Unveiling Modality Bias: Automated Sample-Specific Analysis for Multimodal Misinformation Benchmarks

Anonymous ACL submission

Abstract

Numerous multimodal misinformation benchmarks exhibit bias toward specific modalities, allowing detectors to make predictions based solely on one modality. While previous research has quantified bias at the dataset level or manually identified spurious correlations between modalities and labels, these approaches lack meaningful insights at the sample level and struggle to scale to the vast amount of online information. In this paper, we investigate the design for automated recognition of modality bias at the sample level. Specifically, we propose three bias quantification methods based on theories/views of different levels of granularity: 1) a coarse-grained evaluation of modality benefit; 2) a medium-grained quantification of information flow; and 3) a fine-grained causality analysis. To verify the effectiveness, we conduct a human evaluation on two popular benchmarks. Experimental results reveal three interesting findings that provide potential direction toward future research: 1) Ensembling multiple views is crucial for reliable automated analysis; 2) Automated analysis is prone to detector-induced fluctuations; and 3) Different views produce a higher agreement on modality-balanced samples but diverge on biased ones.

1 Introduction

The proliferation of online social media has accelerated the dissemination of misinformation (Li et al., 2024; Bu et al., 2024; Wang et al., 2024; Yue et al., 2024b; Wan et al., 2024), particularly in multimodal contexts where images and texts mutually reinforce each other, enhancing persuasiveness and deception to pepole (Tahmasebi et al., 2024; Guo et al., 2024; Chen and Shu, 2023; Comito et al., 2023). To verify the ability of Multimodal Misinformation Detection (MMD) models to exploit multimodal information, previous studies have proposed several Multimodal Misinformation Benchmarks (MMBs) such as Fakeddit (Nakamura et al., 2019) and MMFakeBench (Liu et al., 2024b).

However, these benchmarks exhibit bias toward specific modality (Papadopoulos et al., 2024), where one modality may dominate as the primary source of information, thereby diminishing the role of the other modality (Guo et al., 2023; Liang et al., 2024). Such modality bias can lead to serious problems: First, from the training aspect, models trained on biased benchmarks may lack robustness to the variation of that modality (Yang et al., 2024), making them vulnerable to uni-modal attacks. Second, from the evaluation aspect, biased benchmarks may yield incomprehensive measurement of MMD models, e.g., a model might perform well on a text-biased benchmark because it learns spurious text-label correlations instead of effectively integrating multimodal information (Goyal et al., 2017).

Unfortunately, no systematic investigation has been conducted on the modality bias of existing MMBs. Current methods for detecting modality bias on general multimodal benchmarks like visual question answering can be broadly divided into two categories: automated dataset-level quantification and manual identification by human experts. For the former one, Liang et al. (2024) utilize information theory to measure *redundancy*, *uniqueness*, and *synergy* across the entire dataset. However, as illustrated in Figure 1, bias can vary significantly across individual samples within a dataset, suggesting that this approach lacks the granularity needed to fully capture sample-specific biases. The latter one, as demonstrated by Liu et al. (2024a), involves detecting specific issues, such as spurious correlations between text modalities and labels. While manual identification can effectively detect biased samples, it is limited by scalability and is impractical for handling a large volume of online data. This naturally raises the question: **is it possible to automatically measure the modality bias at the sample level without human intervention?**

To this end, we conduct a systematic analysis of modality bias in MMBs and verify whether ma-

Figure 1: The automated analysis of samples from Fakeddit. For biased samples, we can directly infer from the preferred modality like the **Left** (an unreasonable fat cat image) and **Right** (the impossibility of resurrection) one.

chines can automatically provide a reasonable measurement. Modality bias can be classified into three types: Uni-image, Uni-text, and Modality-balance, which indicate image bias, text bias, and no bias. We leverage three quantification methods of different granularities and adapt them to bias identification, i.e., modality benefit, modality flow, and modality causal effect. At a coarse level, **modality benefit** identifies the input modality that contributes the most for final predictions using Shapley values (Wei et al., 2024; Shapley, 1953) from game theory, which fairly assesses individual contributions of different players in cooperative scenarios. At a medium level, **modality flow** utilizes saliency scores (Michel et al., 2019; Wang et al., 2023), which quantify attention interactions between different input modalities and output predictions to inspect the decision-making process and determine the prior modality. At the finest level, **modality causal effect** constructs the causal inference graph of MMD, which contains modality-balanced and biased paths, and traces the path that has the maximal causal effect based on counterfactual reasoning (Chen et al., 2023b, 2024b). We treat these methods as providing different views upon the decision of modality bias and adopt a voting mechanism to integrate these three views to obtain an ensembled multi-view output.

To validate the effectiveness of such automated sample-specific bias analysis, we conduct a human evaluation on 100 samples of Fakeddit (Nakamura et al., 2019) and MMFakeBench (Liu et al., 2024b) respectively. Experimental results reveal three interesting findings that offer potential direction and design consideration toward future automated sample-specific modality bias analysis: 1) Ensembling multiple views is crucial for a reliable automated analysis, which is not possible through single-view analysis, because the intricate nature of automated sample-specific modality bias detection is a complex task for machines. 2) Automated analysis is prone to detector-induced fluctuations. The performance of both single- and multi-view analysis is sensitive to the choice of misinformation detectors. This phenomenon is unavoidable since each view is dependent on the parameters of the chosen detector. Mitigating such sensitivity could enhance its practicality for real-world deployment. 3) Different views produce a higher agreement on modality-balanced samples but diverge on biased ones. Overall, we believe that automated sample-specific analysis has significant practical applications, e.g., cleaning a biased MMB by retaining modality-balanced samples with high consistency. Our contributions are as follows: **Firstly**, we are the first to design an automated sample-specific modality bias analysis for multimodal misinformation benchmarks. **Secondly**, we investigate the effectiveness of the proposed automated analysis via a human evaluation on two multimodal misinformation benchmarks. **Thirdly**, we uncover some interesting findings from empirical experiments, offering potential directions toward future research.

2 Related Work

2.1 Modality Bias

Modality bias is prevalent in various multimodal learning tasks (Papadopoulos et al., 2023; Chen et al., 2022). While there is no systematic analysis of modality bias in MMBs, prior research has uncovered bias patterns in general multimodal benchmarks like visual question answering (VQA). Two common approaches for analyzing modality bias include automated dataset-level quantification and manual identification by human experts. In the case of automated quantification, Liang et al. (2024) measure modality interaction using information theory and propose two PID estimators to evaluate entire datasets. However, bias can vary significantly across individual samples in MMBs, which limits the ability of dataset-level approaches to detect sample-specific biases. Regarding manual identification, Goyal et al. (2017) reveal a spurious correlation between text and labels in the VQA (Antol et al., 2015) dataset, where simply answering "yes" to questions beginning with "Do you see a ..." achieves 87% accuracy without considering the rest of the question or the image. Similarly, Liu et al. (2024a) highlight that over 90% of the answers to questions about whether the audio in the MUSIC-AVQA (Li et al., 2022) dataset matches the instrument shown in the video are "yes". Papadopoulos et al. (2024) simply hypothesize that modality bias in multimodal misinformation benchmarks stems from "asymmetric pairs" and they do not make a systematical analysis on the automated bias quantization. Although manual methods can effectively detect and mitigate bias through techniques like data augmentation or filtering rules, they are impractical for analyzing the vast amount of online multimodal misinformation.

Since bias can vary significantly across individual samples, this paper investigates the feasibility of automated sample-specific modality bias analysis and makes some interesting observations, providing potential direction and design consideration.

2.2 Multimodal Misinformation Benchmarks

Current multimodal misinformation benchmarks can be broadly categorized into two types: real-world and synthetic datasets. Fakeddit (Nakamura et al., 2019), the largest multimodal misinformation dataset, contains over 400k samples sourced from the social networking platform Reddit. Among synthetic datasets, NewsCLIPings (Luo et al., 2021) is constructed using techniques such as scene learning, person matching, and CLIP (Radford et al., 2021) to produce out-of-context samples. MMFakeBench (Liu et al., 2024b) leverages powerful vision-language models like DALL-E3 (Ramesh et al., 2022) to generate AI-based misinformation related to textual veracity, visual veracity, and cross-modal consistency distortion. However, as discussed in the introduction, there exists significant modality bias in these benchmarks, which presents clear drawbacks for both training and evaluating MMD models in real-world deployment.

In this paper, we perform the automated analysis on two multimodal misinformation benchmarks: a real-world dataset Fakeddit, and a synthetic dataset MMFakeBench. By analyzing benchmarks of different scenarios, we seek to comprehensively validate the effectiveness of our automated analysis.

3 Automated Sample-Specific Analysis

3.1 Overview

The overall workflow of automated analysis is illustrated in Figure 2. Several misinformation detectors are used to power the computation of automated analysis, i.e., the Image-only model, Image-text model, Text-only model, and large vision-language model. We need to fine-tune these models for more reliable measurements because existing models lack robust zero-shot capabilities for MMD. For a multimodal misinformation benchmark, we randomly select some samples (Subset1) to fine-tune the models and perform single- and multi-view analysis on the remaining subset (Subset2).

3.2 Modality Benefit

From the view of modality benefit, we introduce a Shapely value-based metric (Wei et al., 2024; Shapley, 1953), which is designed for cooperative games with n players, to observe the uni-modal contribution by comparing the model's prediction with/without specific modality. For generalization, we first illustrate the scenario with n modality and then provide the formula when $n = 2$.

Each sample $x = (x^{m_1}, x^{m_2}, ..., x^{m_n})$ is with n modality, y is the corresponding label, x^{m_i} is the modality m_i of sample x. Let $M = \{m_1, m_2, ..., m_n\}$ be the set of all modalities, M' be the subset of M ($M' \subseteq M$) and $x^{M'}$ be the input sample x with modality set M', we can define a benefit function V that maps the model's prediction with input M' to its benefits: if $\hat{y} = y$, $V(x^{M'}) = |M'|$; otherwise, $V(x^{M'}) = 0$. Here $\hat{y}$ is prediction and $||$ denotes the number of input M', i.e., if the model makes a correct prediction, the benefit will be the number of input modalities.

Since a player can interact with other players, different permutations of input modalities may yield varying outcomes. If we define a certain permutation as a strategy and let $\prod_M$ be the permutation of M, there is $|\prod_M| = n!$ strategies. For a strategy $\pi \in \prod_M$, the marginal benefit of modality m_i of sample x in π can be defined as: $v(\pi; x^{m_i}) = V(\pi(x^{m_i}) \cup x^{m_i}) - V(\pi(x^{m_i}))$, where $\pi(x^{m_i})$ represents all predecessors of x^{m_i} in π. This formula quantifies the increased benefit of modality x^{m_i} compared to its predecessors. Considering the marginal contribution of modality m_i of sample x in all strategies, the final benefit of modality m_i is given by: $\phi_{m_i} = \frac{1}{n!} \sum_{\pi \in \prod_M} v(\pi; x^{m_i})$.

As shown in Figure 2(b), when it comes to the

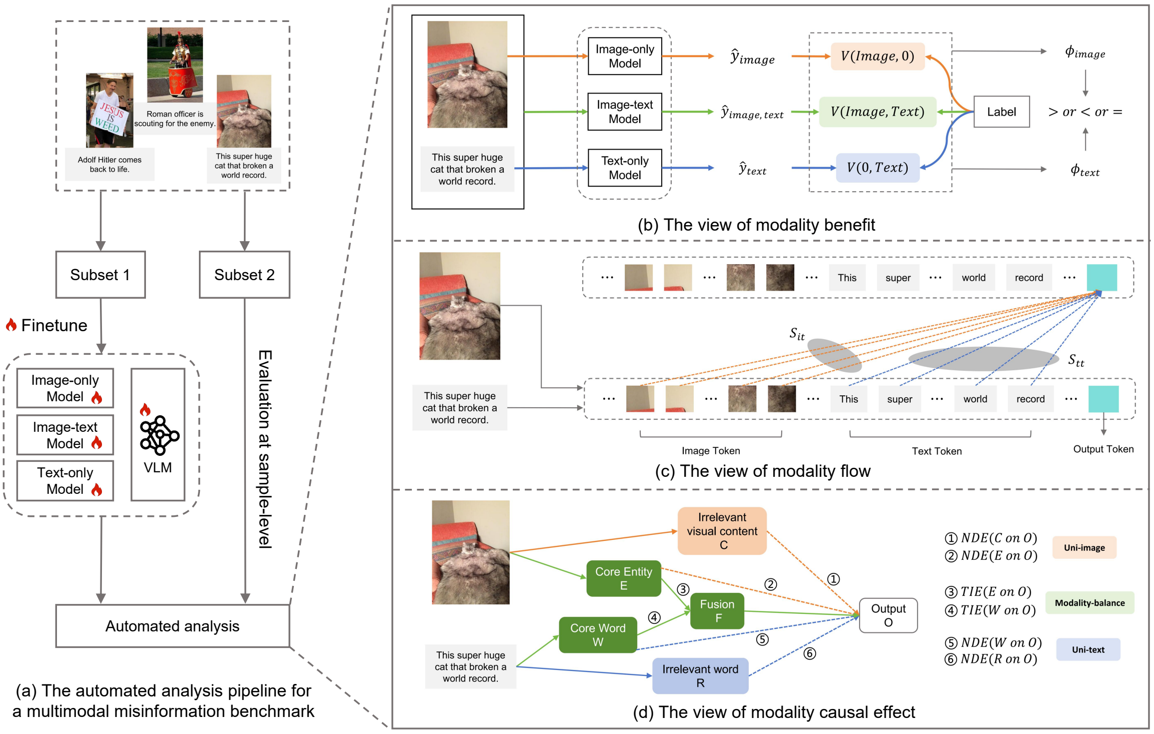

Figure 2: Illustration of proposed automated analysis for modality bias in multimodal misinformation benchmarks.

multimodal misinformation samples with image and text ($n = 2$), there are simply two strategies in $\prod_M = \{\pi_1 = (m_1, m_2), \pi_2 = (m_2, m_1)\}$. The final contribution of such a specific modality m_1 is given by: $\phi_{m_1} = \frac{1}{2}\left[v\left(\pi_1; x^{m_1}\right) + v\left(\pi_2; x^{m_1}\right)\right] = \frac{1}{2}\left[V\left(x^{m_1}, 0^{m_2}\right) - V\left(0^{m_1}, 0^{m_2}\right) + V\left(x^{m_2}, x^{m_1}\right) - V\left(x^{m_2}, 0^{m_1}\right)\right]$, where the above 0^{m_i} denotes the absence of modality m_i. We adopt zero input for image modality and placeholder padding for text modality following Wei et al. (2024). We set $V\left(0^{image}, 0^{text}\right)$ to zero and leverage Image-only, Image-text, and Text-only models to compute $V\left(x^{image}, 0^{text}\right)$, $V\left(x^{text}, x^{image}\right)$, and $V\left(x^{text}, 0^{image}\right)$, respectively. Finally, we can determine the bias type of each sample, i.e., Uni-image: $\phi_{image} > \phi_{text}$, Modality-balance: $\phi_{image} = \phi_{text}$, Uni-text: $\phi_{image} < \phi_{text}$.

3.3 Modality Flow

Figure 2(c) depicts the view of modality flow: comparing the information flow from the image/text to the output token intuitively reveals whether the model relies more on image or text modality when making predictions. Computing accurate attention interactions requires advanced models to provide reliable attention signals, so we leverage a large vision-language model (LVLM) rather than smaller models. Suppose the input prompt for MMD is $P = [..., IT, ..., TT, ..., OT]$, where $IT = (IT_1, IT_2, ..., IT_{n_1})$ is the image token, $TT = (TT_1, TT_2, ..., TT_{n_2})$ is the text token and OT is the output token which is usually the last token. Following Wang et al. (2023), we employ the saliency score to quantify critical token interactions: $S = \left|\sum_h A_h \odot \frac{\partial \mathcal{L}(P)}{\partial A_h}\right|$, where A_h represents the attention matrix of h-th attention head, $\odot$ is Hadamard product, P is the input prompt, $\mathcal{L}(\cdot)$ is the loss function of multimodal misinformation detection. Concretely, $S(j_1, j_2)$ denotes the importance of the information flow from j_2-th token to j_1-th token. Based on the observation that shallow layers are primarily used for token information aggregation and analysis, and deep layers leverage token information for prediction, we only calculate the saliency score for the last attention layer. To study the effect of different saliency calculations, we compare our attention-based saliency score calculation with another perturbation-based method LIME (Ribeiro et al., 2016) in Appendix C.

Generally, the number of image tokens exceeds that of text tokens. For instance, a 224×224 image can be divided into 64 patch tokens, while the corresponding text typically comprises fewer than ten tokens. Since most image tokens may represent background information, their individual contribution may be less significant compared to single text tokens. Therefore, to assess the overall contribution, we adopt the sum of the saliency score as the final significance of information flow from the respective modality to prediction: $S_{it} = \sum_k^{n_1} S(OT, IT_k),\ IT_k \in IT$ and $S_{tt} = \sum_k^{n_2} S(OT, TT_k),\ TT_k \in TT$. we study the effects of different computation strategies of S_{it} and S_{tt} in Appendix D.

Following Jin et al. (2021), we apply a normalization to S_{it} and S_{tt} to map them to the same interval: $S_{it,norm} = \frac{S_{it}}{S_{it}+S_{tt}}, S_{tt,norm} = \frac{S_{tt}}{S_{it}+S_{tt}}$.

In contrast to the discrete space of the Shapely value, the value space of saliency scores is continuous, which means $S_{it,norm} \neq S_{tt,norm}$ even when the sample is modality balanced. Therefore, we define a hyperparameter threshold ϵ to confine the differences of modality-balanced cases. In other words, when $|S_{it,norm} - S_{tt,norm}| < \epsilon$, we consider the sample to be modality-balanced. We conduct a user study to determine the threshold ϵ and a detailed description can be found in Appendix E.

3.4 Modality Causal Effect

The causal mechanisms of MMD problem-solving involve first analyzing the core information, such as primary entities in images and main semantics in text, and then combining them to derive the final prediction. However, biased data can yield predictions directly from a single modality.

In Figure 2(d), we illustrate all possible causal reasoning paths for MMD, where different paths correspond to different types of modality bias. Suppose I is the image, C is the irrelevant visual content of the image, E is the core entity of the image, T is the text, W is the core chunk of the text, R is the irrelevant fragment of the text, F is the information fusion of E and W, and O is the output, we make the following definitions. **Image Bias:** the model may directly predict through $I \rightarrow C \rightarrow O$ and $I \rightarrow E \rightarrow O$. **Text Bias:** the inference paths referred to as text bias include $T \rightarrow R \rightarrow O$ and $T \rightarrow W \rightarrow O$. **Modality Balance:** the desired causal path is via $I \rightarrow E \rightarrow F, T \rightarrow W \rightarrow F$ and $F \rightarrow O$. For core information extraction (C, E, W and R), we utilize MiniCPM-V 2.6 and Llama3-8B (AI@Meta, 2024) to process image and text, respectively. Appendix F provides details of core information extraction. Then we employ counterfactual reasoning to quantify the causal effects of different paths and identify bias types corresponding to the path exhibiting the greatest causal effect.

Counterfactual reasoning can estimate the causal effect of a treatment variable on a response variable by comparing outcomes under conditions that are different from the factual world. We denote the causal mechanism of MMD as: $O_{c,e,w,r,f} = O\left(C=c, E=e, W=w, R=r, F=f\right),\ f = F_{e,w} = F\left(E=e, W=w\right)$.

Consider the variable W as an example. There exist two paths between W and O, namely $W \rightarrow F \rightarrow O$ and $W \rightarrow O$ in the causal inference graph. Following Chen et al. (2023b), we define the total effect (TE) of $W = w$ on O as: $TE(\text{W on O}) = O_{w,f} - O_{w^*,f^*}$, where $*$ denotes the reference value. Total Effect can be interpreted as the comparison between two potential outcomes of W under two distinct treatments w and w^*. Meanwhile, Total Effect can be divided into Natural Direct Effect (NDE) and Total Indirect Effect (TIE). NDE is the causal effect of path $W \rightarrow O$ which means information from W to F has been blocked, while TIE denotes the causal effect of path $W \rightarrow F \rightarrow O$.

In the counterfactual scenario, W is supposed to be the values w and w^* simultaneously, where w^* influences the indirect path $W \rightarrow F \rightarrow O$, while w influences the direct path $W \rightarrow O$. In other words, w^* isolates the influence of W on the intermediate factor F, thereby enabling us to directly observe the effect of W on O. Therefore, $NDE(\text{W on O}) = O_{w,f^*} - O_{w^*,f^*}$ and we have $TIE(\text{W on O}) = TE - NDE = O_{w,f} - O_{w,f^*}$.

Following previous studies (Chen et al., 2023b; Wang et al., 2021), we also set other variables C, E, and R to their reference value c^*, e^*, and r^* when $W = w^*$. For such reference value, we adopt zero input for c^* and e^*, and placeholder padding for w^* and r^*. To obtain the ensemble prediction, we apply a non-linear fusion strategy. For example,$O_{c,e,w,r,f} = \mathcal{F}\left(O_c, O_e, O_w, O_r, O_f\right) = tanh(O_c) + tanh(O_e) + tanh(O_w) + tanh(O_r) + O_f$, where $\mathcal{F}(\cdot)$ is the non-linear fusion strategy, O_c is the output of the irrelevant visual context branch, O_e is the outcome of the core entity branch, O_w is the result of the core semantic words branch, O_r is the output of the irrelevant word branch, O_f is the output of fusion branch. To compute these outputs, we utilize the Image-only model for O_c and O_e, the Text-only model for O_w and O_r, and the Image-text model for O_f. While $\mathcal{F}(\cdot)$ can be any differentiable binary function, Chen et al. (2023b) observe that tanh-sum yields the best performance.

Similarly, we can compute the natural direct effect of variable C, E, and R on O and the total indirect effect of variable E on O, i.e., NDE(C on O), NDE(E on O), NDE(R on O), and TIE(E on O). As shown in Figure 2(d), these causal effect items correspond to the six distinct paths within the inference graph, with each path associated with a specific modality bias type. For each sample, we determine the bias type based on the path exhibiting the greatest causal effect.

Finally, multi-view analysis is derived through a prior majority voting, where the outcome is determined by the majority of three views. In the event of a tie, priority is assigned to the category with the larger number of samples in the human annotation. Discussion of more ensemble strategies is shown in Appendix B.

4 Experiment Setting

4.1 Benchmarks

We conduct the automated sample-specific modality bias analysis on two multimodal misinformation benchmarks, i.e., Fakeddit and MMFakeBench. Fakeddit is a highly diverse real-world benchmark and contains over six hundred thousand multimodal samples. Moreover, MMFakeBench is a synthetic dataset generated by large vision-language models like DALL-E3. These two benchmarks are particularly representative due to their large scale (680K samples) and extensive coverage of diverse domains, including real-world misinformation, AI-generated synthetic content, satire, rumors, face swaps, and Photoshop-edited images. A detailed description of these datasets, along with their statistical distributions, is provided in AppendixG.

4.2 Models

We define the required types of misinformation detection models for our multi-view analysis as {Image-only, Image-text, Text-only, LVLM}. For computational efficiency, we use the first three types of models to support the analysis of modality benefit and modality causal effect (Niu et al., 2021). As for modality flow, computing accurate attention interactions requires advanced models to provide reliable attention signals, so we leverage a large vision-language model (LVLM) rather than smaller models. We select the following models for experimentation, i.e., **Image-only**: UnivFD (Ojha et al., 2023) and DT(I); **Image-text**: HAMMER (Shao et al., 2023) and DT(I, T) (Papadopoulos et al., 2024); **Text-only**: FFNews (Huang et al., 2022) and DT(T); **LVLM**: MiniCPM-V 2.6 (Yao et al., 2024). Since existing models demonstrate limited zero-shot detection performance, we first fine-tune these models to improve their reliability. Appendix H describes details of selected models, the selection criteria, and the fine-tuning process.

4.3 Implement Details

We conduct automated analysis on 100 samples from each benchmark with the following model group: {UnivFD, HAMMER, FFNews, MiniCPM-V 2.6}. All experiments are conducted on one A100 80GB GPU. The approximate inference time of modality benefit, flow, and causal effect: 1 hour, 3 hours, and 2 hours every 60k samples respectively. More experiment details can be found in Appendix E, F.

4.4 Evaluation

We are the first to propose an automated sample-specific modality bias analysis and no existing baselines are available for direct comparison. Therefore, we conduct a human evaluation with three annotators to validate the alignment of single- and multi-view analysis and human judgment. To assess the reliability and agreement of human annotations, we conducted Krippendorff's alpha test (Krippendorff, 2011). Details of annotators' demographic characteristics, annotation procedure, and the result of Krippendorff's alpha test can be found in Appendix I. We report the predicted proportions of each modality bias type and the percentage that aligns with human judgment. For example, $0.84[85.71]$ denotes that multi-view analysis classifies 0.84 of the samples as modality-balance, and among these samples, 85.71% of the results are consistent with human judgment.

5 Experimental Results

This section contains three interesting findings (5.1, 5.2, 5.3) about our proposed automated sample-specific modality bias analysis. More ablation experiments (i.e., the effect of ensemble strategies, saliency score calculations, and computation strategies of S_{it} and S_{tt} in modality flow) and the error analysis can be found in Appendix B, C, D, J.

5.1 Key to Reliable Automated Analysis

Table 1 depicts the quantification comparison of automated analysis and human judgment.

Comparison of Proportion. According to human judgment, most samples are modality-balanced, while only a small proportion are bi-

	Fakeddit				MMFakeBench			
	Uni-image	Modality-balance	Uni-text	Acc	Uni-image	Modality-balance	Uni-text	Acc
Human	0.18	0.78	0.04	-	0.13	0.74	0.13	-
Modality benefit	0.02[0.00]	0.90[78.89]	0.08[37.50]	74.00	0.47[10.64]	0.41[80.49]	0.12[66.67]	46.00
Modality flow	0.15[40.00]	0.52[88.46]	0.33[12.12]	56.00	-	0.67[71.64]	0.33[15.15]	53.00
Modality causal effect	0.40[32.50]	0.56[92.86]	0.04[0.00]	65.00	0.10[40.00]	0.63[82.54]	0.27[40.74]	67.00
Multi-view analysis	0.08[75.00]	0.84[85.71]	0.08[37.50]	**81.00**	0.07[57.14]	0.79[86.08]	0.14[78.57]	**83.00**
Benefit-Flow	0.02[0.00]	0.91[79.12]	0.07[42.86]	75.00	0.16[0.00]	0.82[74.39]	0.02[0.00]	61.00
Benefit-Causal	0.05[0.00]	0.92[79.35]	0.03[0.00]	73.00	0.20[20.00]	0.69[84.06]	0.11[72.73]	70.00
Flow-Causal	0.22[31.82]	0.74[86.49]	0.04[0.00]	71.00	-	0.95[75.79]	0.05[60.00]	75.00

Table 1: The quantification comparison of automated analysis and human judgment. We report the predicted proportion (without []) and accuracy (within []) of different bias types compared to human annotations. Acc denotes the overall accuracy. The proportion ranges from 0 to 1 and the accuracy is presented as percentages (%).

ased. Although single-view analysis generally follows this pattern, notable differences exist in specific numerical values. For example, on Fakeddit, modality benefit classifies 0.02 of the samples as "Uni-image", modality flow classifies 0.33 of the samples as "Uni-text", and modality causal effect classifies 0.40 of the samples as "Uni-image". A similar trend is observed on MMFakeBench. However, multi-view analysis integrates the strengths of each individual view, yielding results that most closely align with human judgment.

Comparison of Accuracy. Different views reveal distinct patterns of bias, and single-view analysis may underperform in certain scenarios. For example, the Modality Benefit analysis shows strong performance (74.00%) on Fakeddit while weak performance (46.00%) on MMFakeBench. However, the ensemble multi-view analysis consistently achieves the highest performance across both datasets, underscoring the stability of multi-view approaches in the complex task of automatically detecting modality bias across diverse scenarios, including both real-world and synthetic samples.

Ablation Study. We also conduct an ablation study on three variants to assess the contribution of each view: (1) Benefit-Flow: Omitting the modality causal effect. (2) Benefit-Causal: Removing the modality flow. (3) Flow-Causal: Excluding the modality benefit. As shown at the bottom of Table 1, each view contributes meaningfully to the multi-view analysis.

Multi-view analysis significantly outperforms the three single-view methods in both performance and stability. Therefore, we conclude that automated sample-specific modality bias analysis is a complex task for machines. While reliable measurements cannot be attained solely through single-view analysis, ensemble multi-view demonstrates promising potential for real-world deployment.

Fakeddit	Group1	Group2	Group3	Group4
Modality benefit	74.00	68.00	74.00	53.00
Modality causal effect	65.00	68.00	62.00	66.00
Multi-view analysis	81.00	72.00	78.00	72.00
MMFakeBench	Group1	Group2	Group3	Group4
Modality benefit	46.00	42.00	46.00	64.00
Modality causal effect	67.00	68.00	49.00	69.00
Multi-view analysis	83.00	73.00	64.00	70.00

Table 2: The accuracy [%] of modality benefit, modality causal effect, and multi-view analysis under different types of misinformation detector.

5.2 Vulnerability to Detector Fluctuations

In the computational process of automated analysis, various misinformation detectors are involved, such as the image-only, image-text, and text-only models utilized in modality benefit and modality causal effect, as well as the LVLM employed in modality flow. A pertinent question arises: **is the automated analysis robust to the different choices of misinformation detectors?**

To answer this question, we evaluate the sensitivity of *modality benefit*, *modality causal effect*, and *multi-view analysis* by altering specific models and observing the change in accuracy based on the same samples selected in Section 5.1[1]. We select four model combinations (across Image-only, Image-text, and Text-only models):

- Group1={UnivFD, HAMMER, FFNews}
- Group2={**DT(I)**, HAMMER, FFNews}
- Group3={UnivFD, **DT(I, T)**, FFNews}

[1] Due to the high computation cost and the strong stability of LVLM compared to small models, we do not study the sensitivity of modality flow.

- Group4={UnivFD, HAMMER, **DT(T)**}

As illustrated in Table 2, when considering the average performance on Fakeddit and MMFakeBench, the maximum fluctuation exceeds 10% for both single-view and multi-view scenarios, indicating that automated analysis is prone to detector-induced fluctuations. We take this phenomenon as unavoidable because each view quantifies modality bias based on models' output, and the performance of different models can vary significantly. Transferring the model for a specific modality inevitably affects the distribution of prediction for that modality, which in turn influences the calculation of modality contribution in each view.

Therefore, in practical applications, certain improvements are necessary to enhance the robustness of automated analysis. On the one hand, the simplest approach is to ensemble various misinformation detectors for each view, thus leveraging the strengths of different types of detectors. However, this method introduces additional computational overhead and is more suitable for scenarios where real-time consideration is low-priority, such as preliminary cleaning of modality-biased benchmarks. On the other hand, model-agnostic features can be incorporated to compute detectors' output, such as edge or texture features for images and TF-IDF features for text. While this reduces reliance on specific model architectures, it requires the design of effective model-agnostic feature extraction methods to ensure that these features can capture the key information related to modality bias.

5.3 Modality-balanced vs. Biased Samples

Table 1 reveals that multi-view analysis achieves high accuracy on modality-balanced samples but exhibits lower accuracy on biased ones. For example, on Fakeddit, the accuracy of multi-view analysis on "Modality-balance" samples is 85.71%, whereas on "Uni-text" samples, the accuracy drops to 37.50%. A similar trend is observed on MMFakeBench, where the accuracy on "Modality-balance" samples is 86.08%, but on "Uni-image" samples, it decreases to 57.14%. **What contribute to this performance discrepancy?**

To answer this question, we use Venn diagrams to visualize the intersections among different views to analyze the consistency of multi-view analysis. It is important to note that this analysis encompasses the entire dataset, rather than those samples from human evaluations. As illustrated in Figure 3, different views exhibit high alignment on modality-balanced samples but significant divergence on biased samples. We attribute this divergence to the fact that different views possess distinct patterns for capturing bias. Generally, higher consistency among views yields higher accuracy, and thus, this divergence leads to suboptimal accuracy on biased samples. In real-world deployment, if our objective is to clean a modality-biased benchmark by retaining only modality-balanced samples, the results of the automated analysis can serve as a robust reference. Conversely, if the focus is on biased samples, it becomes necessary to employ related techniques to mitigate this divergence, thereby ensuring the reliability of the results. For instance, a calibrator could be designed to post-process the predicted probabilities of biased samples of each view.

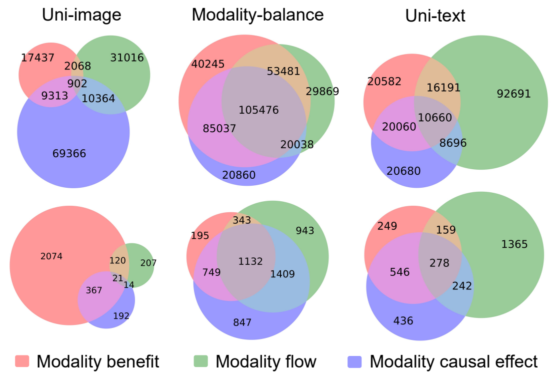

Figure 3: The Venn diagram of three single-views on Fakeddit (top three) and MMFakeBench (bottom three).

6 Conclusion

In this work, we investigate whether it is possible to establish an automated sample-specific modality bias analysis for existing multimodal misinformation benchmarks. We first propose three quantification methods based on different theories and adapt them to bias identification, i.e., the view of modality benefit, modality flow, and modality causal effect. Then we conduct a human evaluation on two multimodal misinformation benchmarks to study the practicability of automated analysis and derive three interesting findings that offer design consideration and improvement direction toward future research. Experimental results indicate that automated sample-specific modality bias analysis holds promising potential for practical applications. This suggests its capability to perform tasks like dataset cleaning (i.e., retaining modality-balanced samples) to mitigate the severity of modality bias.

7 Limitations

There are two limitations in this work. Firstly, due to the substantial workload associated with human evaluation, it is challenging to scale the number of test samples. We randomly selected 100 samples for human evaluation to validate the effectiveness of our proposed multi-view analysis. However, a larger sample size could enhance statistical significance and provide a more robust evaluation. Secondly, we do not study the effect of different large vision-language models (e.g., larger and stronger LVLMs) on modality flow view because of LVLMs' high computation cost of saliency score calculation based on the loss backward process.

References

AI@Meta. 2024. Llama 3 model card.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1351–1360.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. 2020. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 103–120. Springer.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024b. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *arXiv preprint arXiv:2403.18346*.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905.

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. 2023a. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023b. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

Carmela Comito, Luciano Caroprese, and Ester Zumpano. 2023. Multimodal fake news detection on social media: a survey of deep learning techniques. *Social Network Analysis and Mining*, 13(1):101.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *Preprint*, arXiv:2407.11691.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Hao Guo, Zihan Ma, Zhi Zeng, Minnan Luo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2024. Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. *arXiv preprint arXiv:2412.14686*.

Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo. 2023. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–22.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2022. Faking fake news for real fake news detection: Propaganda-loaded training data generation. *arXiv preprint arXiv:2203.05386*.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2021. One map does not fit all: Evaluating saliency map explanation on multi-modal medical images. *arXiv preprint arXiv:2107.05047*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Yupeng Li, Haorui He, Jin Bai, and Dacheng Wen. 2024. Mcfend: a multi-source benchmark dataset for chinese fake news detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4018–4027.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. 2024. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36.

Xiulong Liu, Zhikang Dong, and Peng Zhang. 2024a. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4478–4487.

Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. 2024b. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. *arXiv preprint arXiv:2406.08772*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, Amit K Roy-Chowdhury, and BS Manjunath. 2019. Detecting gan generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*.

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*.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. 2023. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, pages 36–44.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Lloyd S Shapley. 1953. A value for n-person games. *Contribution to the Theory of Games*, 2.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. Multimodal misinformation detection using large vision-language models. *arXiv preprint arXiv:2407.14321*.

Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 134–139.

Herun Wan, Minnan Luo, Zhixiong Su, Guang Dai, and Xiang Zhao. 2024. On the risk of evidence pollution for malicious social text detection in the era of llms. *arXiv preprint arXiv:2410.12600*.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM on Web Conference 2024*, pages 2452–2463.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1288–1297.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. 2024. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27338–27347.

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. 2024. Quantifying and enhancing multi-modal robustness with modality preference. *arXiv preprint arXiv:2402.06244*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024b. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5628–5643.

A Description of Appendix

This appendix contains the investigation of different settings (B, C, D), the detailed information about corresponding processes (E, F, G, H, I), the error analysis of multi-view output (J), and discussion of some considerations (K), which contributes to a comprehensive understanding and evaluation of this paper. Appendix B examines how various methods of combining multi-view can influence performance. Appendix C delves into the effect of different saliency score calculation methods. Appendix D study the effect of different computation strategies of S_{it} and S_{tt} in the view of modality flow. Appendix E describes the determination and impact of super-hyperparameter ϵ. Appendix F focuses on the core information extraction prompts and the effect of different extraction model combinations. Appendix G provides a quantitative overview of multimodal misinformation benchmarks utilized in our work. Appendix H detailedly clarifies the model description, model selection criteria, and fine-tuning details. Appendix I presents the details of human annotation and instruction. Appendix J conducts an error analysis of the ensemble multi-view analysis. Appendix K discusses several considerations of this work, like the versatility of our proposed automated analysis.

B Effect of Ensemble Strategies

We explore the impact of different ensemble strategies in Table 3, including random majority voting, prior majority voting (ours), and weighted voting. The weights assigned to each view are [0.3, 0.2, 0.5], which are determined based on the average performance of single-view analysis. For instance, modality causal effect ranks second on Fakeddit and first on MMFakeBench, demonstrating overall superior performance among three single-view analyses. Therefore, we assign a weight of 0.5 to this view. Different voting strategies exhibit varying performance across different benchmarks. Overall, prior majority voting demonstrates the most stability and optimal performance.

C Effect of Saliency Score Calculations

Table 4 presents the results of our saliency score calculations and LIME for comparative analysis, specifically focusing on multi-view analysis and inference speed. FPS (Frame Per Second) denotes the number of samples that can be processed per second (i.e., a higher value indicates faster). The choice of saliency score calculation method has relatively little impact on the inference speed compared to the performance of multi-view analysis.

D Effect of Computation Strategies

As for the computation strategies of S_{it} and S_{tt}, we report the predicted proportion under sum, average and maximum conditions in Table 6. We observe that the results of average and maximum strategies are highly unreasonable, which exhibits a strong bias toward text modality. We refer to this phenomenon as the modality gap. For instance, the image modality typically contains more tokens than the text modality, but many of these tokens often carry background information with minimal impact on the output. When using the average strategy, the contribution of the text modality is exaggerated. A similar problem arises with the maximum strategy, likely due to inherent differences in how the LVLM assigns attention to individual tokens of different modalities. This could be attributed to the fact that LVLMs consist of a superior language model (>7B) paired with a simple small image encoder (500M).

E Determination of Threshold

We conduct a user study to determine the threshold in the view of modality flow, selecting 20 samples from Fakeddit and MMFakeBench and manually annotating the types of modality bias. It is important to note that these samples are used for tuning the threshold and are different from those used for human evaluation. In this user study, the first author of this paper serves as the data annotator and adopts the same criteria described in Appendix I. By adjusting the threshold from 0 to 0.4 in increments of 0.05, we identify the threshold that achieves the highest accuracy for the modality flow analysis. As shown in Figure 4, we set the threshold as 0.25.

We also present the results of the ensemble multi-view analysis under different threshold ϵ in Table 5. The general trend observed is that, as the threshold increases, accuracy initially rises, then stabilizes, and eventually declines. It is consistent with the findings from the above user study (Figure 4).

	Fakeddit				MMFakeBench			
Ensemble Strategy	Uni-image	Modality-balance	Uni-text	Acc	Uni-image	Modality-balance	Uni-text	Acc
Random majority voting	0.13[46.15]	0.77[84.42]	0.10[30.00]	74.00	0.14[28.57]	0.65[83.08]	0.21[52.38]	69.00
Prior majority voting (Ours)	0.08[75.00]	0.84[85.71]	0.08[37.50]	**81.00**	0.07[57.14]	0.79[86.08]	0.14[78.57]	**83.00**
Weighted voting	0.19[36.84]	0.73[86.30]	0.08[37.50]	73.00	0.07[57.14]	0.69[84.06]	0.24[45.83]	73.00

Table 3: The effect of different ensemble strategies on the multi-view analysis.

	Fakeddit				MMFakeBench				Inference Speed
	Uni-image	Modality-balance	Uni-text	Acc	Uni-image	Modality-balance	Uni-text	Acc	FPS
Ours	0.08[75.00]	0.84[85.71]	0.08[37.50]	**81.00**	0.07[57.14]	0.79[86.08]	0.14[78.57]	**83.00**	**0.4942**
LIME	0.06[66.67]	0.91[80.22]	0.03[0.00]	77.00	0.07[57.14]	0.69[84.06]	0.24[45.83]	73.00	0.3489

Table 4: The effect of different saliency score calculations on the multi-view analysis.

	Fakeddit				MMFakeBench			
ϵ	Uni-image	Modality-balance	Uni-text	Acc	Uni-image	Modality-balance	Uni-text	Acc
0	0.10[60.00]	0.81[85.19]	0.09[33.33]	78.00	0.22[22.73]	0.58[82.76]	0.20[55.00]	64.00
0.05	0.10[60.00]	0.81[85.19]	0.09[33.33]	78.00	0.18[22.22]	0.62[82.26]	0.20[55.00]	66.00
0.10	0.08[75.00]	0.84[85.71]	0.08[37.50]	**81.00**	0.18[22.22]	0.62[82.26]	0.20[55.00]	66.00
0.15	0.08[75.00]	0.84[85.71]	0.08[37.50]	**81.00**	0.18[22.22]	0.67[83.58]	0.15[73.33]	71.00
0.20	0.08[75.00]	0.84[85.71]	0.08[37.50]	**81.00**	0.15[26.67]	0.71[84.51]	0.14[78.57]	75.00
0.25 (Ours)	0.08[75.00]	0.84[85.71]	0.08[37.50]	**81.00**	0.07[57.14]	0.79[86.08]	0.14[78.57]	**83.00**
0.30	0.08[75.00]	0.84[85.71]	0.08[37.50]	**81.00**	0.07[57.14]	0.79[86.08]	0.14[78.57]	**83.00**
0.35	0.06[66.67]	0.87[83.91]	0.07[42.86]	80.00	0.07[57.14]	0.79[86.08]	0.14[78.57]	**83.00**
0.40	0.06[66.67]	0.87[83.91]	0.07[42.86]	80.00	0.07[57.14]	0.82[82.93]	0.11[72.73]	80.00

Table 5: The effect of different threshold ϵ on the multi-view analysis.

	Uni-image	Modality-balance	Uni-text
Sum(Ours)	0.15	0.52	0.33
Avg	0.00	0.00	1.00
Max	0.08	0.00	0.92

Table 6: The predicted proportion [0-1] of modality flow under different aggregation strategies.

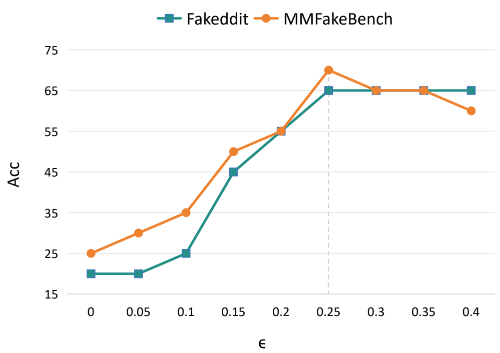

Figure 4: Accuracy of the view of modality flow with varying threshold ϵ on Fakeddit and MMFakeBench.

F Core Information Extraction

In the view of modality causal effect, we first leverage two large models to extract the core information and then construct the causal graph. Specifically, we utilize MiniCPM-V 2.6 to identify the core entity E and irrelevant visual content C of images. Llama3-8B is employed to recognize the core word W and irrelevant word R of texts. Noted that these large models used for core information extraction do not require further fine-tuning. The prompts are as follows:

- **MiniCPM-V 2.6**: $< Image >$ Please identify the core entity in this image. Output the corresponding entity region coordinates in the format of [x1, y1, x2, y2], where (x1, y1) denotes the top-left coordinate and (x2, y2) denotes the bottom-right coordinate. Remember to apply coordinate normalization, which means the coordinate range from 0 to 1.

- **Llama3-8B**: Please identify the keyword that can represent the core semantic information of this sentence: $< Text >$. Output the words in the format of [word1, word2, ..., wordn] if the core semantic is word1, word2, ..., and wordn. Please note that the number of words would not be fixed. It depends on your understanding of the sentence.

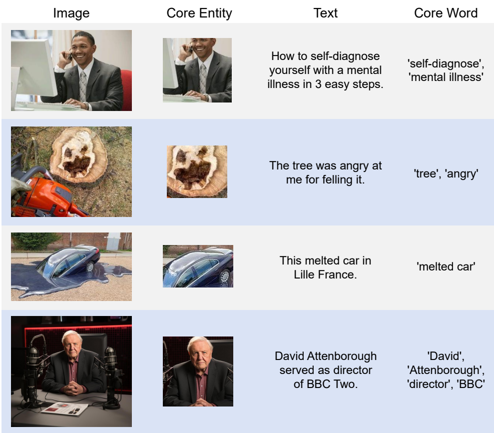

Figure 5: Examples of core information extraction.

Here we provide some examples (Figure 5) to validate the reliability of the extraction results.

To study the effect of different core information extraction models, we adopt additional large models, specifically another LVLM, Ovis1.6-Gemma2-9B (Lu et al., 2024), and another LLM, Yi-1.5-9B (Young et al., 2024). Table 7 depicts the ensemble multi-view analysis of different model combinations. Generally, the stronger a large model's reasoning ability, the more accurately it can extract core information. So the overall accuracy of multi-view analysis will be higher. This phenomenon further corroborates the universality and extensibility of the proposed automated analysis. As the capabilities of large models enhance, the accuracy of our proposed automated sample-specific modality bias analysis is anticipated to improve further.

G Statistics of Benchmarks

Table 8 depicts the statistics of two multimodal misinformation benchmarks, i.e., Fakeddit and MMFakeBench. Specifically, we report the number of each category (i.e., Real or Fake). Constructed from popular online media, Fakeddit is a highly diverse real-world English benchmark and contains over six hundred thousand multimodal samples. In contrast, MMFakeBench is a synthetic English dataset generated by Large Vision-language models (LVLM) like DALL-E3. For a multimodal misinformation benchmark with a predefined partition of "Train", "Valid", and "Test" sets, we first randomly select 40% of the samples from the "Train" set to fine-tune the models, and then perform sample-specific modality bias analysis on the remaining 60% of the "Train" set, the "Valid" set, and the "Test" set. To avoid confusion, we refer to the data used for fine-tuning as "Finetune_train" and "Finetune_valid" , while the remaining subsets used for automated analysis are referred to as "Analysis_train", "Analysis_valid", and "Analysis_test".

H Model Description, Selection Criteria, and Fine-tuning Details

Model Description. We first introduce models utilized in each view. UnivFD (Ojha et al., 2023) is a versatile fake image detector that operates within a feature space not explicitly trained to distinguish real from fake images. HAMMER (Shao et al., 2023), a multimodal detector built on ALBEF (Li et al., 2021), detects manipulation across different multimedia types. FFNews (Huang et al., 2022) specializes in detecting textual fake news, particularly human-generated misinformation. MiniCPM-V 2.6 (Yao et al., 2024) excels in multimodal understanding and outperforms some closed-source LVLMs like Gemini-1.5-Pro (Duan et al., 2024). DT($\cdot$) (Papadopoulos et al., 2024) utilizes CLIP ViT-L/14 (Radford et al., 2021) to extract modality features, with different variants (DT(I), DT(T), DT(I,T)) representing different modality inputs.

Model Selection Criteria. We select these misinformation detection models based on their strong performance and report the detailed quantitative comparison with some other models in Table 9. For **Image-only** models, we show the performance of Patch classifier (Chai et al., 2020), Co-occurence (Nataraj et al., 2019) and UnivFD on FaceForensics++ (Rossler et al., 2019) and LDM (Rombach et al., 2022). For **Image-text** models, we depict the performance of CLIP (Radford et al., 2021), ViLT (Kim et al., 2021) and HAMMER on DGM4 (Shao et al., 2023). For **Text-only** models, we compare the performance of DEFEND (Shu et al., 2019), DualEmo (Vaibhav et al., 2019) and FFNews on PolitiFact (Shu et al., 2020) and LUN (Rashkin et al., 2017). For LVLM, we compare three models of different serials (Ovis1.5-Gemma2-9B (Lu et al., 2024), InternVL2-8B-MPO (Chen et al., 2023a), and MiniCPM-V-2.6) and report the average score of eight evaluation datasets (i.e., MMBench (Liu et al., 2025), MMStar (Chen et al., 2024a), MMMU (Yue et al., 2024a), MathVista (Lu et al., 2023), AI2D (Kembhavi et al., 2016), HallusionBench (Guan et al., 2024), MMVet (Yu et al., 2023), OCRBench (Liu et al., 2024c)) based on VLMEvalKit (Duan et al., 2024). Note that our framework is adaptable to any

	Fakeddit				MMFakeBench			
Model Combination	Uni-image	Modality-balance	Uni-text	Acc	Uni-image	Modality-balance	Uni-text	Acc
MiniCPM-V 2.6, Llama3-8B (Ours)	0.08[75.00]	0.84[85.71]	0.08[37.50]	**81.00**	0.07[57.14]	0.79[86.08]	0.14[78.57]	**83.00**
MiniCPM-V 2.6, Yi-1.5-9B	0.11[54.55]	0.80[86.25]	0.09[33.33]	78.00	0.03[0.00]	0.86[79.07]	0.11[72.73]	76.00
Ovis1.6-Gemma2-9B, Llama3-8B	0.11[54.55]	0.81[85.19]	0.08[37.50]	78.00	0.12[33.33]	0.74[85.14]	0.14[78.57]	78.00
Ovis1.6-Gemma2-9B, Yi-1.5-9B	0.12[50.00]	0.80[85.00]	0.08[37.50]	77.00	0.06[0.00]	0.83[81.93]	0.11[72.73]	76.00

Table 7: The effect of different extraction models on the multi-view analysis.

		Fakeddit	MMFakeBench
Finetune_train	#Real	80465	1044
	#Fake	123281	2556
Finetune_valid	#Real	8796	125
	#Fake	13843	275
Analysis_train	#Real	132820	1831
	#Fake	204409	4169
Analysis_valid	#Real	23320	300
	#Fake	35979	700
Analysis_test	#Real	23507	-
	#Fake	35764	-
Total	#Real	268908	3300
	#Fake	413274	7700

Table 8: Statistics of the Fakeddit and MMFakeBench.

Image-only model	FaceForensics++	LDM
Patch classifier	75.54	79.09
Co-occurence	57.10	70.70
UnivFD	**84.50**	**94.19**
Image-text model	DGM4	
CLIP	76.40	
ViLT	78.38	
HAMMER	**86.39**	
Text-only model	PolitiFact	LUN
DEFEND	82.67	81.33
DualEmo	87.78	81.78
FFNews	**88.00**	**82.53**
LVLM	Param (B)	Avg Score
Ovis1.5-Gemma2-9B	11.4	64.00
InternVL2-8B-MPO	8	64.50
MiniCPM-V-2.6	8	**65.20**

Table 9: Quantitative comparison of misinformation detection models and LVLMs.

	Model	Fakeddit	MMFakeBench
Image-only	UnivFD	79.94	74.25
	DT(I)	88.01	80.75
Image-text	HAMMER	92.41	81.00
	DT(I, T)	93.40	83.75
Text-only	FFNews	89.20	86.04
	DT(T)	88.73	75.50
LVLM	MiniCPM-V 2.6	94.61	95.00

Table 10: The accuracy of tuned models on the "Finetune_valid" set of Fakeddit and MMFakeBench.

misinformation detection method and LVLM.

Fine-tuning Details. Due to the limited performance of existing models in multimodal misinformation detection under zero-shot scenarios, fine-tuning is required for a robust and accurate measurement. Specifically, we apply supervised fine-tuning (SFT) to UnivFD, HAMMER, FFNews, DT(I), DT(I, T), and DT(T) for 10 epochs. As for the MiniCPM-V 2.6, we apply LoRA-based parameter-efficient fine-tuning for 1 epoch considering the balance of resources and accuracy. All hyperparameters are consistent with their original work and experiments are conducted on one A100 80GB GPU. The accuracy of tuned models on the "Finetune_valid" set is shown in Table 10.

I Human Annotation

Liang et al. (2024) show that human judgment can be used as a reliable estimator of multimodal interaction. Following their design, we also conduct a human evaluation with three annotators to demonstrate the effectiveness of multi-view analysis. we recruited the annotators from the local universities of China through public advertisement with a specified pay rate. They are neither the authors nor members of the authors' research group and are all working towards a graduate degree in computer science and possess knowledge of multimodal learning. We pay them 50 CNY an hour. We show both modalities to annotators and ask them to annotate the type of modality bias for each sample. We randomly select 100 samples from each dataset to conduct the experiment. For Fakeddit, there are 60 samples from "Analysis_train", 20 samples from "Analysis_valid", and 20 samples from "Analysis_test". For MMFakeBench, there are 60 samples from "Analysis_train" and 40 samples from "Analysis_valid". We clarify the annotation proce-

	Uni-image	Modality-balance	Uni-text
Fakeddit	0.8251	0.8913	0.8122
MMFakeBench	0.8298	0.8940	0.8031

Table 11: The Krippendorff's alpha test of human annotations.

dure and judgment criteria before annotation.

- Instruction: Given a multimodal news sample, it contains both news caption and news image. You need to rate the following three questions ranging from 0-5.
- Question 1. (Uni-Image): The extent to which **Image** modality enables you to predict without the other modality.
- Question 2. (Uni-Text): The extent to which **Text** modality enables you to predict without the other modality.
- Question 3. (Modality-balance): The extent to which **both** modalities enable you to predict that you would not otherwise make using either modality individually.

For a specific sample, we first average the three scores of each annotator respectively, and then select the type with the highest score as the bias type of this sample.

We conducted Krippendorff's alpha test (Krippendorff, 2011) to assess the reliability and agreement of human annotations. As presented in Table 11, all alpha values exceed 0.8, which demonstrates a high level of agreement among the three annotators and further substantiate the validity of our human annotations.

J Error Analysis

As shown in Figure 6, we conduct an error analysis on the "Uni-image" category, which exhibited the lowest performance in our multi-view analysis. We found that the multi-view analysis struggles to correctly identify well-edited images (Figure 6, left) or images synthesized by large vision-language models (Figure 6, right). Although these images may appear seamless at the pixel level, they contain misinformation at the semantic level. However, the multi-view analysis incorrectly classifies these samples as "Modality-balance". We attribute this issue to the limitations of current MMD models, which are not yet equipped to handle such complex

Figure 6: Error cases of multi-view analysis. The modality bias of these two samples should be "Uni-image".

cases. As more advanced techniques are developed, these types of errors may decrease, improving the accuracy of automated bias evaluation systems.

K Discussion

Firstly, the definition of "modality bias" is derived from (Guo et al., 2023), referring to the tendency of a model to rely on a single modality (e.g., image or text) for decision-making. However, there might be multiple forms of modality bias in practical applications according to varying definitions. Theoretically, each view (i.e., Modality benefit, Modality flow, and Modality causal effect) holds a distinct bias recognition pattern, so the ensemble multi-view analysis is robust to such diverse forms of bias.

Secondly, from the view of modality benefit, we can determine the type of modality bias by comparing the final output benefit of image modality and text modality. Nevertheless, when $V\left(x^{m_1}, 0^{m_2}\right), V\left(0^{m_1}, 0^{m_2}\right), V\left(x^{m_2}, x^{m_1}\right)$, and $V\left(x^{m_2}, 0^{m_1}\right)$ all equal zero, the model is unable to make accurate predictions. In such cases, we hyposize the difficulty of samples exceeds the discriminative capacity of this view, and the Shapely value cannot provide a reasonable classification.

Thirdly, we investigate the automated sample-specific modality bias analysis for multimodal misinformation benchmarks. This deepens our understanding of such benchmarks and provides new insights for online multimodal content analysis. However, this method can be applied not only in the field of misinformation detection. Our automated analysis is broadly applicable to general tasks like visual question answering (VQA) and extends to other modalities like audio.

Fourthly, while our work focuses on identifying and analyzing modality bias, improving misinformation detection based on bias analysis is a direction worthy of in-depth exploration. We encourage future work to improve model training by leveraging modality bias analysis results as auxiliary labels during the optimization process of multimodal misinformation detection.

Fifthly, in real-time applications, the primary computation cost arises from the inference of large models. While the forward of modality flow involves a MiniCPM-V 2.6, the modality causal effect incorporates both MiniCPM-V 2.6 and Llama3-8B. This results in a relatively slower inference speed for these two views. A potential approach is utilizing quantized versions of large models in real-time applications to reduce computational costs.