CIB-MoE: Counterfactual Inconsistency-Bottleneck Mixture-of-Experts for Robust Multimodal Aspect-based Sentiment Analysis and Sarcasm Detection

Anonymous ACL submission

Abstract

Multimodal posts on social media pose a fine-grained affective understanding challenge: the decisive signal often lies in instance-specific discrepancies between text and image, yet models are easily misled by weak cross-modal relevance, heterogeneous mismatch types (semantic, entity-level, and affective), and spurious lexical shortcuts. These issues are central to multimodal aspect-based sentiment analysis (MABSA), which demands aspect-conditioned predictions under noisy visual context, and multimodal sarcasm detection (MMSD), where sarcasm is frequently expressed through cross-modal incongruity rather than surface polarity. We propose **CIB-MoE** (Counterfactual Inconsistency-Bottleneck Mixture-of-Experts), a unified framework that performs discrepancy-aware conditional computation instead of monolithic fusion. CIB-MoE builds lightweight difference experts that quantify complementary mismatch cues—e.g., CLIP-based semantic inconsistency and entity-level misalignment derived from Top-N predicted object labels—and routes them through a two-level gate with an information-bottleneck regularizer for sparse and stable expert usage. To further suppress shortcut-driven routing, we calibrate the gate with realizable counterfactual interventions by substituting the image with neutral (text-aligned) and random (noise) alternatives and imposing ranking/consistency constraints on routing and predictions. Experiments on Twitter-2015/2017 and MMSD/MMSD2.0 show that CIB-MoE achieves state-of-the-art performance while improving robustness under distribution shift and counterfactual evaluation.

1 Introduction

Social media has turned sentiment expression from standalone text into multimodal communication, where users routinely combine language and imagery to convey nuanced and often non-literal meanings. This trend has motivated two fine-grained tasks: **Multimodal Aspect-Based Sentiment Analysis (MABSA)** and **Multimodal Sarcasm Detection (MMSD)**. MABSA predicts the sentiment polarity toward a queried aspect in a post, whereas MMSD identifies sarcasm that arises from an intentional mismatch between literal wording and intended meaning. Despite different objectives, both tasks hinge on a cognitive requirement: **detecting and interpreting cross-modal incongruity**.

Figure 1: Example of MABSA and MMSD tasks.

The key challenge is that the decisive signal seldom resides in a single modality. Instead, it emerges from subtle discrepancies between text and image (semantic, affective, or referential). Early methods largely relied on global feature fusion (Cai et al., 2019; Xu et al., 2020), which struggles to isolate localized, aspect-specific conflicts. Recent work has moved toward structure-aware modeling, including ambiguity-aware multi-level decompositions (Lu et al., 2024; Li et al., 2025) and dynamic routing mechanisms (Tian et al., 2023; Guan et al., 2025), aiming to capture fine-grained incongruity more explicitly.

Figure 1 illustrates that cross-modal interaction is inherently fine-grained and component-dependent. In the MABSA example, the same post expresses different polarities for different aspects (Iniesta: neutral; Real Madrid: negative; La Liga: neutral), while the paired image offers only coarse context and can induce aspect-agnostic shortcuts.

In the MMSD example, "lovely weather" conflicts with a rainy scene, where surface polarity cues become unreliable and sarcasm is triggered by text–image contrast. These cases motivate **CIB-MoE**: a unified framework that detects whether an instance enters a conflict mode and sparsely routes to discrepancy experts under an information-bottleneck gate, calibrated by counterfactual interventions to suppress spurious correlations.

Despite this progress, state-of-the-art models still exhibit three limitations. (i) Many approaches reduce multimodal interaction to implicit, over-parameterized fusion, without categorizing the underlying discrepancy (e.g., semantic vs. entity-level). (ii) Models often exploit dataset-specific shortcuts—especially lexical sentiment cues—rather than genuine cross-modal conflicts, a weakness highlighted by MMSD2.0 (Qin et al., 2023). (iii) Modality imbalance remains prevalent: dominant textual signals can overwhelm visual evidence, leading to biased predictions (Jia et al., 2024; Zhao et al., 2025). Although causal-inspired approaches attempt to mitigate such biases (Zhu et al., 2024b; Wu et al., 2025), their predefined and shallow structures can be insufficient to adapt to diverse discrepancy patterns.

These observations raise the following question:

> *Can we design a unified framework that (i) explicitly models diverse types of cross-modal differences through specialized expertise, while (ii) dynamically routes information through a calibrated gate that suppresses spurious shortcuts via counterfactual reasoning?*

We answer this question with **CIB-MoE** (Counterfactual Inconsistency-Bottleneck Mixture-of-Experts). CIB-MoE formulates multimodal understanding as discrepancy-aware inference: it introduces four **Difference Experts** to characterize semantic alignment, sentiment/prediction disagreement, entity overlap, and image-reliance cues. A **Two-Level Inconsistency-Bottleneck Gate** first detects conflict mode and then sparsely selects experts under an information bottleneck, preventing uninformed feature aggregation. Finally, **Counterfactual Difference Supervision** (neutral vs. random image interventions) calibrates routing behavior, encouraging reliance on task-relevant cross-modal evidence rather than incidental correlations.

Our contributions are:

- **Explicit Difference Modeling:** We propose specialized experts that decompose cross-modal interaction into semantic, predictive, entity, and image-reliance dimensions, improving interpretability.
- **Two-Level IB Gating:** We design a hierarchical, bottlenecked router that detects conflict mode and sparsely selects informative experts, mitigating aspect-agnostic fusion shortcuts.
- **Counterfactual Calibration:** We introduce counterfactual interventions to explicitly supervise routing, suppressing spurious correlations and improving robustness.
- **Strong Empirical Results:** Experiments on MABSA and MMSD benchmarks demonstrate that CIB-MoE achieves superior performance and out-of-distribution generalization.

2 Related Work

2.1 Multimodal Aspect-Based Sentiment Analysis

Surveys consistently identify modality imbalance, weak image–aspect relevance, and spurious correlations as major obstacles for multimodal sentiment analysis and MABSA (Zhao et al., 2024b). Correspondingly, recent MABSA models shift from coarse fusion to structured, fine-grained interaction, including alignment- and multi-view-based designs (e.g., CMFFA, AMIFN) (Xiao et al., 2023; Yang et al., 2024b) and hierarchical structural/semantic alignment (e.g., Atlantis, VLHA) (Xiao et al., 2024; Zou et al., 2025); robustness methods further address noisy images via denoising curricula (e.g., M2DF) (Zhao et al., 2023). Our work departs from implicit attention-based handling by treating cross-modal discrepancies as explicit routing signals.

2.2 Multimodal Sarcasm Detection

MMSD hinges on identifying cross-modal contrast beyond text sentiment alone (Xu et al., 2020), and debiased settings such as MMSD2.0 expose strong reliance on spurious textual cues (Farabi et al., 2024; Qin et al., 2023). Recent approaches leverage VLP models to build visuo-textual representations and quantify incongruity (Qin et al., 2023; Wang et al., 2024), and further decompose inconsistency into factual and affective (sentiment) mismatch (Lu et al., 2024). In parallel, debiasing-oriented methods introduce adaptive routing for sample-specific bias suppression (Wu et al., 2025)

or apply training-free counterfactual debiasing at inference time (Zhu et al., 2024b). Our work complements these lines by explicitly learning discrepancy-aware routing that calibrates modality weighting via realizable interventions.

2.3 MoE and Causal Robustness

Sparse MoE provides conditional computation and has recently been adapted to multimodal and LLM settings (Shazeer et al., 2017; Shen et al., 2023, 2024; Wu et al., 2024; Zhao et al., 2024a; Lo et al., 2025; Fang et al., 2025). Causal and counterfactual approaches debias multimodal affective models via causal graphs or intervention-based objectives (Yang et al., 2024a; Chen et al., 2024; Kim et al., 2024; Patil et al., 2023), but typically require structural assumptions or treat causality as an auxiliary regularizer. CIB-MoE instantiates a "weak-causal" alternative by leveraging realizable neutral/random image replacements, discrepancy-aware experts, and an information-bottlenecked gate to calibrate routing without specifying a full structural causal model.

3 Methodology

We propose **CIB-MoE** (Counterfactual Inconsistency-Bottleneck Mixture-of-Experts), a unified framework for robust multimodal aspect-based sentiment analysis (MABSA) and multimodal sarcasm detection (MMSD). CIB-MoE treats multimodal understanding as discrepancy-aware inference: instead of indiscriminately fusing modalities, it quantifies multiple types of cross-modal differences with lightweight experts, then routes these discrepancy signals through a two-level bottlenecked gate. The routing behavior is calibrated via counterfactual image interventions so that the model relies on task-relevant cross-modal evidence rather than spurious shortcuts. An overview of the CIB-MoE framework and its key components is shown in Figure 2.

3.1 Problem Formulation and Backbone Encoder

Each instance consists of a text sequence $\mathbf{x} = (x_1, \ldots, x_T)$, an image $\mathbf{v}$, and (for MABSA) a queried aspect a. The label y satisfies $y \in \mathcal{Y}_{\text{sent}}$ for MABSA and $y \in \mathcal{Y}_{\text{sar}} = \{\text{non-sar}, \text{sar}\}$ for MMSD. We denote the task posterior by $p_\theta(y \mid \mathbf{x}, \mathbf{v}, a)$ for MABSA and $p_\theta(y \mid \mathbf{x}, \mathbf{v})$ for MMSD.

Backbone representations. We encode text with RoBERTa(Liu et al., 2019) and image with a Vision Transformer (ViT)(Dosovitskiy et al., 2021):

$$\mathbf{H}_t = \text{BERT}(\mathbf{x}) \in \mathbb{R}^{T \times d}, \quad (1)$$

$$\mathbf{H}_v = \text{ViT}(\mathbf{v}) \in \mathbb{R}^{M \times d}, \quad (2)$$

where M is the number of visual patch tokens. We fuse modalities using a multimodal Transformer with cross-attention:

$$\mathbf{H}_m = \text{MMT}(\mathbf{H}_t, \mathbf{H}_v) \in \mathbb{R}^{L \times d}. \quad (3)$$

We take the [CLS] token of $\mathbf{H}_m$ as the pooled multimodal representation $\mathbf{h} \in \mathbb{R}^d$. For MABSA, we produce an aspect-aware pooled representation $\mathbf{h}_a$ via target marking or aspect-conditioned pooling. A task-specific classifier is

$$p_\theta(y \mid \mathbf{x}, \mathbf{v}, a) = \text{softmax}\big(W_y \mathbf{h}_a + \mathbf{b}_y\big), \quad (4)$$

and for MMSD we use the same form with $\mathbf{h}$ in place of $\mathbf{h}_a$. CIB-MoE augments $\mathbf{h}_a$ (or $\mathbf{h}$) with discrepancy-aware expert evidence before prediction.

3.2 Difference Experts

We instantiate $K = 4$ difference experts $\{E_k\}_{k=1}^K$. Each expert first computes a scalar discrepancy statistic and then maps it to an expert feature $\mathbf{e}_k \in \mathbb{R}^{d_e}$ through a lightweight MLP. The resulting expert features are later routed by a sparse gate.

(1) Semantic Inconsistency Expert E_{sem}. We use CLIP to measure global semantic alignment between the text and image. Let $\mathbf{z}_t$ and $\mathbf{z}_v$ denote the CLIP text and image embeddings. We define

$$s_{\text{sem}} = 1 - \cos(\mathbf{z}_t, \mathbf{z}_v), \quad (5)$$

and map it to

$$\mathbf{e}_{\text{sem}} = E_{\text{sem}}(s_{\text{sem}}) \in \mathbb{R}^{d_e}. \quad (6)$$

(2) Unimodal Predictive Discrepancy Expert E_{sent}. We attach a text-only head C_t(RoBERTa) on top of $\mathbf{H}_t$ and an image-only head C_v(ViT) on top of $\mathbf{H}_v$. Both heads predict distributions over the current task label space. Denote their outputs by $\mathbf{p}_t$ and $\mathbf{p}_v$. We use symmetric KL divergence:

$$s_{\text{sent}} = \text{KL}(\mathbf{p}_t \| \mathbf{p}_v) + \text{KL}(\mathbf{p}_v \| \mathbf{p}_t), \quad (7)$$

and obtain

$$\mathbf{e}_{\text{sent}} = E_{\text{sent}}(s_{\text{sent}}) \in \mathbb{R}^{d_e}. \quad (8)$$

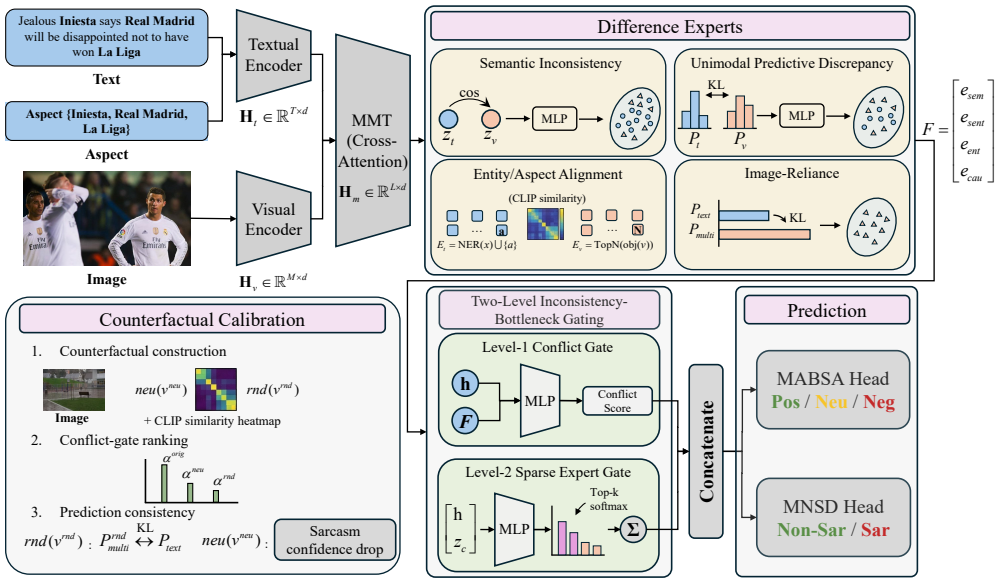

Figure 2: **CIB-MoE framework.** RoBERTa/ViT encode $\mathbf{x}, \mathbf{v}$ and an MMT yields $\mathbf{h}$. CIB-MoE routes four discrepancy cues (CLIP mismatch, unimodal divergence, Top-20 entity overlap, image-induced shift) through a two-level sparse gate, with neutral/random image interventions for calibration.

(3) Entity Alignment Expert E_{ent}. We extract a set of textual entities/aspects $\mathcal{E}_t$ from $\mathbf{x}$ and a set of visual entities $\mathcal{E}_v$ from $\mathbf{v}$. We compute an overlap-based mismatch score:

$$s_{\text{ent}} = 1 - \text{overlap}(\mathcal{E}_t, \mathcal{E}_v), \tag{9}$$

and map it to

$$\mathbf{e}_{\text{ent}} = E_{\text{ent}}(s_{\text{ent}}) \in \mathbb{R}^{d_e}. \tag{10}$$

(4) Image-Reliance (Shortcut) Expert E_{cau}. To quantify reliance on visual evidence, we compare a text-only classifier C_{text} (RoBERTa) with a multimodal classifier C_{multi} based on $\mathbf{h}$ (or $\mathbf{h}_a$ for MABSA). Let their outputs be $\mathbf{p}_{\text{text}}$ and $\mathbf{p}_{\text{multi}}$, respectively. We define

$$s_{\text{cau}} = \text{KL}(\mathbf{p}_{\text{multi}} \| \mathbf{p}_{\text{text}}), \tag{11}$$

and map it to

$$\mathbf{e}_{\text{cau}} = E_{\text{cau}}(s_{\text{cau}}) \in \mathbb{R}^{d_e}. \tag{12}$$

We stack expert features row-wise as

$$\mathbf{F} = [\mathbf{e}_{\text{sem}}; \mathbf{e}_{\text{sent}}; \mathbf{e}_{\text{ent}}; \mathbf{e}_{\text{cau}}] \in \mathbb{R}^{K \times d_e}. \tag{13}$$

3.3 Two-Level Inconsistency-Bottleneck Gating

Not all instances require discrepancy reasoning. We therefore employ a two-level gate: a conflict gate detects whether discrepancy is informative, and an expert gate sparsely selects relevant experts under a bottleneck constraint.

Level-1: Conflict Gate. Let $\text{pool}(\mathbf{F}) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{e}_k$ be mean pooling over experts. We compute a conflict embedding

$$\mathbf{z}_c = \phi_c([\mathbf{h}; \text{pool}(\mathbf{F})]) \in \mathbb{R}^{d_c}, \tag{14}$$

and obtain a scalar conflict score

$$\alpha = \sigma(w_c^\top \mathbf{z}_c + b_c) \in (0, 1). \tag{15}$$

Level-2: Sparse Expert Gate. Conditioned on $(\mathbf{h}, \mathbf{z}_c)$, the expert gate produces routing logits and a sparse mixture:

$$\mathbf{g} = \phi_e([\mathbf{h}; \mathbf{z}_c]) \in \mathbb{R}^K, \tag{16}$$

$$\mathbf{w} = \text{Top-}k\text{-softmax}(\mathbf{g}) \in \Delta^{K-1}, \tag{17}$$

where Top-k-softmax keeps the top-k entries of $\mathbf{g}$, sets the rest to $-\infty$, and applies softmax. The routed expert summary is

$$\mathbf{m} = \sum_{k=1}^{K} w_k \mathbf{e}_k \in \mathbb{R}^{d_e}. \tag{18}$$

Prediction Head. We augment the backbone representation with routed discrepancy evidence:

$$\tilde{\mathbf{h}} = [\mathbf{h}; \mathbf{m}], \tag{19}$$

and predict

$$p_\theta(y \mid \mathbf{x}, \mathbf{v}) = \text{softmax}\big(W_y' \tilde{\mathbf{h}} + \mathbf{b}_y'\big). \tag{20}$$

For MABSA, we replace $\mathbf{h}$ with $\mathbf{h}_a$ in the above equations.

Information Bottleneck Regularization. To discourage dataset-specific routing shortcuts, we regularize the routing distribution:

$$\mathcal{L}_{\text{IB}} = \lambda_{\text{ent}}\,\mathbb{E}[\mathrm{H}(\mathbf{w})] + \lambda_{\text{bal}} \sum_{k=1}^{K} \Big(\frac{1}{B} \sum_{i=1}^{B} w_k^{(i)} - \frac{1}{K} \Big)^2, \tag{21}$$

where minimizing entropy promotes sparse per-instance routing and the balancing term prevents expert collapse.

3.4 Counterfactual Difference Supervision

We calibrate the gate with counterfactual image interventions to distinguish meaningful conflicts from incidental mismatches.

Counterfactual Construction. For each sample $(\mathbf{x}, \mathbf{v})$, we construct a semantically aligned neutral-image counterfactual $(\mathbf{x}, \mathbf{v}^{\text{neu}})$ by retrieval with high CLIP similarity, and a random-image counterfactual $(\mathbf{x}, \mathbf{v}^{\text{rnd}})$ by replacing $\mathbf{v}$ with an unrelated image.

Conflict-Gate Ranking. We impose margin-based ranking constraints:

$$\begin{aligned}\mathcal{L}_{\text{cf\_gate}} = \mathbb{E}\big[\max(0, \alpha^{\text{neu}} - \alpha^{\text{orig}} + \delta)\big] \\ + \mathbb{E}\big[\max(0, \alpha^{\text{rnd}} - \alpha^{\text{orig}} + \delta')\big].\end{aligned} \tag{22}$$

Prediction Consistency. For random-image counterfactuals, the multimodal prediction should revert toward the text-only prediction:

$$\mathcal{L}_{\text{cf\_pred}} = \mathbb{E}\Big[\mathrm{KL}\big(\mathbf{p}^{\text{rnd}}_{\text{multi}} \| \mathbf{p}_{\text{text}}\big)\Big]. \tag{23}$$

For sarcasm detection, removing the conflict should reduce sarcasm confidence:

$$\begin{aligned}\mathcal{L}_{\text{cf\_sar}} = \mathbb{E}\Big[\max\big(0,\, p_\theta(\text{sar} \mid \mathbf{x}, \mathbf{v}^{\text{neu}}) \\ - p_\theta(\text{sar} \mid \mathbf{x}, \mathbf{v}) + \gamma\big)\Big].\end{aligned} \tag{24}$$

3.5 Multi-Task Training Objective

We jointly train on MABSA and MMSD with a shared backbone and a shared CIB-MoE module. Let $\mathcal{L}^{\text{sent}}_{\text{sup}}$ and $\mathcal{L}^{\text{sar}}_{\text{sup}}$ be cross-entropy losses for aspect sentiment and sarcasm prediction. The full objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}^{\text{sent}}_{\text{sup}} + \mathcal{L}^{\text{sar}}_{\text{sup}} + \lambda_{\text{IB}}\mathcal{L}_{\text{IB}} + \lambda_{\text{cf\_gate}}\mathcal{L}_{\text{cf\_gate}} \\ + \lambda_{\text{cf\_pred}}\mathcal{L}_{\text{cf\_pred}} + \lambda_{\text{cf\_sar}}\mathcal{L}_{\text{cf\_sar}}.\end{aligned} \tag{25}$$

Where $\lambda_\cdot$ are hyperparameters. This counterfactually calibrated training encourages sparse, task-relevant routing over discrepancy experts and improves robustness under spurious correlations and modality imbalance.

4 Experiments

In this section, we first describe the tasks and datasets (Sec. 4.1), baselines (Sec. 4.2) and implementation details (Sec. 4.3). We then present our main in-domain results (Sec. 4.4), followed by a series of robustness and generalization analyses (Sec. 4.5), ablation studies (Sec. 4.6), and case studies (Sec. 4.7).

4.1 Tasks and Datasets

We evaluate CIB-MoE on two fine-grained multimodal benchmarks that require modeling cross-modal incongruity but differ in label spaces and dataset biases: MABSA and MMSD.

MABSA. We utilize Twitter-2015/2017, which provide image–tweet pairs with aspect terms and aspect-level sentiment labels. Following standard practice, we split multi-aspect posts into aspect instances and perform three-way classification over $\{\text{POS}, \text{NEU}, \text{NEG}\}$; statistics are in Table 1.

MMSD. We utilize MMSD and MMSD2.0 (Qin et al., 2023) for binary sarcasm detection over $\{\text{SAR}, \text{NON-SAR}\}$ (Table 2). MMSD2.0 reduces lexical shortcuts, and we additionally test cross-dataset shift by training on MMSD and evaluating on MMSD2.0.

Table 1: Statistics of Twitter2015 & Twitter2017 datasets.

	Twitter2015			**Twitter2017**		
Split	**POS**	**NEU**	**NEG**	**POS**	**NEU**	**NEG**
Train	928	1883	368	1508	1638	416
Dev	303	670	149	515	517	144
Test	317	607	113	493	573	168
Total	1548	3160	630	2516	2728	728

Table 2: Data composition of MMSD and MMSD2.0.

	MMSD			MMSD2.0		
	Non-Sar	Sar	Total	Non-Sar	Sar	Total
Train	8,642	11,174	19,816	9,572	10,240	19,816
Validation	959	1,451	2,410	1,042	1,368	2,410
Test	959	1,450	2,409	1,037	1,372	2,409

4.2 Baselines

We benchmark CIB-MoE against baselines that progressively increase modeling capacity from unimodal prediction to structure-aware incongruity modeling and debiasing, covering both MABSA and MMSD.

Unimodal baselines. We include **Text-only (BERT)** and **Image-only (ViT)** as lower bounds to quantify the standalone contribution of each modality. **Traditional multimodal fusion.** We benchmark representative early/late fusion variants and dataset-proposed models such as **HFM** and **D&R Net**, which mainly rely on global cross-modal interaction without explicit discrepancy decomposition. **Task-specific multimodal SOTA.** For **MABSA** (Twitter-2015/2017), we compare against recent strong systems including ITOAOF, AMIFN, AESAL, DEQA, DPCI, and CORSA (Table 3). For **MMSD** (MMSD/MMSD2.0), we include graph-based and incongruity-aware models (e.g., CMGCN, InCrossMGs), CLIP- and routing-based methods, and retrieval/MLLM-style classifiers (e.g., Multi-view CLIP, MILNet, FSICN, VIDR-MLLM, MICL, SCI-GDFN; Table 4). **Debiasing/causal baselines.** We further consider debiasing methods that target spurious correlations via counterfactual augmentation or inference-time bias subtraction (e.g., DMSD, TFCD; Table 4).

4.3 Implementation Details

We adopt RoBERTa-base and ViT-B/32 as backbone encoders. Each discrepancy expert is a two-layer MLP (d_e=128), producing $\mathbf{E} \in \mathbb{R}^{4\times d_e}$. Routing is performed by a two-level gate with a conflict embedding (d_c=128) and Top-2 sparse selection, regularized by an information bottleneck (λ_{ent}=0.01, λ_{bal}=0.1, λ_{IB}=0.5); counterfactual losses share a fixed weight of 0.1. Semantic discrepancy uses CLIP ViT-B/32 embeddings, and entity alignment is computed over the Top-20 predicted object labels. Training uses AdamW (lr 2×10^{-5}, batch size 16 and training for 50 epochs, 10% warmup) with interleaved MABSA/MMSD mini-batches; neutral and random-image interventions are constructed on-the-fly via CLIP retrieval and uniform sampling. We report Acc/Macro-F1 for MABSA and Acc/P/R/F1 for MMSD/MMSD2.0, averaged over five seeds, with paired t-tests against the strongest baseline when applicable (p<0.05). Experiments are conducted on a PyTorch framework using a single RTX A6000 GPU with 48GB of memory.

Table 3: Experimental results on Twitter-2015 and Twitter-2017 for MABSA.

Method	Twitter2015		Twitter2017	
	Acc.	Mac-F1	Acc.	Mac-F1
BERT(Yu and Jiang, 2019)	74.15	68.86	68.15	65.23
ViLBERT (Yu et al., 2022)	73.76	69.85	67.42	64.87
TomBERT (Yu and Jiang, 2019)	77.15	71.75	70.34	68.03
ESAFN (Yu et al., 2019)	73.38	67.37	67.83	64.22
EF-CapTrBERT-DE (Khan and Fu, 2021)	77.92	73.9	72.3	70.2
FITE-DE-Large (Yang et al., 2022)	78.76	74.79	73.87	73.03
ITOAOF (Wang et al., 2023)	79.45	75.11	74.47	73.05
AMIFN (Yang et al., 2024b)	78.69	75.50	72.29	70.21
AESAL (Zhu et al., 2024a)	80.1	75.2	78.8	75.9
DEQA (Han et al., 2025)	82.1	77.6	75.8	75.1
DPCI (Liu et al., 2025a)	80.42	76.39	75.20	74.73
CORSA (Liu et al., 2025b)	81.1	77.7	76.6	74.5
CIB-MoE	**82.34**	**78.29**	**78.96**	**76.05**

4.4 Main Results

Table 3 reports MABSA performance on Twitter-2015/2017, and Table 4 summarizes results on MMSD and MMSD2.0. Overall, CIB-MoE achieves consistent improvements over strong task-specific baselines, suggesting that explicitly modeling and calibrating cross-modal discrepancies is beneficial across both sentiment and sarcasm settings.

Results on MABSA. Table 3 shows that CIB-MoE establishes new best results on both Twitter benchmarks. On Twitter-2015, it achieves **82.34** Acc / **78.29** Macro-F1, exceeding CORSA by **+1.24** Acc and **+0.59** Macro-F1. On Twitter-2017, it reaches **78.96** Acc / **76.05** Macro-F1, improving over CORSA by **+2.36** Acc and **+1.55** Macro-F1. The larger gains on Twitter-2017 are consistent with the setting where images are more weakly related and discrepancy-aware routing is more critical.

Results on MMSD and MMSD2.0. As reported in Table 4, CIB-MoE achieves **94.14** F1 on MMSD, outperforming strong recent baselines such as MICL (+**3.81** F1) and SCI-GDFN (+**0.34** F1). On the debiased MMSD2.0 benchmark, it attains **89.68** F1 / **91.33** Acc, surpassing robustness-oriented methods including TFCD and VIDR-MLLM. Overall, the consistent improvements across in-domain and debiased settings indicate that calibrated discrepancy modeling benefits both aspect-level sentiment prediction and sarcasm detection.

4.5 Robustness and Generalization Analysis

We assess whether CIB-MoE generalizes beyond benchmark-specific shortcuts using cross-dataset

Table 4: Performance Comparison on MMSD and MMSD2.0

Method	MMSD				MMSD2.0			
	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Acc (%)	Pre (%)	Rec (%)	F1 (%)
Unimodal Baselines								
Text-Only (BERT)	83.85	81.24	79.15	80.22	74.78	75.12	71.58	73.34
Image-Only (ViT)	67.83	65.92	61.12	63.43	72.02	71.48	68.05	69.72
Traditional Multimodal								
HFM (Cai et al., 2019)	83.44	76.57	84.15	80.18	70.57	64.84	69.05	66.88
D&R Net (Xu et al., 2020)	84.02	77.97	83.42	80.60	-	-	-	-
Graph-Based								
CMGCN (Liang et al., 2022)	87.55	83.63	84.69	84.16	79.83	78.45	75.42	76.90
InCrossMGs (Liang et al., 2021)	86.10	81.38	84.36	82.84	-	-	-	-
Recent SOTA								
Multi-view CLIP (Qin et al., 2023)	88.33	82.66	88.65	85.55	85.64	80.33	88.24	84.10
HKE (Liu et al., 2022)	87.36	82.97	86.48	84.89	-	-	-	-
MILNet (Qiao et al., 2023)	89.50	88.24	85.99	87.11	-	-	-	-
FSICN (Lu et al., 2024)	90.55	90.12	89.32	89.72	-	-	-	-
DIP (Wen et al., 2023)	89.59	87.76	86.58	87.17	-	-	-	-
DMSD (Jia et al., 2024)	88.95	84.89	87.90	86.37	-	-	-	-
G^2SAM (Wei et al., 2024)	90.48	87.95	89.02	88.48	-	-	-	-
VIDR-MLLM (Tang et al., 2024)	89.97	89.26	89.58	89.42	86.43	87.00	86.30	86.34
TFCD (Zhu et al., 2024b)	89.57	84.83	89.43	88.13	86.54	82.46	87.95	84.31
MICL (Guo et al., 2025)	92.08	90.05	90.61	90.33	-	-	-	-
AMSK (Dong et al., 2026)	89.84	87.51	87.60	87.56	86.39	79.74	**89.37**	84.28
SCI-GDFN (Xi et al., 2025)	94.06	93.50	94.17	93.80	-	-	-	-
CIB-MoE	**94.82**	**93.98**	**94.31**	**94.14**	**91.33**	**90.65**	88.74	**89.68**

Table 5: Cross-dataset generalization performance.

Method (MMSD / MABSA)	MMSD → MMSD2.0 (ΔF1)	Tw15 → Tw17 (ΔF1)	Tw17 → Tw15 (ΔF1)
Text-only (BERT)	73.34 (−6.88)	66.45 (−2.41)	68.90 (+3.67)
HFM / TomBERT	66.88 (−13.30)	64.12 (−7.63)	67.55 (−4.20)
Multi-view CLIP / AESAL	84.10 (−1.45)	71.80 (−3.40)	74.20 (−1.00)
TFCD / CORSA	84.31 (−3.82)	73.15 (−4.55)	76.85 (+0.25)
CIB-MoE (ours)	**89.68 (−4.46)**	**74.92 (−3.37)**	**79.15 (+3.10)**

Table 6: Counterfactual robustness analysis on MMSD2.0.

Method	Random-image Consistency ↑	Neutral-image Monotonicity ↑
Text-only (BERT)	1.00	0.00
Multi-view CLIP	0.76	0.22
VIDR-MLLM	0.82	0.31
TFCD	0.88	0.45
SCI-GDFN	0.85	0.38
CIB-MoE (ours)	**0.96**	**0.64**

transfer (Table 5) and counterfactual robustness tests (Table 6).

Cross-dataset generalization. As reported in Table 5, CIB-MoE yields strong target-domain performance under distribution shift. On MMSD→MMSD2.0, it reaches 89.68 F1 with a 4.46-point drop, outperforming unimodal BERT (73.34, −6.88) and the fusion baseline HFM (66.88, −13.30). While Multi-view CLIP exhibits a smaller drop (84.10, −1.45), its MMSD2.0 F1 remains notably lower than CIB-MoE; TFCD also lags on MMSD2.0 (84.31 vs. 89.68). CIB-MoE further shows stable transfer on Twitter-2015/2017 (Tw15→Tw17: 74.92, −3.37; Tw17→Tw15: 79.15, +3.10).

Counterfactual robustness. Table 6 confirms that CIB-MoE is both noise-invariant and conflict-sensitive: it achieves 0.96 random-image consistency and 0.64 neutral-image monotonicity, the best among the compared methods. This behavior is consistent with counterfactually calibrated routing that suppresses non-informative visual cues while reacting to conflict removal.

4.6 Ablation Studies

Table 7 shows that **difference experts** are essential: removing them causes the largest degradation (Tw17: −2.20; MMSD2.0: −4.26). Counterfactual training is the primary driver of calibrated routing (w/o CF losses: 74.95/87.25), while the conflict gate and IB regularizer provide smaller but consistent gains. Single-expert settings underperform the full model; the entity expert is most effective on Tw17 (75.15), whereas sentiment and causal experts are strongest on MMSD2.0 (88.10/88.50).

We observe three consistent trends (Table 7): (i) removing all difference experts brings performance close to or slightly above the best generic MoE baseline, confirming that the gains of CIB-MoE come primarily from structured difference modeling; (ii) the conflict gate and IB regularization both contribute positively, with the largest degradation observed when all counterfactual losses are removed, indicating that intervention-based supervision is crucial for robust gating; (iii) among single-

Image			
Text	RT @ SportsCenter : BREAKING : One . More . Win . **[Kentucky Derby]**$_{\text{Neu}}$ winner **[American Pharoah]**$_{\text{Pos}}$ crosses finish line first in **[Preakness Stakes]**$_{\text{Neu}}$.	Found this board game at the **[Chelsea Flea Market]**$_{\text{Neu}}$. Oh how the times have changed for **[Bruce]**$_{\text{Pos}}$	lucky me ! two city cat journeys in one day
ITOAOF\DIP	**Kentucky Derby : Positive (×)** **American Pharoah : Positive (√)** **Preakness Stakes : Neutral (√)**	**Chelsea Flea Market : Positive (×)** **Bruce : Positive (√)**	**Non-Sarcasm**
CORSA\MICL	**Kentucky Derby : Neutral (√)** **American Pharoah : Positive (√)** **Preakness Stakes : Positive (×)**	**Chelsea Flea Market : Neutral (√)** **Bruce : Neutral (×)**	**Non-Sarcasm**
CIB-MoE	**Kentucky Derby : Neutral (√)** **American Pharoah : Positive (√)** **Preakness Stakes : Neutral (√)**	**Chelsea Flea Market : Neutral (√)** **Bruce : Positive (√)**	**Sarcasm**

Figure 3: A comparison of our model with baseline models on three samples from the MMSD dataset.

Table 7: Ablation study on Twitter-2017 and MMSD2.0. We report Macro-F1 for Twitter-2017 and F1 for MMSD2.0. The full model yields the best performance, validating the synergy of all components.

Variant	Tw17 Mac-F1	MMSD2.0 F1
Full CIB-MoE	**76.05**	**89.68**
w/o Diff-Experts	73.85	85.42
w/o Conflict Gate	75.42	88.95
w/o IB	75.78	89.15
w/o CF losses	74.95	87.25
Single Expert Analysis		
Only E_{sem} (Semantic)	74.60	87.51
Only E_{sent} (Sentiment)	74.45	88.16
Only E_{ent} (Entity)	75.15	86.88
Only E_{cau} (Causal)	74.88	88.57

expert variants, the causal and sentiment inconsistency experts are most impactful on MMSD2.0, while entity alignment plays a more important role for aspect-level sentiment in Twitter-2015/2017, aligning with our intuitive understanding of the tasks.

4.7 Case Studies

Figure 3 summarizes three instances where discrepancy-aware routing yields clearer aspect- and intent-level decisions.

MABSA: mitigating sentiment leakage. In the horse-racing example, ITOAOF/DIP predicts POS for Kentucky Derby and CORSA predicts POS for Preakness Stakes, although both are NEU. CIB-MoE correctly classifies all three entities, indicating reduced sentiment spillover from the winner-related cue to co-mentioned event mentions.

MABSA: target-specific grounding under weak visual relevance. In the second post, ITOAOF/DIP assigns POS to the neutral location (Chelsea Flea Market) and CORSA misses the POS label for Bruce. CIB-MoE separates the neutral context entity from the opinion target, supporting aspect-conditioned prediction under weak image–text coupling.

MMSD: beyond literal polarity. For the sarcasm case ("lucky me! two city cat journeys in one day"), baselines output NON-SAR while CIB-MoE predicts SAR, suggesting that discrepancy-aware routing can attenuate locally positive wording when it conflicts with post-level context.

5 Conclusion

We propose **CIB-MoE** for robust MABSA and MMSD, casting multimodal prediction as discrepancy-aware routing with typed experts and a two-level information-bottlenecked gate. We further calibrate routing using neutral/random image interventions as counterfactual supervision. Experiments on Twitter-2015/2017 and MMSD/MMSD2.0 show consistent gains in accuracy and robustness, and case studies confirm reduced sentiment leakage across co-mentioned entities and improved handling of subtle sarcasm beyond literal polarity.

Limitations

CIB-MoE relies on several practical components whose quality bounds overall performance. First, the entity-alignment signal depends on the accuracy and calibration of the image-side label predictions (Top-N object labels) and text-side entity/aspect extraction; errors in either stage may weaken or misdirect routing. Second, the neutral-image intervention uses CLIP-based retrieval, which may return images that are only superficially aligned with the text, introducing imperfect counterfactuals and limiting the strength of supervision. Third, our discrepancy experts are intentionally lightweight and capture a predefined set of difference types; although this improves interpretability and stability, it may miss rarer forms of incongruity (e.g., fine-grained pragmatics or culture-specific sarcasm) that require richer world knowledge. Finally, while the two-level gate adds modest overhead, the additional experts and counterfactual sampling increase training-time computation, and the current implementation evaluates robustness on established Twitter-style benchmarks; broader validation on other domains (e-commerce reviews, multilingual settings, or video-based sarcasm) remains for future work.

References

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics (ACL 2019)*, pages 2506–2515.

Fuhai Chen, Pengpeng Huang, Xuri Ge, Jie Huang, and Zishuo Bao. 2024. Multimodal sentiment analysis based on causal reasoning. *arXiv preprint arXiv:2412.07292*.

Jing Dong, Yu Sui, Qiang Zhang, Hui Fang, Gerald Schaefer, Rui Liu, Pengfei Yi, and Xiaoyong Fang. 2026. An adaptive multimodal semantic knowledge enhanced framework for sarcasm detection. *Expert Systems with Applications (ESWA)*, 298:129773.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR 2021)*.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. 2025. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14314–14324.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, Jeju, Korea. IJCAI.

Xin Guan, Jiuxin Cao, Hui Zhang, Biwei Cao, and Bo Liu. 2025. Mian: Multi-head incongruity aware attention network with transfer learning for sarcasm detection. *Expert Systems with Applications (ESWA)*, 263:125702.

Diandian Guo, Cong Cao, Fangfang Yuan, Yanbing Liu, Guangjie Zeng, Xiaoyan Yu, Hao Peng, and S. Yu Philip. 2025. Multi-view incongruity learning for multimodal sarcasm detection. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 1754–1766.

Zhixin Han, Mengting Hu, Yinhao Bai, Xunzhi Wang, and Bitong Luo. 2025. Deqa: Descriptions enhanced question-answering framework for multimodal aspect-based sentiment analysis. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2025)*, volume 39, pages 23987–23995.

Mengzhao Jia, Tao Wang, Liqiang Zhang, and Xinyu Chen. 2024. Debiasing multimodal sarcasm detection with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2024)*, volume 38, pages 18354–18362.

Zaid Khan and Yun Fu. 2021. Exploiting BERT for multimodal target sentiment classification through input space translation. In *Proceedings of the ACM International Conference on Multimedia (ACM MM2021)*, pages 3034–3042.

Juyeon Kim, Juyoung Hong, and Yukyung Choi. 2024. Causal inference for modality debiasing in multimodal emotion recognition. *Applied Sciences*, 14(23):11397.

Kuntao Li, Yifan Chen, Qiaofeng Wu, Weixing Mai, Fenghuan Li, and Yun Xue. 2025. Ambiguity-aware multi-level incongruity fusion network for multimodal sarcasm detection. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 380–391.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM 2021)*, pages 4707–4715.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Volume 1: Long Papers*, pages 1767–1777.

Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 4995–5006.

Rui Liu, Jiahao Cao, Jiaqian Ren, Xu Bai, and Yanan Cao. 2025a. Dual-path counterfactual integration for multimodal aspect-based sentiment classification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, pages 22759–22769.

Xinjing Liu, Ruifan Li, Shuqin Ye, Guangwei Zhang, and Xiaojie Wang. 2025b. Multimodal aspect-based sentiment analysis under conditional relation. In *Proceedings of the International Conference on Computational Linguistics (COLING 2025)*, pages 313–323.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. 2025. A closer look into mixture-of-experts in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4427–4447.

Qiang Lu, Yunfei Long, Xia Sun, Jun Feng, and Hao Zhang. 2024. Fact-sentiment incongruity combination network for multimodal sarcasm detection. *Information Fusion*, 104:102203.

Vaidehi Patil, Adyasha Maharana, and Mohit Bansal. 2023. Debiasing multimodal models via causal information minimization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4108–4123.

Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2023)*, volume 37, pages 9507–9515.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023 (Findings of ACL 2023)*, pages 10834–10845.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. 2024. Mome: Mixture of multimodal experts for generalist multimodal large language models. *Advances in neural information processing systems*, 37:42048–42070.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. 2023. Scaling vision-language models with sparse mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11329–11344.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024) (Volume 1: Long Papers)*, pages 1732–1742.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Volume 1: Long Papers*, pages 2468–2480.

Jie Wang, Yan Yang, Yongquan Jiang, Minbo Ma, Zhuyang Xie, and Tianrui Li. 2024. Cross-modal incongruity aligning and collaborating for multi-modal sarcasm detection. *Information Fusion*, 103:102132.

Qianlong Wang, Hongling Xu, Zhiyuan Wen, Bin Liang, Min Yang, Bing Qin, and Ruifeng Xu. 2023. Image-to-text conversion and aspect-oriented filtration for multimodal aspect-based sentiment analysis. *IEEE Transactions on Affective Computing (TAFFC)*, 15(3):1264–1278.

Yiwei Wei, Shaozu Yuan, Hengyang Zhou, Longbiao Wang, Zhiling Yan, Ruosong Yang, and Meng Chen. 2024. G2SAM: Graph-based global semantic awareness method for multimodal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2024)*, volume 38, pages 9151–9159.

Changsong Wen, Guoli Jia, and Jufeng Yang. 2023. Dip: Dual incongruity perceiving network for sarcasm detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)*, pages 2540–2550.

Xun Wu, Shaohan Huang, Wenhui Wang, Shuming Ma, Li Dong, and Furu Wei. 2024. Multi-head mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:94073–94096.

Zichen Wu, Hsiu-Yuan Huang, and Yunfang Wu. 2025. Beyond spurious signals: Debiasing multimodal large language models via counterfactual inference and adaptive expert routing. In *Findings of the Association for Computational Linguistics: EMNLP 2025 (Findings of EMNLP 2025)*, pages 3805–3825.

Zhonghao Xi, Bengong Yu, and Haoyu Wang. 2025. Multimodal sarcasm detection based on sentiment-clue inconsistency global detection fusion network. *Expert Systems with Applications (ESWA)*, 275:127020.

Luwei Xiao, Xingjiao Wu, Junjie Xu, Weijie Li, Cheng Jin, and Liang He. 2024. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Information Fusion*, page 102304.

Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Jie Zhou, and Liang He. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management (IP&M)*, 60(6):103508.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 3777–3786.

Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. 2024a. Towards multimodal sentiment analysis debiasing via bias purification. In *European Conference on Computer Vision*, pages 464–481.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 3324–3335.

Juan Yang, Mengya Xu, Yali Xiao, and Xu Du. 2024b. AMIFN: Aspect-guided multi-view interactions and fusion network for multimodal aspect-based sentiment analysis. *Neurocomputing*, 573:127222.

Jianfei Yu, Kai Chen, and Rui Xia. 2022. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing (TAFFC)*, 14(3):1966–1978.

Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2019)*.

Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Fei Zhao, Chunhui Li, Zhen Wu, Yawen Ouyang, Jianbing Zhang, and Xinyu Dai. 2023. M2df: Multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis. In *EMNLP 2023*, pages 9057–9070.

Guilong Zhao, Yixia Zhao, Xiangrong Yin, Lei Lin, and Jizhao Zhu. 2025. Beyond spurious cues: Adaptive multi-modal fusion via mixture-of-experts for robust sarcasm detection. *Mathematics*, 13(20):3250.

Hao Zhao, Zihan Qiu, Huijia Wu, Zili Wang, Zhaofeng He, and Jie Fu. 2024a. Hypermoe: Towards better mixture of experts via transferring among experts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10605–10618.

Tianyu Zhao, Ling-ang Meng, and Dawei Song. 2024b. Multimodal aspect-based sentiment analysis: A survey of tasks, methods, challenges and future directions. *Information Fusion*, 112:102552.

Linlin Zhu, Heli Sun, Qunshu Gao, Tingzhou Yi, and Liang He. 2024a. Joint multimodal aspect sentiment analysis with aspect enhancement and syntactic adaptive learning. In *Proceedings of the thirty-third international joint conference on artificial intelligence (IJCAI 2024)*, pages 6678–6686.

Zhihong Zhu, Xianwei Zhuang, Yunyan Zhang, Derong Xu, Guimin Hu, Xian Wu, and Yefeng Zheng. 2024b. TFCD: Towards multi-modal sarcasm detection via training-free counterfactual debiasing. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2024)*, pages 6687–6695.

Wang Zou, Xia Sun, Qiang Lu, Xuxin Wang, and Jun Feng. 2025. A vision and language hierarchical alignment for multimodal aspect-based sentiment analysis. *Pattern Recognition*, 162:111369.