbors sampling. *IEEE transactions on signal processing*, 69: 766–780.

OpenAI: Aaron Hurst, A. L.; and et al., A. P. G. 2024. GPT-4o System Card. arXiv:2410.21276.

Pandey, A.; and Vishwakarma, D. K. 2025. VyAnG-Net: A novel multi-modal sarcasm recognition model by uncovering visual, acoustic and glossary features. *Intelligent Data Analysis*, 1088467X251315637.

Ray, A.; Mishra, S.; Nunna, A.; and Bhattacharyya, P. 2022. A Multimodal Corpus for Emotion Recognition in Sarcasm. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6992–7003.

Shah, S.; Reddy, S.; and Bhattacharyya, P. 2022. Emotion enriched retrofitted word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, 4136–4148.

Slonim, N.; and Tishby, N. 1999. Agglomerative information bottleneck. *Advances in neural information processing systems*, 12.

Song, L.; Zhao, Z.; Ma, Y.; Liu, Y.; and Li, J. 2024. Utterance-Level Incongruity Learning Network for Multimodal Sarcasm Detection. In *2024 26th International Conference on Advanced Communications Technology (ICACT)*, 43–49. IEEE.

Tiwari, D.; Kanojia, D.; Ray, A.; Nunna, A.; and Bhattacharyya, P. 2023. Predict and Use: Harnessing Predicted Gaze to Improve Multimodal Sarcasm Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15933–15948.

Tiwari, P.; Zhang, L.; Qu, Z.; and Muhammad, G. 2024. Quantum fuzzy neural network for multimodal sentiment and sarcasm detection. *Information Fusion*, 103: 102085.

Tomar, M.; Tiwari, A.; Saha, T.; and Saha, S. 2023. Your tone speaks louder than your face! Modality Order Infused Multi-modal Sarcasm Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3926–3933.

Tomar, M. S.; Saha, T.; Tiwari, A.; and Saha, S. 2024. Action and Reaction Go Hand in Hand! a Multi-modal Dialogue Act Aided Sarcasm Identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 298–309.

Wang, J.; Yang, Y.; Jiang, Y.; Ma, M.; Xie, Z.; and Li, T. 2024. Cross-modal incongruity aligning and collaborating for multi-modal sarcasm detection. *Information Fusion*, 103: 102132.

Wen, C.; Jia, G.; and Yang, J. 2023. Dip: Dual incongruity perceiving network for sarcasm detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2540–2550.

Wyner, A. D. 1978. A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38(1): 51–59.

Xiao, X.; Liu, G.; Gupta, G.; Cao, D.; Li, S.; Li, Y.; Fang, T.; Cheng, M.; and Bogdan, P. 2024. Neuro-Inspired Information-Theoretic Hierarchical Perception for Multimodal Learning. arXiv:2404.09403.

Yan, X.; Mao, Y.; Ye, Y.; and Yu, H. 2023. Cross-modal clustering with deep correlated information bottleneck method. *IEEE Transactions on Neural Networks and Learning Systems*.

Yue, T.; Mao, R.; Wang, H.; Hu, Z.; and Cambria, E. 2023. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion*, 100: 101921.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114.

Zhang, T.; Dong, C.; Su, J.; Zhang, H.; and Li, Y. 2022a. Unimodal and Multimodal Integrated Representation Learning via Improved Information Bottleneck for Multimodal Sentiment Analysis. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 564–576. Springer.

Zhang, T.; Zhang, H.; Xiang, S.; and Wu, T. 2022b. Information Bottleneck based Representation Learning for Multimodal Sentiment Analysis. In *Proceedings of the 6th International Conference on Control Engineering and Artificial Intelligence*, 7–11.

Zhang, Y.; Yu, Y.; Wang, M.; Huang, M.; and Hossain, M. S. 2024. Self-Adaptive Representation Learning Model for Multi-Modal Sentiment and Sarcasm Joint Analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5): 1–17.

Zhang, Y.; Yu, Y.; Zhao, D.; Li, Z.; Wang, B.; Hou, Y.; Tiwari, P.; and Qin, J. 2023. Learning multi-task commonness and uniqueness for multi-modal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence*.