# Beyond Spurious Signals: Debiasing Multimodal Large Language Models via Counterfactual Inference and Adaptive Expert Routing

**Zichen Wu, Hsiu-Yuan Huang, Yunfang Wu***
School of Computer Science, Peking University
MOE Key Laboratory of Computational Linguistics, Peking University
National Key Laboratory for Multimedia Information Processing, Peking University
wuzichen@pku.edu.cn, wuyf@pku.edu.cn

## Abstract

Multimodal Large Language Models (MLLMs) have shown substantial capabilities in integrating visual and textual information, yet frequently rely on spurious correlations, undermining their robustness and generalization in complex multimodal reasoning tasks. This paper addresses the critical challenge of superficial correlation bias in MLLMs through a novel causal mediation-based debiasing framework. Specially, we distinguishing core semantics from spurious textual and visual contexts via counterfactual examples to activate training-stage debiasing and employ a Mixture-of-Experts (MoE) architecture with dynamic routing to selectively engages modality-specific debiasing experts. Empirical evaluation on multimodal sarcasm detection and sentiment analysis tasks demonstrates that our framework significantly surpasses unimodal debiasing strategies and existing state-of-the-art models. For further research, we release the training/evaluation pipelines at Github[1].

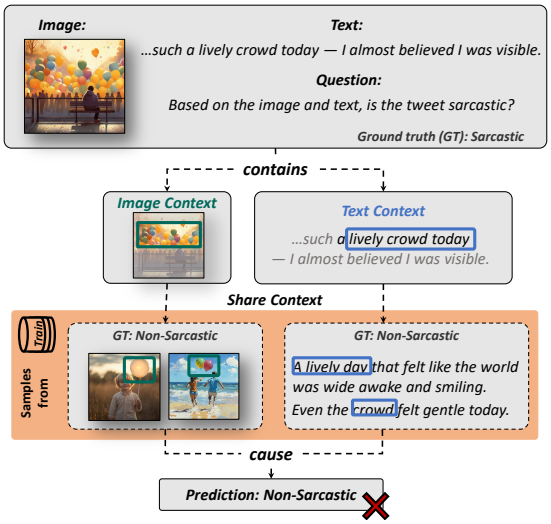


Figure 1: A diagram shows how a large model is negatively impacted by learning spurious correlations during training. In the depicted example, a non-sarcastic test sample (GT) is misclassified as sarcastic because its features frequently co-occurred with sarcastic labels in the training data.

## 1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated significant capabilities in integrating information from various modalities, such as vision and language, achieving notable success in multimodal tasks (OpenAI et al., 2024; Liu et al., 2023; Chen et al., 2024d; Wang et al., 2024b). By unifying modalities, MLLMs can capture richer semantic information than unimodal approaches, establishing them as a promising direction for complex, real-world applications (Wang et al., 2024a).

Despite this progress, current MLLMs remain unreliable on tasks requiring nuanced semantic understanding and reasoning. In practice, they often over-rely on spurious correlations in one modality instead of truly fusing information, which leads to hallucinations and poor generalization (Hosseini et al., 2025; Zhang et al., 2024b). Prior studies have found that these models may latch onto superficial cues present in the training data, leading to biases that are then amplified by large-scale pretraining (Zhao et al., 2024). For example, in sarcasm detection dataset MMSD2.0 (Qin et al., 2023), certain words (e.g., "weather") or objects in an image (e.g., "ceramic mug") appeared frequently with the sarcastic label (Fig. 2), inadvertently cueing the model to predict "sarcasm" whenever it encounters those features. After supervised training, the model learns such spurious associations instead of genuine cross-modal reasoning, resulting in brittle performance on test data where the same shortcuts do not hold. These observations highlight the need for methods to make MLLMs focus on core semantics rather than incidental correlations.

To address this "superficial correlation" bias,

* Corresponding author.
[1] https://github.com/Zichen-Wu/Multimodal-Mixture-of-Expert-Debiasing.



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3805–3825
November 4-9, 2025 

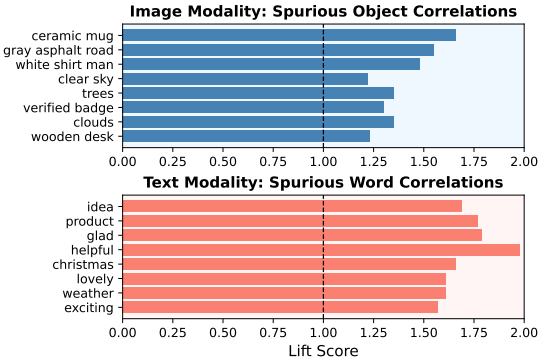


Figure 2: Lift scores for spurious correlations in image (objects) and text (words) modalities. Scores above 1.0 indicate positive spurious correlations.

causal mediation analysis (Pearl, 2022) offers a powerful framework for enhancing the reliability of multimodal semantic understanding. Its core principle involves modeling the multimodal task as a causal graphical model. By generating counterfactual inputs (isolating suspected biased elements) and comparing the model's predictions on these versus original inputs, it becomes possible to quantify and thereby mitigate spurious correlations between input features and the model's output.

However, existing approaches applying causality to MLLMs face limitations. Some methods perform debiasing operations on only a single modality (Niu et al., 2021; Agarwal et al., 2020; Patil et al., 2023) or apply corrections solely during the inference stage (Zhu et al., 2024b; Yang et al., 2024a), neglecting the significant impact of multi-modality and debiasing the model during the training. Other works focus on debiasing learned representations in a more general sense and have not been effectively adapted to modern MLLMs (Zhang et al., 2024a; Chen et al., 2024a). These shortcomings often fail to comprehensively address biases arising from intricate multimodal interactions, leaving the model's internal representations suboptimal. Based on these observations, we pose the following research question: **How can causal mediation analysis be employed to jointly debias the superficial correlations within both textual and visual pathways of an MLLM, at both the training and inference stages?**

To address this issue, we first propose a multimodal causal analysis framework that explicitly distinguishes core semantic information from spurious contextual cues within each modality. Leveraging large pretrained models, we automatically extract superficial context from both textual and visual modalities to construct counterfactual samples. By penalizing model reliance on these counterfactual samples during training, we effectively integrate debiasing into the learning stage, thereby enhancing the robustness and representational power of the main multimodal model.

Recognizing that the automatically extracted superficial contexts may contain inaccuracies and that not all samples require uniform debiasing treatment, we further introduce a routing mechanism combined with a Mixture-of-Experts (MoE) architecture. Under this framework, dedicated expert models specifically handle the debiasing tasks for textual and visual modalities independently. The router dynamically learns and determines which debiasing expert(s) should be activated for the given sample, ensuring tailored and efficient debiasing. During training, these experts are exposed to modality-specific counterfactual data, compelling them to specialize in identifying and mitigating spurious influences. At inference time, the integrated system combines predictions from the main model and the appropriate bias experts, as guided by the router, to yield robust and accurate final outcomes.

We validate the proposed approach on two challenging multimodal semantic understanding tasks: sarcasm recognition (MSD) and sentiment analysis (MSA). Our experiments demonstrate that the causal debiasing framework substantially improves reliability and accuracy compared to both unimodal debiasing baselines and state-of-the-art task-specific models. Notably, the model achieves better generalization to different debiasing categories where naive finetuning falters, confirming that it learned to discount superficial correlations and focus on true multimodal semantics. These results underscore the benefits of a principled, joint debiasing strategy for MLLMs. In summary, our contributions include:

- We formulate a multimodal debiasing approach grounded in causal mediation analysis, which simultaneously addresses both text and image biases.
- We propose a novel expert-based architecture with a gating mechanism to isolate and remove spurious influences during both training and inference.
- Our approach achieves state-of-the-art results on sarcasm and sentiment tasks, enhancing robustness and interpretability.

## 2 Related Work

### 2.1 Bias in MLLMs

Multimodal vision-language models often learn unintended spurious correlations between modality-specific cues and target outputs, causing biased predictions and poor generalization. For instance, Visual QA models frequently exploit dataset biases, such as responding *yes* to questions starting with *Do you see* due to learned shortcuts (Niu et al., 2021; Kim et al., 2023). These biases may arise from coincidental visual or textual patterns. MLLMs trained on web-scale data can further inherit and amplify such modality-specific biases, which drives addressing these biases crucial for robust multimodal understanding (Ye et al., 2024; Zhang et al., 2024b).

### 2.2 Causal Mediation Analysis and Interventions

Structural Causal Models (SCMs) offer a principled approach to explicitly model and mitigate spurious correlations by using interventions and counterfactual reasoning (Pearl, 2022), a direction that has attracted increasing attention (Shekhar et al., 2017; Thrush et al., 2022; Le et al., 2023; Zheng et al., 2024a; Leng et al., 2024; Zhu et al., 2024a).

Recent studies apply causal inference frameworks in multimodal tasks to quantify and remove modality-induced biases (Palit et al., 2023; Golovanevsky et al., 2025; Yu et al., 2024b). For instance, Zhu et al. (2024b) used counterfactual text generation to address textual biases in sarcasm detection, and Yang et al. (2024a) adopted a similar approach for sentiment analysis. Chen et al. (2024a) extended these strategies by explicitly modeling biases in both image and text modalities, but did not investigate the bias originating from spurious parts in a fine-grained manner. In VQA tasks, causal reasoning methods include the Chain-of-Thought (CoT)-based CAVE module by Chen et al. (2024b), and Liu et al. (2024)'s entropy-based counterfactual debiasing strategy for video-grounded QA. Additionally, Yang et al. (2024b) applied causal inference in emotion recognition by decomposing context into relevant and irrelevant cues, eliminating distracting signals during inference. Unlike previous methods, our proposed approach explicitly constructs fine-grained counterfactual samples, integrating causal debiasing directly into both training and inference to comprehensively address multimodal biases.

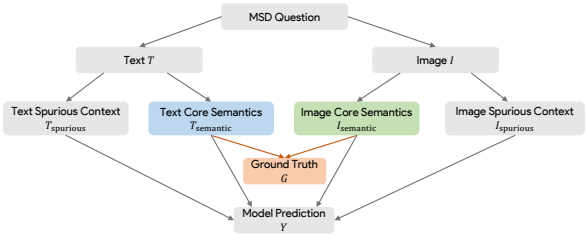


Figure 3: Causal Graph for Multimodal Sarcasm Detection. The non-grey regions indicate the ideal causal mechanisms. Unbiased prediction results could be achieved by separating image and text inputs into semantic and spurious components and performing controlled interventions.

## 3 Preliminaries

### 3.1 Multimodal Causal Mediation Framework

MLLMs often suffer biases from spurious textual and visual contexts, degrading prediction accuracy. We propose a multimodal causal mediation framework to explicitly model causal relationships among multimodal inputs, bias-mediating variables, and outputs.

Taking sarcasm detection as an example, inputs include text ($T$) and image ($I$), ideally generating semantic features ($T_{semantic}$, $I_{semantic}$). However, irrelevant contextual features ($T_{spurious}$, $I_{spurious}$) may bias outcomes. We illustrate these effects via a multimodal causal graph (Fig. 3), distinguishing:
**Unbiased path:** $T \rightarrow T_{semantic} \rightarrow Y$ and $I \rightarrow I_{semantic} \rightarrow Y$, capturing desired accurate semantic understanding.
**Biased path:** $T \rightarrow T_{spurious} \rightarrow Y$ and $I \rightarrow I_{spurious} \rightarrow Y$, capturing biases introduced by spurious information.

We capture the Natural Direct Effect (NDE) of inputs on the prediction outcome as the unbiased results. In causal mediation theory, it refers to isolating the direct influence of the relevant semantic information ($T_{semantic}, I_{semantic}$), while holding mediating biases constant at baseline levels. Based on the multimodal causal graph, we quantify these effects through three scenarios:

• Textual Counterfactual Scenario: Isolate $T_{spurious}$ and mask visual input:

$$Y_t = Y(T_{spurious}, \phi). \quad (1)$$

• Visual Counterfactual Scenario: Isolate $I_{spurious}$ and mask textual input:

$$Y_i = Y(\phi, I_{spurious}). \quad (2)$$

• Original Scenario: Use the full original input:

$$Y_0 = Y(T, I). \tag{3}$$

We then extract the unbiased output by computing the difference between the original and counterfactual predictions:

$$Y_{\text{unbiased}} = \text{DIFF}(Y_0, Y_t, Y_i), \tag{4}$$

This provides a principled estimate of prediction free from modality-specific spurious bias.

### 3.2 Counterfactual Content Construction

Building upon our multimodal causal mediation framework (Sec. 3.1), the ability to isolate specific causal factors necessitates the construction of targeted counterfactual contents. Unlike traditional methods relying on heuristic rules or dataset annotations, we automate the generation of fine-grained multimodal counterfactual inputs ($T_{spurious}, I_{spurious}$) using MLLMs. For textual inputs, this involves identifying and masking core semantic segments. For visual inputs, we utilize attention mechanisms to pinpoint and modify salient regions. Due to space constraints, a detailed methodological description is provided in App. B. These meticulously constructed counterfactuals are instrumental for the causal debiasing techniques presented in following sections.

## 4 Multimodal Debiasing Methods

In this section, we propose methods to mitigate spurious multimodal biases, targeting both inference and training stages:

1. **Multimodal Inference Debiasing (MID):** An inference-stage, plug-and-play method for external bias correction without altering model parameters.

2. **Multimodal Training Debiasing:** To embed robust representations, we integrate debiasing directly into model training. This involves foundational principles of counterfactual-aware training. which uses counterfactual samples to reduce spurious correlations. Building upon these, we introduce our primary training-stage method, **Multimodal Mixture-of-Experts Joint Debiasing (MME-JD)**. This advanced MoEs approach uses a learned router to adaptively dispatch samples to specialized expert branches, enabling fine-grained debiasing. These methods are detailed in the subsequent sections.

### 4.1 MID: Multimodal Inference Debiasing

The simplest debiasing approach applies only at inference, without any additional training procedure. Leveraging the proposed multimodal causal mediation anlysis framework, we could easily adapt it to MLLMs by using counterfactual samples in an inference-time plug-and-play manner.

Given original multimodal inputs (image $I$, text $T$), we first derive the original prediction probabilities $p_0$. We then generate counterfactual samples icluding text-only context ($\hat{T}$) and image-only context ($\hat{I}$), to isolate modality-specific biases, yielding probability distributions $p_t$ and $p_i$, respectively. Each $\{p_0, p_t, p_i\}$ is a vector of length $K$ (for a $K$-class prediction task).

To suppress the bias introduced by spurious text and/or image context, we perform a linear correction on $p_0$:

$$\tilde{p} = p_0 - \alpha p_i - \beta p_t, \tag{5}$$

where $\alpha$ and $\beta$ are hyperparameters controlling how aggressively we subtract the counterfactual probabilities from the original prediction.

Choosing $\alpha, \beta$ is non-trivial: different datasets and tasks may require stronger or weaker correction. We follow a validation-set-based approach, searching over $\alpha, \beta \in [0, 1]$ to maximize a performance metric Eval (We use F1 in experiments):

$$\hat{\alpha}, \hat{\beta} = \arg \max_{\alpha, \beta \in [A, B]} Eval(D|f, \alpha, \beta). \tag{6}$$

We employ Bayesian optimization to efficiently find the best-performing parameters.

### 4.2 Intergate Causal Debiasing into Training

While MID provides a practical inference-time fix, it operates externally and does not fundamentally alter the learned representations or the model's reliance on spurious features. To encourage the model to learn representations that are intrinsically more robust to spurious multimodal correlations, we shift our focus from inference-time correction to integrating counterfactual information directly into the training process.

#### 4.2.1 Foundations

A direct approach to incorporate the causal effect of spurious contexts is to model a bias-removed prediction (Eq. 5) into the loss function:

$$\mathcal{L} = -\mathbb{E}[\log \text{norm}(P(y|i,t) - \alpha \cdot P(y|\hat{i}) - \beta \cdot P(y|\hat{t}))], \tag{7}$$

where $\text{norm}(\cdot)$ is a softmax normalization. However, this approach is computationally expensive

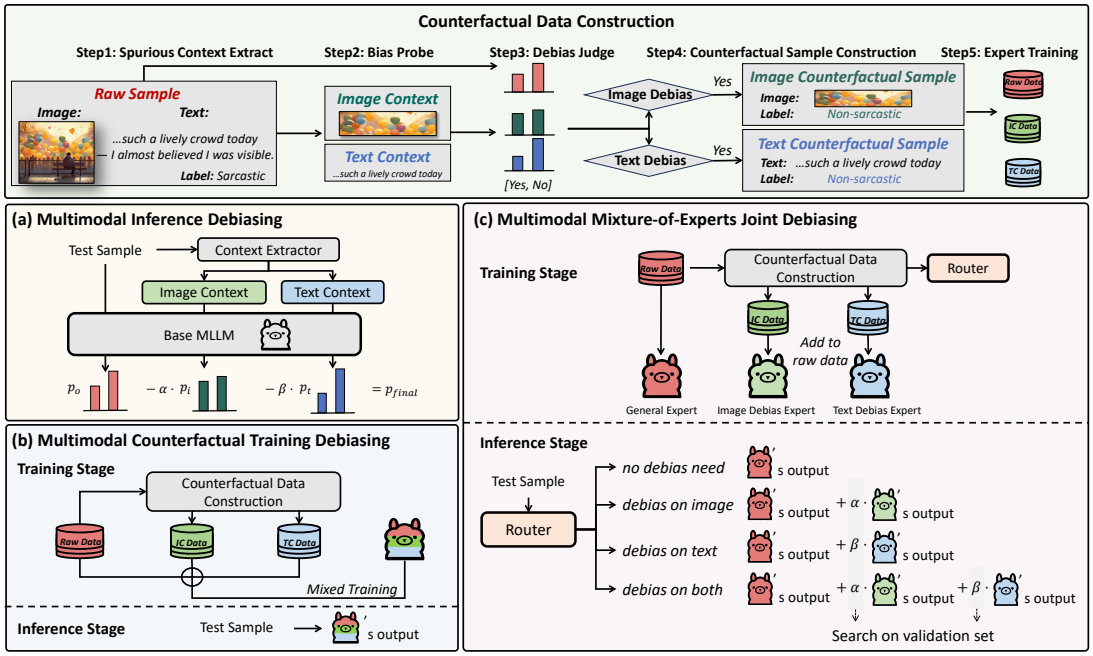


Figure 4: Overview of proposed multimodal debiasing frameworks: (a) inference-only debiasing using context extraction, (b) counterfactual training debiasing via mixed data augmentation, and (c) Mixture-of-Experts joint debiasing incorporating a dynamic router mechanism.

(three forward passes per step) and can suffer from numerical instability due to its internal subtraction.

To address these issues while maintaining the same fundamental training goals, we propose an alternative strategy centered on constructing counterfactual training objectives using reversed labels. The core objective remains consistent with $\mathcal{L}$ to maximize $P(y|i,t)$ on original inputs while minimizing reliance on spurious contexts. This is practically implemented by encouraging the model to predict an incorrect label $\hat{y}$, when presented with only the spurious context. This leads to the following training objective, $\mathcal{L}'$:

$$\mathcal{L}' = \underbrace{-\mathbb{E}[\log P(y|i,t)]}_{\text{(I) Maximize original accuracy}} + \underbrace{\mathbb{E}[\log P(y|\hat{i}) + \log P(y|\hat{t})]}_{\text{(II) Penalize spurious reliance via } y} \tag{8}$$

$$\overset{\text{approx.}}{=\!=} \text{(I)} \underbrace{-\mathbb{E}[\log P(\hat{y}|\hat{i})] - \mathbb{E}[\log P(\hat{y}|\hat{t})]}_{\text{(III) Minimize spurious accuracy via } \hat{y} \text{ (reversed label)}} . \tag{9}$$

This formulation provides an efficient way for training-time debiasing by incorporating counterfactual samples into the training data, requiring only a single forward pass per instance and thus enhancing efficiency.

While $\mathcal{L}'$ (Eq. 9) enables counterfactual training, MLLMs may hallucinate or misattribute spurious correlations. Indiscriminate use of naively generated counterfactuals can therefore introduce noise. To selectively identify informative samples for debiasing, we compare the original prediction $p_0(y)$ with counterfactual predictions $p_t(y), p_i(y)$ obtained from the base model (as in Sec. 4.1) on truth $y$ using a tolerance $\varepsilon$:

a. **No debiasing**: $p_t(y) > p_0(y) + \varepsilon \wedge p_i(y) > p_0(y) + \varepsilon$ (i.e., both contexts aid prediction).

b. **Visual debiasing only**: $p_i(y) + \varepsilon < p_0(y) < p_t(y) - \varepsilon$ (i.e., visual context harms, text aids/is neutral). Construct visual counterfactual.

c. **Textual debiasing only**: $p_t(y) + \varepsilon < p_0(y) < p_i(y) - \varepsilon$ (i.e., textual context harms, visual aids/is neutral). Construct textual counterfactual.

d. **Debias both modalities**: $p_t(y) + \varepsilon < p_0(y) \wedge p_i(y) + \varepsilon < p_0(y)$ (i.e., both contexts harm). Construct both counterfactuals.

e. **Exclude**: All other cases (e.g., within uncertainty margin).

A smaller $\varepsilon$ includes more samples for debiasing but can increase ambiguity near decision boundaries. We use $\varepsilon = 0.1$ in our experiments.

### 4.2.2 Training-Stage Only Debiasing

Building upon $\mathcal{L}'$ and counterfactual sample selection strategy, we could establish a specific training-time debiasing approach, **MCTD** (**Multimodal Counterfactual Training Debiasing**), illustrated in Fig. 4 (b). During inference, MCTD directly accepts test samples $(i,t)$ to produce outputs, re-

quiring no post-hoc adjustment.[2]

### 4.3 MME-JD: Multimodal Mixture-of-Experts Joint Debiasing

While the basic Counterfactual Training (Sec.4.2.2) improves internal representation robustness, its uniform application of counterfactual augmentation may not be ideal for all samples. To accommodate diverse bias types and intensities with sample-specific treatment, we propose MME-JD, a Multimodal Mixture-of-Experts Joint Debiasing approach, where each sample dynamically selects appropriate debiasing experts through a learned routing mechanism (as in Fig. 4 (c)).

#### 4.3.1 Architecture and Objectives

We employ three parallel expert branches:
1. General Expert, trained conventionally on original samples:

$$\mathcal{L}_{\text{GE}} = -\mathbb{E}[\log P(y|i,t)]. \tag{10}$$

2. Image Debiasing Expert, expliciting incorporating counterfactual visual examples selected by Sec 4.2.2 criterion b,d into training:

$$\mathcal{L}_{\text{IDE}} = -\mathbb{E}[\log P(y|i,t)] - \mathbb{E}[\log P(\hat{y}|\hat{i})]. \tag{11}$$

3. Text Debiasing Expert (TDE), analogously reducing textual bias via:

$$\mathcal{L}_{\text{TDE}} = -\mathbb{E}[\log P(y|i,t)] - \mathbb{E}[\log P(\hat{y}|\hat{t})]. \tag{12}$$

#### 4.3.2 Router and Inference-time Combination

A key component in MME-JD is a router module designed to dynamically determine the most suitable expert combination for any given input sample. The router takes the original and counterfactual inputs $(i, t, \hat{i}, \hat{t})$ and is trained as a classifier to predict an optimal expert strategy label, $c$, for each training instance. This strategy label $c$ is assigned by applying the heuristic counterfactual sample selection criteria detailed previously in Sec. 4.2.1. We assign $c \in \{0, 1, 2, 3\}$ corresponding to the required expert combination: *GE only* (0), *GE + IDE* (1), *GE + TDE* (2), and *GE + IDE + TDE* (3).

The Router is trained as a classifier to predict the expert strategy label c for each sample. Its training objective is to minimize the cross-entropy loss:

$$\mathcal{L}_{routing} = -\mathbb{E} \log Q(c|i,t,\hat{i},\hat{t}). \tag{13}$$

[2]We include MCTD as a distinct method here for experimental comparison, allowing us to assess its training-stage debiasing performance against inference-stage approaches.

| Method | Training Cost | Inference Cost |
|---|---|---|
| **Base Model** | 1× | 1× |
| **TFCD** (text debias) | 1× | ∼2× |
| **MID** | 1× | 3× |
| **MCTD** | ∼1× | 1× |
| **MME-JD** | ∼3× | 3× |

Table 1: Relative computational overhead vs. the base model; TFCD debiases only the text side.

During inference, for a given sample $(i, t)$, after generating $\hat{t}, \hat{i}$, the router first predicts the optimal expert strategy by

$$c^* = \arg\max Q(c|i,t,\hat{i},\hat{t}). \tag{14}$$

Then, the relevant experts (GE, and TDE/IDE if selected by $c^*$) process their respective inputs to produce outputs $p_{\text{GE}}, p_{\text{TDE}}, p_{\text{IDE}}$. The final output logits are computed by combining the outputs of the selected experts based on the strategy $c^*$.

$$\tilde{p} = \begin{cases} p_{\text{GE}}, & c^* = 0 \\ p_{\text{GE}} + \alpha_1 \, p_{\text{IDE}}, & c^* = 1 \\ p_{\text{GE}} + \beta_2 \, p_{\text{TDE}}, & c^* = 2 \\ p_{\text{GE}} + \alpha_3 \, p_{\text{IDE}} + \beta_3 \, p_{\text{TDE}}. & c^* = 3 \end{cases} \tag{15}$$

where $\alpha_c$ and $\beta_c$ are scalar weighting parameters that could be searched on validation set as Sec. 4.1.

### 4.4 Computational Overhead

Table 1 reports relative training and inference costs (normalized to the base model). Inference cost mainly reflects the number of forward passes over the VLM: TFCD adds one textual counterfactual (∼2×), MID adds two counterfactuals (3×). MCTD folds counterfactual supervision into training, keeping inference at 1×; MME-JD trains experts and routes at test time, yielding ∼3× for both training and inference.

## 5 Experimental Setup

In this section, we outline the experimental framework used to evaluate our proposed methods. Comprehensive details regarding datasets, evaluation metrics, implementation specifics, and all comparative methods are provided in App. C.

### 5.1 Datasets and Implementation Details

We evaluated our methods on two representative multimodal tasks: sarcasm detection using MMSD2.0 (Qin et al., 2023), and sentiment analysis using MVSA-Multi (Niu et al., 2016). Standard metrics Accuracy, Precision, Recall, and F1 were

| Method | Acc. | Prec. | Recall | F1 |
|---|---|---|---|---|
| HFM | 71.04 ± 0.29 | 64.92 ± 1.36 | 69.63 ± 0.93 | 67.01 ± 0.86 |
| Att-BERT | 80.10 ± 1.34 | 76.35 ± 2.73 | 77.76 ± 0.54 | 77.14 ± 1.46 |
| CMGCN | 79.92 ± 1.40 | 75.84 ± 1.16 | 78.10 ± 1.78 | 76.86 ± 1.45 |
| HKE | 76.47 ± 1.31 | 73.51 ± 1.00 | 71.62 ± 2.62 | 72.40 ± 1.72 |
| Multi-view CLIP | 85.35 ± 0.37 | 81.37 ± 1.12 | 87.05 ± 0.60 | 83.28 ± 1.10 |
| InternVL2.5 | 85.76 ± 0.57 | 79.91 ± 1.51 | 89.39 ± 0.28 | 84.38 ± 0.61 |
| InternVL2.5 + TFCD | 85.96 ± 0.34 | 80.00 ± 0.69 | 89.87 ± 1.02 | 84.65 ± 0.45 |
| InternVL2.5 + MID | 86.34 ± 0.59 | 80.62 ± 1.75 | 89.88 ± 0.63 | 85.00 ± 0.45 |
| InternVL2.5 + MCTD | 86.26 ± 0.39 | 80.33 ± 1.03 | 90.16 ± 0.51 | 84.96 ± 0.25 |
| InternVL2.5 + MME-JD | **86.84** ± 0.41 | **81.25** ± 1.06 | **90.26** ± 1.04 | **85.52** ± 0.41 |
| Qwen2-VL | 86.74 ± 0.29 | 81.14 ± 0.88 | 90.45 ± 0.81 | 85.54 ± 0.32 |
| Qwen2-VL + TFCD | 87.55 ± 0.35 | 82.70 ± 0.72 | 89.87 ± 0.53 | 86.14 ± 0.26 |
| Qwen2-VL + MID | 87.68 ± 0.51 | 83.14 ± 1.01 | 89.52 ± 0.84 | 86.21 ± 0.47 |
| Qwen2-VL + MCTD | 88.13 ± 0.34 | 82.51 ± 1.15 | **91.90** ± 0.62 | 86.95 ± 0.41 |
| Qwen2-VL + MME-JD | **88.42** ± 0.16 | **83.19** ± 0.80 | 91.61 ± 0.90 | **87.20** ± 0.23 |

Table 2: Comparison of our proposed methods (including MID, MCTD and MME-JD) with existing methods on dataset MMSD2.0.

| Method | Acc. | M-F1 | W-F1 |
|---|---|---|---|
| MVAN | 66.06 ± 0.97 | 54.45 ± 1.02 | 64.01 ± 1.12 |
| MGNNS | 67.49 ± 0.31 | 54.74 ± 1.72 | 64.37 ± 0.80 |
| CLMLF | 66.80 ± 0.71 | 54.93 ± 1.39 | 64.63 ± 0.89 |
| MDSE | 66.82 ± 1.26 | 55.12 ± 3.25 | 64.77 ± 0.92 |
| CF-MSA | 67.12 ± 1.28 | 55.18 ± 1.14 | 64.92 ± 1.01 |
| InternVL2.5 | 71.02 ± 0.35 | 58.37 ± 1.36 | 69.66 ± 0.35 |
| InternVL2.5 + MCIS | 71.08 ± 0.54 | 59.23 ± 0.81 | 70.00 ± 0.53 |
| InternVL2.5 + MID | 71.52 ± 0.95 | 60.20 ± 1.24 | 70.60 ± 0.84 |
| InternVL2.5 + MCTD | 71.52 ± 0.75 | 60.08 ± 1.30 | 70.52 ± 1.04 |
| InternVL2.5 + MME-JD | **71.86** ± 0.76 | **60.64** ± 1.02 | **70.87** ± 0.68 |
| Qwen2-VL | 69.79 ± 0.44 | 61.15 ± 0.71 | 69.98 ± 0.31 |
| Qwen2-VL + MCIS | 71.02 ± 0.47 | 60.35 ± 0.58 | 70.43 ± 0.33 |
| Qwen2-VL + MID | 71.46 ± 0.87 | 60.68 ± 0.93 | 70.92 ± 0.52 |
| Qwen2-VL + MCTD | 70.79 ± 0.66 | 60.55 ± 1.40 | 70.77 ± 0.73 |
| Qwen2-VL + MME-JD | **72.08** ± 0.46 | **62.42** ± 0.69 | **71.95** ± 0.35 |

Table 3: Comparison of our proposed methods with existing methods on MVSA-Multi. M-F1 denotes Macro-F1, while W-F1 refers to Weighted-F1.

utilized for sarcasm detection, while Accuracy, Macro-F1, and Weighted-F1 were used for sentiment analysis. Our experiments employed the large models Qwen2-VL-7B (Wang et al., 2024b) and InternVL2.5-4B (Chen et al., 2025), fine-tuned using LoRA (Hu et al., 2022). Our router model was based on CLIP (Radford et al., 2021). Inference-time debiasing hyperparameters were tuned via Bayesian optimization (Snoek et al., 2012).

### 5.2 Comparing Methods

We compared against task-specific multimodal baselines: Sarcasm Detection: **HFM**, **Attn-BERT**, **CMGCN**, **HKE**, **Multi-view CLIP**. Sentiment Analysis: **MVAN**, **MGNNS**, **CLMLF**, **MDSE**. and multimodal causal debiasing methods: **TFCD**, **MCIS**, **CF-MSA**. Detailed descriptions and selection reason are in Appendix C.4.

## 6 Results and Analysis

### 6.1 Main Results

Tab. 2 and 3 present our proposed methods' performance comparisons on two datasets against strong MLLM baselines (InternVL2.5, Qwen2-VL) and existing task-specific methods. While existing text-based counterfactual debiasing (e.g., TFCD, MCIS) offers slight improvements, our proposed MID, which extends counterfactual inference to the multimodal level, further enhances performance. Integrating counterfactual samples during training (MCTD) also yields gains, though not consistently surpassing MID, suggesting the continued necessity for inference-time debiasing.

Crucially, our comprehensive MME-JD model, which combines counterfactual training with a router and Mixture-of-Experts, consistently de-

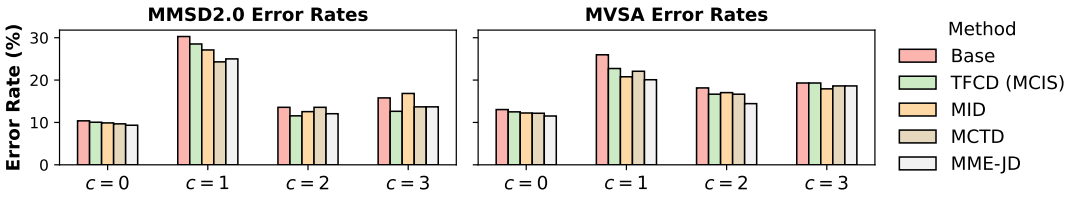


Figure 5: Comparison of error rates (%) across different debiasing methods on MMSD2.0 and MVSA datasets regarding to debiasing category. 0/1/2/3 refer to no/image/text/both debias need.

| Methods | R. | C. | M. | MMSD2.0 | | MVSA-Multi | |
|---|---|---|---|---|---|---|---|
| | | | | Acc. | F1 | Acc. | M-F1 |
| *Inference-only* | | | | | | | |
| Qwen2-VL | | | | 71.94 | 65.89 | 67.95 | 49.28 |
| + MID | ✗ | ✗ | ✗ | 73.22 | 66.97 | 69.12 | 51.41 |
| + MRID | ✓ | ✗ | ✗ | **74.88** | **68.21** | **69.29** | **52.83** |
| *Training backbone* | | | | | | | |
| Qwen2-VL | | | | 86.74 | 85.54 | 69.79 | 61.15 |
| + MID | ✗ | ✗ | ✗ | 87.68 | 86.21 | 71.46 | 60.68 |
| + MRID | ✓ | ✗ | ✗ | 88.09 | 86.58 | **72.58** | 61.94 |
| + MCTD | ✗ | ✓ | ✗ | 88.13 | 86.95 | 70.79 | 60.55 |
| + MME-JD | ✓ | ✓ | ✓ | **88.42** | **87.20** | 72.08 | **62.42** |

Table 4: Performance comparison when different components employed in base model. "R.", "C.", and "M." refer to the Routing mechanism, Counterfactual training, and Mixture-of-Experts, respectively. We introduce an additional Router component to MID (MRID) to specifically assess the Router's contribution. See Appendix D for implementation details.

livered the most substantial improvements. It achieved top F1 scores of 85.52% (InternVL2.5) and 87.20% (Qwen2-VL) on MMSD2.0. These results confirm MME-JD's effectiveness in mitigating multimodal spurious correlations through the synergy of training-time strategies, expert routing, and inference-time debiasing.

### 6.2 Ablation Study

Ablation studies (Tab. 4) demonstrate the individual and combined contributions of our framework's components. For inference-time debiasing, the Router (MRID) significantly enhanced performance over MID. For training debiasing, counterfactual training (MCTD) alone improved F1 scores (e.g., +1.4 on MMSD2.0 over baseline), and the complete MME-JD model, integrating the Router, Counterfactual Training, and Mixture-of-Experts (MoE), achieved the highest F1 scores (87.20 on MMSD2.0, 62.42 on MVSA-Multi). This underscores that the deep integration of all components is crucial for optimal performance and learning robust, debiased representations.

Further analysis (Tab. 5) on modality-specific debiasing shows that while individual image or text debiasing provides modest gains, neither matches the performance of the fully integrated MME-JD, which highlights the necessity of simultaneously addressing biases across both modalities for optimal multimodal understanding.

### 6.3 Analysis on Error Rate on Categories

To further understand the impact of our methods, we analyzed error rates across sample categories with distinct bias types following 4.3.2. Results shown in Fig. 5 indicate that samples requiring any form of debiasing generally presented higher difficulty than those without debiasing needs ($c = 0$). Among these, text-only biased samples ($c = 2$) consistently showed better debiasing effectiveness compared to image-only biased samples ($c = 1$). TFCD exhibited notable limitations when addressing image-biased samples ($c = 1$), aligning with our expectations given its text-focused design. MID and MCTD each demonstrated advantages in different categories, underscoring the importance of integrating both inference and training-based debiasing approaches. The proposed MME-JD method, although not always outperforming all other methods within each individual category, achieved the best overall performance by effectively integrating multiple debiasing strategies.

### 6.4 Analysis on Router

The router component dynamically assigns input samples to suitable experts within our MME-JD framework. To evaluate its effectiveness, we compared a trained router against an oracle router (Tab. 6). The oracle router, serving as an ideal upper bound (following Sec. 4.3.2), demonstrated notably superior performance, highlighting both the substantial potential of the MME-JD expert architecture and the trained router's accuracy as a key performance bottleneck.

| Methods | Acc | Prec | Recall | F1 |
|---|---|---|---|---|
| Qwen2VL | **86.74** | **81.14** | **90.45** | **85.54** |
| + IDE | 87.13 | 81.2 | 91.13 | 85.89 |
| + TDE | 87.05 | 81.22 | 90.93 | 85.8 |
| + MME-JD | **88.42** | **83.19** | **91.61** | **87.2** |
| InternVL2.5 | **85.76** | **79.91** | **89.39** | **84.38** |
| + IDE | 86.17 | 80.26 | 90.15 | 84.92 |
| + TDE | 86.21 | 80.36 | 89.97 | 84.89 |
| + MME-JD | **86.84** | **81.25** | **90.26** | **85.52** |

Table 5: Ablation study on the effectiveness of image and text debiasing experts within the proposed MME-JD framework.

| Method | Router | Acc. | Prec. | Recall | F1 |
|---|---|---|---|---|---|
| Qwen2VL | trained | 88.42 | 83.19 | 91.61 | 87.20 |
| + MME-JD | *oracle* | *92.40* | *87.46* | *96.14* | *91.59* |
| InternVL2.5 | trained | 86.84 | 81.25 | 90.26 | 85.52 |
| + MME-JD | *oracle* | *89.72* | *83.93* | *93.45* | *88.43* |

Table 6: Comparison of model performance using a trained router versus an oracle router in MME-JD.

| Model | Cate. | Prec. | Recall | F-0.5 |
|---|---|---|---|---|
| Router for Qwen2VL | $c$=0 | 77.55 | 96.40 | 80.7 |
| | $c$=1 | 45.54 | 13.46 | 30.83 |
| | $c$=2 | 32.00 | 12.72 | 18.20 |
| Router for InternVL2.5 | $c$=0 | 81.36 | 77.12 | 80.48 |
| | $c$=1 | 27.86 | 13.04 | 22.71 |
| | $c$=2 | 29.82 | 42.89 | 31.75 |

Table 7: Router classification performance for different debiasing types. $c$ indicates debiasing category: 0 (No Debias), 1 (Image Debias), and 2 (Text Debias).

| True \ Pred | No Debias | Image | Text | Both |
|---|---|---|---|---|
| No Debias | 1,734 | 29 | 72 | 5 |
| Image Debias | 140 | 10 | 3 | 1 |
| Text Debias | 230 | 3 | 36 | 1 |
| Double Debias | 129 | 3 | 2 | 11 |

Table 8: Confusion matrix of the trained router.

Tab. 7 further shows that while the trained router effectively identifies samples not requiring debiasing ($c$=0), its performance degrades sharply for image ($c$=1) and text ($c$=2) debiasing cases. We attribute this weakness primarily to severe training-data imbalance—only a small minority of samples require specialized debiasing—so errors on these critical cases disproportionately limit overall MME-JD effectiveness.

To diagnose the failure modes more precisely, we analyze the class-wise confusion matrix in Tab. 8. The router exhibits a strong conservative bias toward *No Debias*: 92.7% of all predictions fall into this branch (2233/2409), yielding high recall for *No Debias* (94.2%) but low recall for *Image*/*Text*/*Both* (6.5%/13.3%/7.6%). Among all non-*No Debias* ground-truth instances, 97.5% of the errors are conservative *No Debias* predictions (499 out of 512), indicating that the router abstains when uncertain rather than confusing debiasing types. Indeed, after excluding the *No Debias* branch, the largest entry in each remaining row lies on the diagonal (10/36/11 for *Image*/*Text*/*Both*), implying limited cross-type confusion once the router commits to debiasing. Class-precision by prediction is 77.7% (*No*), 22.2% (*Image*), 31.9% (*Text*), and 61.1% (*Both*). Together with the distribution statistics (Appendix, Tab. 10), these results suggest that threshold calibration or cost-sensitive training (e.g., class-balanced losses or utility-aware decision thresholds) could trade additional recall on debiasing branches for a modest precision drop, improving end-to-end routing without substantially increasing cross-class swaps, which drives our future work.

### Qualitative evidence

We provide two representative sarcasm–detection case studies in Appendix F. Both examples apply our masking-based debiasing to discount background-only evidence while preserving task-relevant semantics.

## 7 Conclusion

In this paper, we addressed the critical issue of spurious multimodal biases that impair the robustness and generalization of MLLMs. Prior approaches lacked comprehensive multimodal analysis and training-time adaptability. We first developed a fine-grained multimodal causal framework, explicitly distinguishing spurious context from semantic content, thereby extending previous inference-time adjustments to multimodal scenarios. Building upon this, we introduced MME-JD, incorporating counterfactual content into training and employing a Mixture-of-Experts architecture with dynamic routing for adaptive, sample-specific debiasing. Extensive experiments demonstrate that MME-JD significantly surpasses simpler debiasing methods and existing benchmarks. Future research will explore enhancing counterfactual generation techniques and refining router designs for enhancement.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (62076008).

## Limitation

Despite the improvements demonstrated, our approach has several limitations that open avenues for future research:

**Accuracy of Spurious Context Identification:** The efficacy of our framework critically depends on the precise identification of spurious contexts versus core semantic content. Our current methodology, employing prompt engineering and attention-based mechanisms for semantic extraction, faces inherent challenges. Prompt engineering can inadvertently introduce new biases, while attention scores may not always reliably pinpoint genuinely spurious regions. Therefore, despite the advanced nature of LLMs, inaccuracies in this initial extraction phase can propagate the effectiveness of subsequent debiasing efforts. Future studies should integrate additional robust and objective methodologies alongside these advanced models to further enhance extraction accuracy and reliability.

**Router Mechanism Complexity and Stability:** While the router mechanism in MME-JD is designed to manage the variability of counterfactual sample quality, its own training and calibration require careful attention and can be intricate. This complexity might risk it becoming a performance bottleneck or introducing sensitivities. For instance, ensuring the router generalizes well and avoids 'dependency loops' (where its performance is overly reliant on specific states of other components it's trying to manage) is crucial. Developing more inherently robust router architectures, along with streamlined and stable training strategies, presents a key direction for future improvement.

In addition, our router faces severe class imbalance, since only around 10% of MMSD2.0 samples require explicit debiasing, as detailed in Table 10. We did not apply resampling, loss weighting, or focal loss during router training, prioritizing precision over recall. This conservative choice avoids mismatched debiasing but leads to weaker performance on minority categories (e.g., image-only debias). Exploring class-aware or curriculum-based router training strategies is an important direction for future work.

**Linearity Assumption in Causal Mediation:** Our current causal mediation framework approximates bias removal using linear relationships. However, the interactions between biases and core semantics within complex Multimodal Large Language Models (MLLMs) are likely to be highly nonlinear. This linearity assumption may therefore limit our model's capacity to fully neutralize biases, particularly in scenarios requiring simultaneous debiasing of both textual and visual modalities (e.g., in Fig 5, MME-JD did not perform the best, and sometimes text-specific methods (TFCD) could surpass multimodal methods). A significant avenue for future research lies in investigating and integrating nonlinear debiasing strategies to more comprehensively model and mitigate these intricate multimodal interactions, thereby further enhancing model robustness.

**Inference Efficiency of MME-JD** The MME-JD framework's adaptive inference process, while effective for debiasing, can impact its overall efficiency. Key steps, including the on-the-fly generation of counterfactual inputs $(\hat{t}, \hat{i})$, router processing, and the subsequent execution of selected expert(s), collectively contribute to increased latency and computational resource demands compared to simpler methods. While we currently employ independent models for each expert, we have tried MoE-LoRA (Luo et al., 2024) to reduce the training and inference cost, but it did not yield comparable debiasing performance. How to develop more successful parameter-efficient expert architectures that maintain high debiasing capabilities is worth considering in the future.

## Ethical Considerations

**Potential Risks** While our methods effectively reduce known biases, they depend on identifying spurious correlations accurately. Errors or oversights in distinguishing between semantic and spurious contexts could inadvertently reinforce existing biases or introduce new ones. Practitioners must be cautious and continually validate debiasing processes to prevent unintended consequences.

**Use of AI Assistants** We have employed ChatGPT as a writing assistant, primarily for polishing the text after the initial composition.

## References


Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

Alexandru-Costin Băroiu and Ștefan Trăușan-Matu. 2022. Automatic sarcasm detection: Systematic literature review. *Information*, 13(8).

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*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Koyel Chakraborty, Siddhartha Bhattacharyya, and Rajib Bag. 2020. A survey of sentiment analysis from social media data. *IEEE Transactions on Computational Social Systems*, 7(2):450–464.

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

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024b. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, Miami, Florida, USA. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8020–8028. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Michal Golovanevsky, William Rudman, Vedant Palit, Ritambhara Singh, and Carsten Eickhoff. 2025. What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. *Preprint*, arXiv:2406.16320.

Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Wang. 2022. CPL: Counterfactual prompt learning for vision and language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3407–3418, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Parsa Hosseini, Sumit Nawathe, Mazda Moayeri, Sriram Balasubramanian, and Soheil Feizi. 2025. Seeing

what's not there: Spurious correlation in multimodal llms. *Preprint*, arXiv:2503.08884.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2024. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11975–11985.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jae Myung Kim, A. Sophia Koepke, Cordelia Schmid, and Zeynep Akata. 2023. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2585–2595.

Tiep Le, VASUDEV LAL, and Phillip Howard. 2023. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In *Advances in Neural Information Processing Systems*, volume 36, pages 71195–71221. Curran Associates, Inc.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.

Jingzhe Li, Chengji Wang, Zhiming Luo, Yuxian Wu, and Xingpeng Jiang. 2024. Modality-dependent sentiments exploring for multi-modal sentiment classification. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7930–7934.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2282–2294, Seattle, United States. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Hongcheng Liu, Pingjie Wang, Zhiyuan Zhu, Yanfeng Wang, and Yu Wang. 2024. CE-VDG: Counterfactual entropy-based bias reduction for video-grounded dialogue generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2958–2968, Torino, Italia. ELRA and ICCL.

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*, pages 4995–5006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *Preprint*, arXiv:2402.12851.

Badr-Eddine Marani, Mohamed Hanini, Nihitha Malayarukil, Stergios Christodoulidis, Maria Vakalopoulou, and Enzo Ferrante. 2024. *ViG-Bias: Visually Grounded Bias Discovery and Mitigation*, page 414–429. Springer Nature Switzerland.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling*, page 15–27.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12700–12710.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. 2023. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2856–2861.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, Online. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Judea Pearl. 2022. *Causal Diagrams for Empirical Research (With Discussions)*, 1 edition, page 255–316. Association for Computing Machinery, New York, NY, USA.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *Preprint*, arXiv:2204.03162.

Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. 2024. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 82235–82264. Curran Associates, Inc.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, and 5 others. 2024a. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *Preprint*, arXiv:2408.01319.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shuqi Wang, Xufeng Duan, and Zhenguang Cai. 2024c. A multimodal large language model "foresees" objects based on verb information but not gender. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 435–441, Miami, FL, USA. Association for Computational Linguistics.

Wei Wang, Boxin Wang, Ning Shi, Jinfeng Li, Bingyu Zhu, Xiangyu Liu, and Rong Zhang. 2021. Counterfactual adversarial learning with representation interpolation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4809–4820, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. 2024. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15669–15680, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for large language models. *Preprint*, arXiv:2410.15319.

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 *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LVIII*, page 464–481, Berlin, Heidelberg. Springer-Verlag.

Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang. 2024b. Robust emotion recognition in context debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12447–12457.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2021a. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021b. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339, Online. Association for Computational Linguistics.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M. Rehg, and Aidong Zhang. 2024. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *Preprint*, arXiv:2406.17126.

Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024a. Api: Attention prompting on image for large vision-language models.

Tao Yu, Yi-Fan Zhang, Chaoyou Fu, Junkang Wu, Jinda Lu, Kun Wang, Xingyu Lu, Yunhang Shen, Guibin Zhang, Dingjie Song, Yibo Yan, Tianlong Xu, Qingsong Wen, Zhang Zhang, Yan Huang, Liang Wang, and Tieniu Tan. 2025. Aligning multimodal llm with human preference: A survey. *Preprint*, arXiv:2503.14504.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13807–13816.

Jie Zhang, Xiaosong Ma, Song Guo, Peng Li, Wenchao Xu, Xueyang Tang, and Zicong Hong. 2024a. Amend to alignment: Decoupled prompt tuning for mitigating spurious correlation in vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59505–59519. PMLR.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024b. Debiasing multimodal large language models. *Preprint*, arXiv:2403.05262.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Mingjia Zhang, and Baobao Chang. 2024. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. *Preprint*, arXiv:2411.14279.

Guangmin Zheng, Jin Wang, Xiaobing Zhou, and Xuejie Zhang. 2024a. Enhancing semantics in multimodal chain of thought via soft negative sampling. *Preprint*, arXiv:2405.09848.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024b. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. 2024a. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 291–300, New York, NY, USA. Association for Computing Machinery.

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 Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6687–6695. International Joint Conferences on Artificial Intelligence Organization. Main Track.

## A Causal Inference Theory and Counterfactual Causal Inference

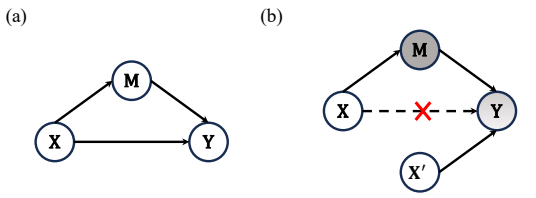


Figure 6: (a) An example of a causal graph, where $X$ transmits information to the outcome $Y$ via the mediator variable $M$. (b) The process of counterfactual intervention, observing the influence of $X$ on $Y$ via the mediator variable $M$ by interfering the link between $X$ and $Y$.

This section first introduces the theory of causal inference, which serves as the cornerstone for multimodal causal mediation analysis in Sec. 3.1.

Causal graphs are highly generalized analytical tools used to reveal causal dependencies between variables. The prediction process of a model can be defined as a Directed Acyclic Graph (DAG) $G = V, E$, where the set of nodes $V$ represents a series of intermediate factors involved in the prediction process, and the set of edges $E$ denotes the causal effects between them. Figure 6 presents an example of a causal graph with three variables. In this graph, uppercase letters denote random variables, and lowercase letters denote their actual observed values. The causal relationship from a reference variable $X$ to an outcome variable $Y$ comprises two paths:

- $X \rightarrow Y$: The direct effect path, representing the direct effect of $X$ on $Y$.
- $X \rightarrow M \rightarrow Y$: The indirect effect path, representing the effect of $X$ on $Y$ mediated by the intermediate variable $M$.

Counterfactual inference involves applying counterfactual treatment conditions to the model and observing the outcome of the original input under these new conditions, thereby deducing the influence of the counterfactual factors. This process can be formally expressed. The outcome for the original input is:

$$Y_{x,m} = Y(X = x, M = m), \tag{16}$$

where $m = M_x = M(X = x)$. By applying a counterfactual treatment to $x$ to yield $x^*$, we obtain the counterfactual outcome:

$$Y_{x^*,M_{x^*}} = Y(X = x^*, M = M(X = x^*)), \tag{17}$$

where $M_{x^*}$ represents the value of the mediator $M$ when $X$ is changed to $x^*$. Furthermore, a scenario can be constructed where X is counterfactually set to $x^*$, but the mediator $M$ retains the value it would have taken under the original input $x$. The outcome in this case is $Y_{x^*,M_x}$.

Causal effects quantify the difference between the outcome after intervening on a reference variable and the outcome in its natural (pre-intervention) state. According to causal theory, when the variable $X$ is changed from $x$ to $x^*$, the Total Effect (TE) on the model's outcome is defined as:

$$TE = \text{DIFF}(Y_{x,M_x}, Y_{x^*,M_{x^*}}), \tag{18}$$

where DIFF represents a function that measures the difference between the predicted outcomes before and after the intervention. The Total Effect (TE) can be decomposed into the Natural Direct Effect (NDE) and the Total Indirect Effect (TIE). The NDE refers to the impact of a change in the causal variable $X$ (from $x$ to $x^*$) on the outcome $Y$ when the mediator variable $M$ is held at the level it would naturally assume if $X$ were counterfactually set:

$$NDE = \text{DIFF}(Y_{x,M_{x^*}}, Y_{x^*,M_{x^*}}). \tag{19}$$

This aims to isolate the impact of the reference variable on the outcome variable, holding the mediator constant at $M_{x^*}$.

The Total Indirect Effect (TIE) represents the effect of the reference variable $X$ indirectly influencing the outcome variable $Y$ through the mediator $M$. It is calculated by subtracting NDE from TE:

$$TIE = TE - NDE = \text{DIFF}(Y_{x,M_x}, Y_{x,M_{x^*}}). \tag{20}$$

Thus, depending on the nature of the mediator $M$, NDE or TIE can be leveraged for unbiased estimation regarding the outcome variable Y:

- If the mediator $M$ is identified as a variable that potentially introduces bias and is not conducive to the final prediction. In this case, we hold $M$ at the value $M_{x^*}$, and observe the unbiased effect of $X$ on $Y$. This process is achieved through NDE.
- If the mediator $M$ represents the pathway or features of interest, and only the effect mediated through $M$ is desired for an accurate prediction. Here, TIE can be used to calculate the indirect effect of $X$ on $Y$ via $M$, yielding an unbiased estimate of this mediated effect.

In this paper, we define M as the bias variable (representing spurious textual and visual context), and therefore, we employ NDE for unbiased estimation.

## B Counterfactual Content Construction

Guided by the multimodal causal mediation framework, we explicitly construct counterfactual samples that isolate textual and visual spurious contexts. Existing counterfactual generation approaches typically rely on rule-based heuristics, manipulation of latent representations or pre-existing annotations in datasets, and thus lack the granularity and flexibility for our detailed causal framework. In contrast, we propose leveraging large MLLMs to automate counterfactual input generation at the raw input level.

### B.1 Counterfactual Text Input Construction

For textual inputs, we aim to clearly distinguish between key semantic content ($T_{semantic}$) and spurious context which is potentially bias-inducing ($T_{spurious}$). The counterfactual textual input is generated through the following procedure:

**Identifying Semantic Content via Prompting.** We prompt a large language model to extract primary semantic content relevant to the prediction task. We avoid directly identifying spurious context since irrelevant segments often lack stable semantic indicators, causing imprecise model responses. Instead, extracting key semantic segments, which inherently possess clear and consistent semantic features, provides a more precise foundation, allowing us to reliably obtain spurious context by subtracting the identified semantic content from the original text.

**Generating Context-only Counterfactuals.** Once critical semantic components are identified, we replace these segments with a neutral placeholder token (`[MASK]`), ensuring grammatical fluency while preserving only the bias-prone content:

$$T_{spurious} = T \backslash T_{semantic} \tag{21}$$

### B.2 Counterfactual Image Input Construction

Analogous to the textual scenario, we need to isolate visual content relevant to the prediction task ($I_{semantic}$) from spurious visual contexts ($I_{spurious}$). Due to the complexity of accurately distinguishing these regions, we propose an automated approach based on attention mechanisms from MLLM, enabling fine-grained visual counterfactual generation.

**Extracting Visual Attention via Neutral Prompts.** Analogous to our textual approach, we prompt the vision-language model with neutral descriptions rather than task-specific instructions. Unlike prior methods, which often rely on coarse or manual visual annotations, our method proactively extracts fine-grained visual content critical to the prediction task. Motivated by Yu et al. (2024a), task-specific prompts would introduce semantic biases related to internal task priors, potentially distorting the saliency distribution. In contrast, neutral prompting allows the model to identify image regions purely based on intrinsic visual importance.

Formally, given textual tokens $X_{text}$ and image patches $X_{image}$, the model generates descriptive tokens $Y_{out}$. Attention scores from $Y_{out}$ to $X_{image}$ are extracted from deeper transformer layers (e.g., layers 29–31 of the 32-layer Qwen2-VL-7B) to capture richer, more fine-grained multimodal interactions (Wang et al., 2024c). The attention score for the image patch at spatial position $(i, j)$ is computed as:

$$\phi_{i,j} = \sum_{h,m} A^{h}_{m,t}, \quad t = j + P \cdot (i - 1). \tag{22}$$

Here, $A^{h}_{m,t}$ denotes attention weights from the $m$-th token of the generated sequence to the $t$-th image patch across head $h$, and $P$ denotes the number of patches per row in the patch sequence.

**Mask Enhancement for Counterfactual Image Generation** To generate visually coherent and semantically meaningful counterfactual images from the raw attention masks, we apply several post-processing steps. First, the attention scores are normalized into a standardized range [0, 1] to stabilize mask values and facilitate subsequent manipulation. Second, we enhance the contrast of these normalized masks by scaling, thus emphasizing visually salient regions and clearly distinguishing critical visual elements from background noise. Third, a spatial smoothing operation (convolutional filtering) is performed to prevent abrupt transitions between masked and unmasked regions, maintaining visual continuity. Finally, we interpolate the smoothed masks back to the original image dimensions and

**Algorithm 1** Counterfactual Image Generation via Attention Masks

**Require:** MLLM $model$, image $I$, prompt $P$, number of model layers $L$, image dimensions $H, W$, image patch grid dimensions $(patch_h, patch_w)$, enhancement factor $\alpha$
**Ensure:** counterfactual image $I_{cf}$

$Y_{out} \leftarrow model$.generate($I$, $P$) ▷ Generate the neutral image analysis
attns$\leftarrow model$.extract_attentions($I$, $P$, $Y_{out}$) ▷ Fetch attentions from $Y_{out}$ to $I$
focused_attns$\leftarrow$ attns[L-4:L-1, :, :] ▷ Select certain layers
$\phi \leftarrow$ MeanPool(focused_attns) ▷ Keep only final dimension only
$\text{mask}_{\text{2d}} \leftarrow$ Reshape($\phi$, $(patch_h, patch_w)$) ▷ Reshape to 2 dimensional
$\text{mask}_{\text{norm}} \leftarrow$ Normalize($\text{mask}_{\text{2d}}$) ▷ Normalization
$\text{mask}_{\text{enhanced}} \leftarrow \alpha \cdot \text{mask}_{\text{norm}}$ ▷ Enhancement
$\text{mask}_{\text{smooth}} \leftarrow$ Conv2D($\text{mask}_{\text{enhanced}}$, kernel) ▷ Smmothing
$\text{mask}_{\text{resized}} \leftarrow$ Interpolate($\text{mask}_{\text{smooth}}$, size=$(H, W)$)
$I_{cf} \leftarrow$ AlphaBlend($I$, $\text{mask}_{\text{resized}}$) ▷ Blending with original image
**return** $I_{cf}$

blend them seamlessly with the original images via alpha blending. This procedure ensures that the resulting counterfactual images preserve contextual coherence while effectively highlighting the identified visual regions. Alg. 1 summarizes this detailed workflow precisely.

### B.3 Quality Control via Router

While the proposed automated approach efficiently constructs fine-grained counterfactual inputs, the inherent complexity and variability of multimodal content might occasionally lead to suboptimal or noisy counterfactual samples. To mitigate potential quality issues arising from these automatically generated inputs, we incorporate a router mechanism that dynamically assesses their suitability for subsequent causal analysis. Specifically, the router determines whether each generated counterfactual input should be utilized or discarded based on its semantic reliability and consistency. Further details regarding the design and operational logic of the router are elaborated in later sections.

## C Detailed Experimental Setup

### C.1 Datasets and Evaluation Metrics

Our experiments were conducted on two distinct multimodal datasets to address sarcasm detection and sentiment analysis. Statistical details for both datasets are presented in Tab. 9. For multimodal sarcasm detection, we utilized the MMSD2.0 dataset (Qin et al., 2023). We directly adopted the official dataset partitions for training, validation, and testing as provided by the original authors. For multimodal sentiment analysis, we employed the MVSA-Multi subset of the MVSA dataset (Niu et al., 2016). MVSA is a widely recognized benchmark in this domain, constructed from Twitter posts. The MVSA dataset consists of two subsets: MVSA-Single (MVSA-S), which contains 5,129 samples, each annotated by a single annotator, and MVSA-Multi (MVSA-M), which includes 19,600 samples, with each sample receiving three independent annotations. We selected MVSA-Multi for our experiments due to its superior annotation reliability and lower noise levels, thereby supporting more robust experimental findings. For the MVSA-Multi dataset, we randomly partitioned the data into training, validation, and testing sets, adhering to a 7.5:1:1 ratio.

To comprehensively assess model efficacy, we used specific evaluation metrics for each task. For sarcasm detection (MMSD2.0), where "Non-sar" instances outnumber "Sarcasm," metrics like Precision, Recall, and F1-score were used alongside Accuracy to ensure the model's ability to identify the minority sarcastic class wasn't obscured. Similarly, for the more significantly imbalanced MVSA-Multi sentiment dataset (with "Negative" as a clear minority), Accuracy was supplemented with Macro-F1 and Weighted-F1 to provide a fairer assessment of performance across all classes

### C.2 Implementation Details

Our experiments utilized two powerful large multimodal model series, Qwen2-VL (Wang et al., 2024b) and InternVL2.5 (Chen et al., 2025), specifically the Qwen2-VL-7B and InternVL2.5-4B versions, trainable on a single NVIDIA A100 80G GPU. We employed LoRA (Hu et al., 2022)

| | MMSD2.0 | | MVSA-Multi | | |
|---|---|---|---|---|---|
| | #Sarcasm | #Non-sar | #Positive | #Neutral | #Negative |
| **Train** | 8642 | 11174 | 8954 | 3461 | 1023 |
| **Valid** | 959 | 1451 | 1186 | 483 | 124 |
| **Test** | 959 | 1450 | 1177 | 463 | 151 |

Table 9: Statistics of dataset MMSD2.0 and MVSA-Multi

for fine-tuning, using the LLaMA-Factory framework (Zheng et al., 2024b) for Qwen2-VL-7B (with data in dialogue format) and official scripts for InternVL2.5-4B. Key LoRA parameters included a rank of 16, 10 epochs, a learning rate of 4e-5, weight decay of 0.01, a batch size of 8, and 2 gradient accumulation steps, accelerated with DeepSpeed Stage 1 (Rajbhandari et al., 2020). Checkpoints were saved every 500 steps, with the best model selected based on minimum validation loss. Each expert training requires 1 day around.

We choose CLIP (Radford et al., 2021) to serve as the backbone for our router model. This Router was trained for 15 epochs using the Transformers library (Wolf et al., 2020), with the best checkpoint chosen based on the F-0.5 score on the validation set to prioritize precision in identifying samples needing correction. Optimal hyperparameters for the inference-time debiasing process were identified using Bayesian optimization (Snoek et al., 2012) via the skopt package, employing a Gaussian process model with a maximum of 50 function evaluations. To ensure the robustness and reliability of our experimental results, we conducted each proposed method five times independently and reported the average performance across these trials.

### C.3 Debiasing Category Distribution

For completeness, we summarize the training-set distribution of debiasing categories (InternVL2.5). The majority of samples are labeled *None*, yet a non-trivial portion requires single- or dual-modality debiasing, motivating adaptive expert routing.

### C.4 Comparing Methods

In order to comprehensively evaluate the performance of our proposed methods, we considered several settings and selected representative methods for comparison:

(1) Methods specific for downstream tasks. For multimodal sarcasm detection, we choose **HFM** (Cai et al., 2019), which introduced a hierarchical fusion approach for multi-modal sarcasm detection, distinctively treating image attributes as a third modality alongside text and image features; **Attn-BERT** (Pan et al., 2020), which applied BERT-based architectures with a significant attention mechanism to model user expressions in multi-modal content; **CMGCN** (Liang et al., 2022), constructing of an instance-specific cross-modal graph to explicitly map relationships between image objects and textual words. It then employed a graph convolutional network to learn and identify incongruity within these structures; **HKE** (Liu et al., 2022) focused on advancing representation learning for heterogeneous knowledge graphs, embed diverse types of entities and relations, thus capturing complex semantics; **Multi-view CLIP** (Qin et al., 2023) utilized CLIP by processing information from multiple views (text, image, and their interaction) to capture multi-grained cues.

For multimodal sentiment analysis, we choose **MVAN** (Yang et al., 2021a), which introduced a multi-view attention mechanism that captures image features from both object and scene perspectives, and employed a memory network that is continually updated to obtain deep semantic features; **MGNNS** (Yang et al., 2021b), constructed separate graphs for text and image modalities to capture the global co-occurrence characteristics of the dataset and utilized multi-channel graph neural networks to learn multimodal representations with a multi-head attention mechanism for in-depth fusion; **CLMLF** (Li et al., 2022) combined contrastive learning with a multi-layer fusion strategy to help the model learn common sentiment-related features across modalities; **MDSE** (Li et al., 2024) focused on identifying sentiment expressions specific to individual modalities by leveraging semi-supervised variational autoencoders.

(2) Methods for multimodal causal debias. **TFCD** (Zhu et al., 2024b) targeted at biases in multi-modal sarcasm detection, specifically the model's over-reliance on frequently occurring non-sarcastic words and static co-occurrences between training data labels and modal features. The challenge of non-sarcastic word bias in the tex-

| Dataset / Category | None | Image | Text | Both |
|---|---|---|---|---|
| MMSD2.0 (n=19,816) | 86.8% | 4.2% | 7.0% | 2.0% |
| MVSA-Multi (n=13,438) | 93.2% | 2.5% | 2.1% | 2.2% |

Table 10: Summary of debiasing-category prevalence (percentage of training set).

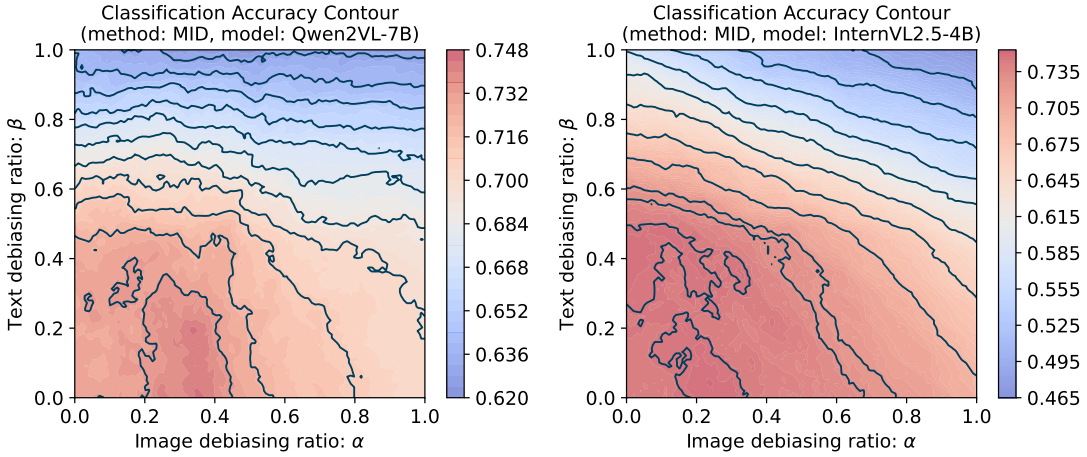


Figure 7: Contour plots illustrating the influence of image debiasing ratio ($\alpha$) and text debiasing ratio ($\beta$) on classification accuracy when applying MID to Qwen2-VL-7B and InternVL2.5-4B models (vanilla model without training).

tual modality, as identified by TFCD, aligns with our the text branch in our proposed causal graph. **MCIS** (Yang et al., 2024a) designed for multi-modal sentiment analysis and explicitly differentiates between two common types of biases: utterance-level label bias and word-level context bias. The approach adopted by MCIS shares significant similarities with TFCD, with the primary distinction being the application task. **CF-MSA** (Chen et al., 2024a) addressed biases in both textual and visual modalities for multi-modal sentiment analysis. While CF-MSA shares our focus on addressing biases in both modalities, it is implemented based on BERT and operates at the feature representation level, and estimated unimodal bias by completely masking one modality.

## D MRID: Multimodal Router-Guided Inference Debiasing

To further analyze the effectiveness of the dynamic routing mechanism from MME-JD in an inference-only setting, we introduce Multimodal Router-Guided Inference Debiasing (MRID). This approach adapts the standard MID framework by incorporating the router to selectively apply modality-specific debiasing, rather than uniformly correcting for both modalities.

Specifically, following Sec. 4.1, we obtain outputs under three scenarios: original prediction ($p_0$), text-spurious prediction $p_t$ and image-spurious prediction ($p_i$). The MME-JD router (as described in Section 4.3.2), which was trained to predict an optimal expert strategy is then employed. The router takes the set of inputs $(i, t, \hat{i}, \hat{t})$and outputs a strategy $c^*$. Based on the router's decision $c^*$, the final debiased prediction $\tilde{p}$ for MRID is computed conditionally:

$$\tilde{p} = \begin{cases} p_0, & c^* = 0 \\ p_0 - \alpha_1 \cdot p_i, & c^* = 1 \\ p_0 - \beta_2 \cdot p_t, & c^* = 2 \\ p_0 - \alpha_3 \cdot p_i - \beta_3 \cdot p_t. & c^* = 3 \end{cases} \quad (23)$$

The hyperparameters $\alpha_c, \beta_c$ are searched on validation set as Sec. 4.1.

## E Analysis on Hyperparam for MID

To examine the usage of linear mode inference-time debaising, we analyzed the classification accuracy sensitivity to the coefficients $\alpha$ (text debias degree) and $\beta$ (image debias degree) in MID. The contour plots presented in Fig. 7. For both models, the resulting accuracy is clearly dependent on the specific choices of $\alpha, \beta$. It is evident that optimal performance is typically achieved when both $\alpha, \beta$ are non-zero. For the Qwen2VL-7B, peak

accuracy is observed around (0.4, 0.2), while the highest accuracy for InternVL2.5-4B is achieved in $\alpha \in [0.1, 0.4], \beta \in [0.3, 0.5]$. Setting these coefficients to extreme values (1.0) can lead to suboptimal performance, potentially due to over-correction which might suppress true signal along with bias. The slightly different optimal regions and peak accuracies between the two base models also suggest that these hyperparameters may benefit from model-specific tuning for best results.

## F Case Study

Below we present two samples selected from training set to illustrate the usage of multimodal debias.

**Enhancing judgement by removing multimodal bias** In case 1 (Fig. 8(a–b)), the caption "nothing says equality like discrimination" expresses sarcasm through a polarity-incongruent pairing and a fixed "nothing-says-X-like-Y" template. On the text side we mask the trigger content words *equality* and *discrimination*; on the image side we occlude the headline strip and adjacent high-saliency area, leaving a neutral studio background. The model's output changes from $[0.21, 0.79]$ (non-sarcastic, sarcastic) to $[0.01, 0.99]$. This shift indicates that a non-trivial portion of the initial non-sarcastic probability was supported by background cues rather than the ironic semantics, and that discounting background-only evidence yields a more calibrated prediction without altering the correct class.

**Correcting judgement by mitigating textual bias** In case 2 (Fig. 8(c–d)), the ground-truth label is non-sarcastic, yet the original prediction is borderline ($[0.47, 0.53]$). Here the dominant source of spurious evidence lies in the text: quoted evaluatives such as "extra," "too much," and "enough" are frequent correlates of sarcasm in web corpora and thus behave as prior-driven indicators. We mask these tokens (and lightly occlude the highest-saliency facial region in the image), which removes much of the sarcasm prior while preserving the author's non-ironic intent. The distribution moves to $[0.72, 0.28]$, aligning with the label and illustrating that attenuating text-side priors is effective when surface lexical markers, rather than semantics, are responsible for the erroneous bias.

## G Prompt Templates

**Image Analysis Prompt**

You are provided with an image from a tweet with the associated text: "%s". Analyze the image and categorize its visual elements based on their semantic relevance:

- **Main Content Elements:** Identify visual elements in the image that provide meaningful semantic clues, such as emotionally charged objects, facial expressions, or thematic components. These elements should align with or contradict the textual information.
- **Context Elements:** Identify generic, non-essential visual details (e.g., background patterns, irrelevant objects) that do not contribute significant semantic value to understanding the text-image relationship.

**Output Format:**

- **Main Content Elements:** [List of visual elements]
- **Context Elements:** [List of visual elements]

**Analysis Process:** Provide a brief explanation of how the main content elements were identified and their connection (or contradiction) with the text. Highlight how the context elements were separated based on their lack of semantic importance.

**Text Analysis Prompt**

You are provided with an image from a tweet with the associated text: "%s". Identify and categorize the words in the text based on their semantic relevance:

- **Main Content Words:** Extract words or phrases that provide meaningful semantic clues, such as emotional, thematic, or descriptive elements.
- **Context Words:** Extract words or phrases that are generic, stylistic, or non-essential (e.g., stop words, filler adjectives) and do not contribute significant semantic value.

**Output Format:**

- **Analysis Process:** Provide a brief explanation of how the main content words were identified and their relation with the image. Highlight how the context words were separated based on their lack of semantic importance.
- **Main Content Words:** [List of words/phrases]
- **Context Words:** [List of words/phrases]


(a) Origin Text: nothing says equality like discrimination


(b) Masked Text: nothing says [MASK] like [MASK]

(c) Origin Text: before u say jongin 's " extra " and " too much " , ask urself ... are ur favs even " enough " ?

(d) Masked Text: before u say jongin 's " [MASK] " and " [MASK] " , ask urself ... are ur favs even " [MASK] " ?

Figure 8: The original samples versus the masked samples (image and text inputs after masking major semantic content) with core semantic and contextual information removed.

| Case | Ground Truth | Before Debias | After Debias |
|---|---|---|---|
| Fig. 8 a,b | sarcastic | [0.21, **0.79**] | [0.01, **0.99**] |
| Fig. 8 c,d | non-sarcastic | [0.47, **0.53**] | [**0.72**, 0.28] |

Table 11: Probability distributions over {non-sarcastic, sarcastic} before and after debiasing.