Rethink to Check: Mitigating Confirmation Bias for End-to-End Multimodal Fact-Checking

Anonymous ACL submission

Abstract

End-to-end multimodal fact-checking (MFC) aims to assess the truthfulness of claims using retrieved multimodal evidence. Existing methods rely on the stance extracted from the evidence, achieving good performance with annotated gold evidence, but performing poorly with system-retrieved evidence. The key issue is that the existing model is only exposed to annotated gold evidence during training, inevitably leading to confirmation bias. Such bias refers to that the model tends to treat low-quality system-retrieved evidence as high-quality gold evidence during testing, thus resulting in low robustness and generalization of the model. To mitigate the bias, we propose a novel multi-check framework with causal intervention and counterfactual reasoning. It incorporates three independent checkers to verify claims from diverse perspectives, thereby ensuring a more balanced and accurate fact-checking. Specifically, we first construct two distinct types of counterfactual instances via causal intervention. Then, we apply counterfactual reasoning to train three independent checkers with tailored counterfactual instances or annotated samples. During inference, we eliminate confirmation bias by synthesizing the verification results of all checkers. Experimental results demonstrate the superiority of our proposed framework to state-of-the-art methods, showing performance improvements of 5.5% and 16.9% with annotated and system-retrieved evidence, respectively. Our code will be released once the paper is accepted.

1 Introduction

Fact-checking aims to assess the authenticity of a claim by analyzing the relevant evidence (Guo et al., 2022). It can significantly mitigate the serious social harm inflicted by misinformation, such as the crisis of medical trust during COVID-19 (Islam et al., 2020) and interference in the 2016 U.S. presidential election (Bovet and Makse, 2019).

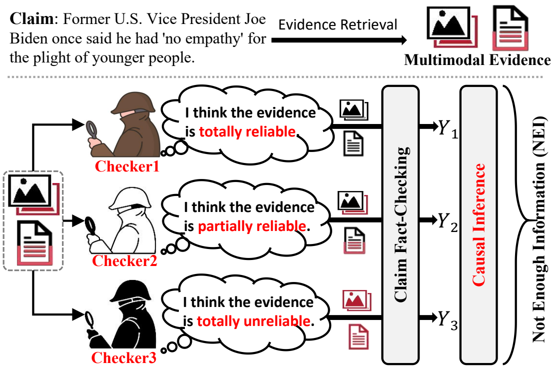


Figure 1: Illustration of evidence retrieval (top) and our multi-check method (bottom).

However, current fact-checking requires analyzing intricate multimodal evidence, and relying on manual fact-checking is inefficient (Schlichtkrull et al., 2023). Thus, it is crucial and urgent to develop automated multimodal fact-checking (MFC).

The current MFC efforts include out-of-context (OOC) detection (Luo et al., 2021) and end-to-end scenarios (Yao et al., 2023). The former is an extension of the image repurposing detection task (Sabir et al., 2018), which requires determining whether an image corresponds to the text. The latter is an expansion of textual fact-checking into multimodal scenarios and consists of multimodal evidence retrieval and fact-checking (Akhtar et al., 2023). Compared to single OOC detection, end-to-end MFC is more challenging and can be adapted to more scenarios (including OOC (Geng et al., 2024)), which is closely aligned with real-world fact-checking. Thus, this work focuses on the end-to-end MFC, which leverages retrieved multimodal evidence to verify the claims.

The focus of existing MFC methods is verifying the given claims according to the stance of retrieved evidence (Yao et al., 2023; Yuan et al., 2023). Unfortunately, the quality of retrieved evidence often varies significantly, sometimes including conflicting information with different stances or false information. The unreliable evidence poses great challenges to MFC and limits fact-checking performance. The underlying reason is that the existing model is only exposed to authentic evidence (gold evidence) during model training, leading to the model suffering from **confirmation bias** (Nickerson, 1998). Specifically, this bias refers to the model's tendency to treat system evidence as high-quality gold evidence during testing ($checker1$ in Figure 1), which inevitably introduces the possible conflicting or false information in system evidence into fact-checking, thereby affecting the model's robustness and generalizability.

In this paper, we propose a multi-check framework, introducing causal intervention and counterfactual reasoning to alleviate the above confirmation bias. Our key motivation is to rethink the evidence and check the claims from different perspectives. Specifically, we imagine a counterfactual world where each claim is verified by three independent fact-checkers, treating the same evidence from different perspectives. As illustrated in Figure 1, $checker1$ considers the evidence reliable while $checker2$ considers the evidence partially reliable, and $checker3$ considers the evidence unreliable. During verification, $checker2$ and $checker3$ are used to model possible conflicting and false information in system evidence and eliminate confirmation bias in $checker1$ from a causal perspective.

Driven by the aforementioned motivation, the proposed multi-check frame is divided into three main steps: multimodal counterfactual instance construction, multi-check training, and multi-check reasoning. Specifically, 1) To effectively train diverse checkers, we leverage a causal model to construct two distinct types of counterfactual instances by intervening on the original training samples. 2) Considering the causal effect of counterfactual instances, we tailor distinct training objectives for individual checkers. 3) During reasoning, we feed retrieved evidence into all checkers and fuse all verification results as the final prediction. Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to investigate the confirmation bias under real-world end-to-end MFC. We provide the theoretical foundation from the causal perspective to analyze the confirmation bias.
- We propose a causal intervention and counterfactual reasoning based framework that introduces a novel multi-check process to mitigate confirmation bias.
- Extensive experiments demonstrate the effectiveness of our model compared to the state-of-the-art (SOTA) MFC methods and LLMs (GPT-3.5 and GPT-4o).

2 Related Work

2.1 Multi-modal Fact-Checking

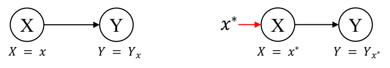
Some multimodal fact-checking (MFC) works (Abdelnabi et al., 2022; Yuan et al., 2023; Zhang et al., 2023; Papadopoulos et al., 2023) focus on the out-of-context (OOC) misinformation and serve it as an image-text mismatch checking task. (Abdelnabi et al., 2022) first introduce the multi-modal cycle-consistency to detect the mis- or disinformation of image-text pairs. (Yuan et al., 2023) models the stance of external evidence to aid misinformation detection. (Zhang et al., 2023) introduce an improved attention network to facilitate a comprehensive understanding of contextual information. To foster MFC, (Yao et al., 2023) propose end-to-end MFC, Mocheg, which encompasses the complete phases of fact-checking and more closely aligns with real-world MFC. Specifically, end-to-end MFC requires automatically retrieving evidence relevant to the claim and predicting the label based on system-retrieved evidence.

However, due to the low accuracy of evidence retrieval, existing methods are plagued by incomplete and unreliable evidence, which leads to poor generalization performance of the models in practical application. In other words, current methods overfit the gold evidence in the training phase and exhibit low robustness during real-world testing with system-retrieved evidence.

2.2 Confirmation Bias

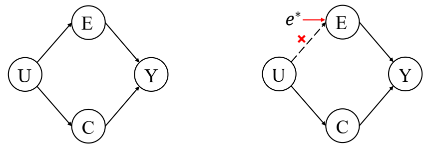
Confirmation bias (Nickerson, 1998) is a psychological concept referring to the inclination to favor information that aligns with one's preexisting beliefs while disregarding conflicting information. Such bias often occurs in semi-supervised or unsupervised learning, referring to the noise accumulation when the model is trained using incorrect predictions (Tarvainen and Valpola, 2017).

However, in real-world end-to-end MFC, confirmation bias has not yet been studied or defined. We are the first to investigate the confirmation bias in this field. Specifically, we observed confirmation bias during training which can lead the model to



(a) (b)

Figure 2: Example of the causal graph where X and Y represent the cause and effect respectively, with $*$ denoting reference values.



(a) Causal graph with confounder (Factual world).

(b) Counterfactual world by causal intervention.

Figure 3: The causal graphs for fact-checking. E: multimodal evidence, C: claim, Y: label of claim, U: confounder. $*$ denotes the reference value.

treat the system-retrieved evidence as normal annotated evidence during real-world testing, reducing the model's robustness and generalizability.

2.3 Causal Inference

Recently, causal inference (Pearl et al., 2016) has been widely used in various deep-learning tasks, such as visual question answering (Niu et al., 2021), multimodal information extraction (Zhou et al., 2024), fake news detection (Tian et al., 2022; Chen et al., 2023), etc. As for fact-checking, (Tian et al., 2022) formulate dataset biases as causal effects and debias it based on counterfactual reasoning.

Unlike debiasing dataset biases, we discover the gap between the evidence used in training and testing. To address this, we construct two types of counterfactual instances to train multiple checkers to rethink the evidence and recheck the claim.

3 Preliminaries

3.1 Causal Graph

Causal graph (Pearl et al., 2016) is used to help analyze the causal effects between different variables, represented by a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$. $\mathcal{N}$ represents the set of variables, and $\mathcal{E}$ represents directed causal edges between variables. As shown in Figure 2(a), $X \rightarrow Y$ denotes the causal pathway between two variables X and Y, where X is the cause and Y is the effect.

3.2 Counterfactual Reasoning and Causal Effect

Counterfactual reasoning (Pearl, 2009) is a statistical inference technique employed to infer potential outcomes under hypothetical circumstances diverging from the factual world. For instance, Figure 2(a) is a factual world where the calculation of effect Y is denoted as $Y_x = Y(X = x)$.

To estimate the causal effect (Pearl, 2022) of a treatment variable X on a response variable Y, we conduct the counterfactual reasoning by causal intervention. As shown in Figure 2(b), we construct a counterfactual world where variable X is intervened to be reference value x^*. Empirically, we denote the intervention operation as $do(\cdot)$. And we define the causal effect (CE) of X on Y as:

$$\begin{aligned} CE_{X \rightarrow Y} &= Y_x - Y_{x^*} \\ &= Y(X = x) - Y(do(X = x^*)) \end{aligned} \quad (1)$$

4 Method

We first formalize the fact-checking task into a causal graph to analyze confirmation bias and causal effects between different factors in Section 4.1. Then we present our multi-check framework consisting of multimodal counterfactual instance construction (4.2), multi-check training (4.3), and multi-check reasoning (4.4).

4.1 Causal Graph of Fact-checking

Figure 3(a) shows the causal graph of the fact-checking process. Nodes E and C denote the multimodal evidence features and claim features respectively. Node Y is the task label and $E \rightarrow Y$ represents the causation from variable E to variable Y. Notable, U denotes the confounder variable that influences both variables E and C, which implies evidence annotator to collect claim-evidence pairs (i.e., $U \rightarrow (C, E)$). During training, U represents the annotator to collect gold evidence (high quality), while during testing, U denotes the evidence retriever to retrieval system evidence (low quality). Confirmation bias arises when the model treats system evidence as gold evidence during testing, leading to low robustness and poor generalization.

4.2 Counterfactual Instance Construction

To alleviate the aforementioned confirmation bias, we cut off the link $U \rightarrow E$ as depicted in Figure 3(b), and construct a counterfactual world by forcibly changing the value of variable E through intervention operation $do(E = e^*)$.

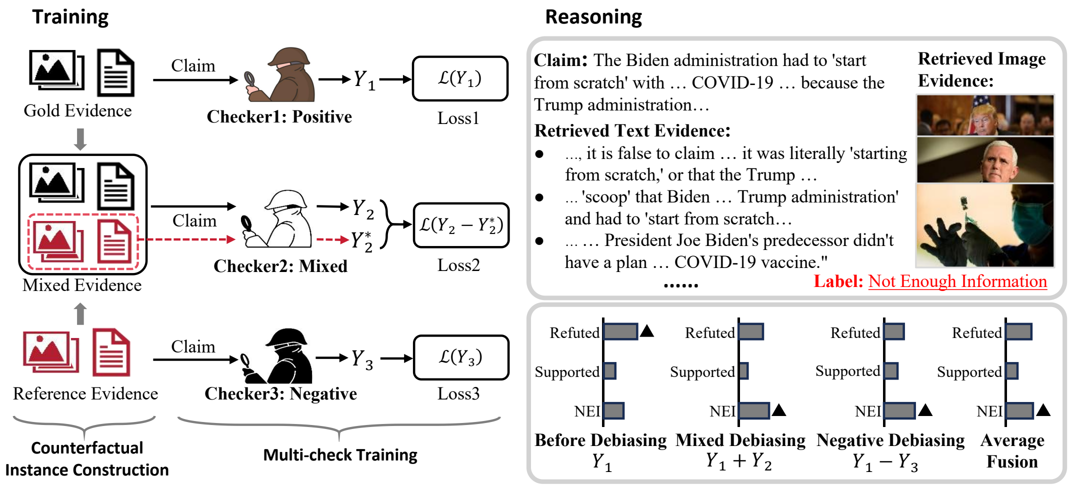


Figure 4: Illustration of the training and reasoning processes of our multi-check framework.

As shown in Figure 4, in the counterfactual world, we proposed a multi-check framework that introduces three independent fact-checkers (positive $checker1$, mixed $checker2$, and negative $checker3$) to rethink evidence and recheck claims from different perspectives. To train different checkers, we require corresponding training data, where $checkers1$ is trained with the gold data to reflect the factual world. For $checkers2$ and $checkers3$, we construct two distinct counterfactual instances for their training.

Given the raw gold sample (c, e_t, e_i) which denotes claim, text evidence, and image evidence respectively. The gold evidence (e_t, e_i) in the training set is reliable. To introduce unreliable evidence during training, we randomly select K irrelevant text and image evidence from the multimodal evidence set as reference unreliable evidence (e_t^*, e_i^*). For the counterfactual instance of $checker3$, we do the interventions $do(E_t = e_t^*)$ and $do(E_i = e_i^*)$ on the variable E to cut off the link $U \rightarrow E$. Empirically, the intervention can be operated by replacing the gold evidence $e_{i,t}$ with the reference evidence (false evidence) $e_{i,t}^*$ to construct the counterfactual sample (c, e_t^*, e_i^*). Similarly, we do the interventions $do(E_t = e_t + e_t^*)$ and $do(E_i = e_i + e_i^*)$, replacing the gold evidence with conflicting evidence with different stances to construct the counterfactual sample $(c, e_t + e_t^*, e_i + e_i^*)$ for $checker2$. Through the above process, we obtain the training samples required for multi-check training.

4.3 Multi-check Training

After obtaining training data including counterfactual instances, we train our multi-check framework. For each checker, given a claim c and multimodal evidence $\{e_t^1, e_t^2, ...\}\&\{e_i^1, e_i^2 ...\}$. Following (Yao et al., 2023), we use CLIP to extract fine-grained representations and detect stance representation from each claim-evidence pair. Finally, all stance representations are used to predict the label of C. Model details can be found in (Yao et al., 2023). Notable, we use the same model architecture but different training objectives for different checkers.

Checker1. To learn the mapping between gold samples and their truthfulness labels, we feed (c, e_t, e_i) into $checker1$, obtain the output Y_1, and use the cross-entropy loss as the loss function:

$$Y_1 = Y(C = c, E_t = e_t, E_i = e_i), \tag{2}$$

$$\mathcal{L}_1 = -\log\left(\frac{\exp(Y_{1,i})}{\sum_{j=0}^{2}\exp(Y_{1,j})}\right), \tag{3}$$

where i denotes the index of the truthfulness label.

Checker2. As discussed in Section 1, $checker2$ aims to enhance the model performance under conflicting evidence with different stances towards to the claim. We hope $checker2$ can assist the model in identifying partial reliable evidence during testing. Based on counterfactual reasoning, we feed $(c, e_t + e_t^*, e_i + e_i^*)$ and obtain Y_2 as follow:

$$Y_2 = Y(C = c, do(E_t = e_t + e_t^*), do(E_i = e_i + e_i^*)). \tag{4}$$

To avoid $checker2$ learning the wrong mapping between unreliable evidence and truthfulness labels, we eliminate the causal effect of unreliable evidence on the truthfulness label by subtraction from the causal perspective. Specifically, we input (c, e_t^*, e_i^*) and subtract the output Y_2^*, and then compute the cross-entropy loss as follow:

$$Y_2^* = Y(C = c, do(E_t = e_t^*), do(E_i = e_i^*)), \quad (5)$$

$$\mathcal{L}_2 = -\log \left(\frac{\exp((Y_2 - Y_2^*)_i)}{\sum_{j=0}^{2} \exp((Y_2 - Y_2^*)_j)} \right). \quad (6)$$

Checker3. To further reduce confirmation bias, we propose $checker3$ to capture the wrong mapping between unreliable evidence (i.e., false information) and truthfulness labels. Therefore, during training, we maximize the confirmation bias, i.e., we hope $checker3$ treats system evidence as unreliable evidence (see Figure 1) to verify the claim. Such wrong mapping will be reduced during inference via subtraction. To do this, we feed (c, e_t^*, e_i^*) into $checker3$ and obtain Y_3. The training loss is calculated as follows:

$$Y_3 = Y(C = c, do(E_t = e_t^*), do(E_i = e_i^*)), \quad (7)$$

$$\mathcal{L}_3 = -\log \left(\frac{\exp(Y_{3,i})}{\sum_{j=0}^{2} \exp(Y_{3,j})} \right). \quad (8)$$

Note that the three checkers mentioned in our framework represent three sub-models that have the same model structure but do not share parameters. Therefore, they have high flexibility in training and can be trained together or separately. To learn the model parameters, we minimize a multi-check training objective as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3, \quad (9)$$

where λ_1, λ_2, and λ_3 are the trade-off hyperparameters to adjust the effect of different views.

4.4 Multi-check Reasoning

During reasoning, we have claim c as well as the multimodal evidence (e_t^s, e_i^s) retrieved by the system. To verify c, we feed (c, e_t^s, e_i^s) into our multi-check MFC framework and obtain three outputs (Y_1, Y_2, Y_3) from different checkers. Y_1 as the output of $checker1$, we use it as a benchmark output with confirmation bias and employ Y_2 and Y_3 to mitigate such bias. Specifically, for the output of $checker2$, we employ addition $(Y_1 + Y_2)$ to enhance the causal effect of reliable evidence within the system evidence on the truthfulness label. In addition to the output of $checker3$, we use subtraction $(Y_1 - Y_3)$ to reduce the aforementioned wrong mappings between unreliable evidence (i.e., false information) within the system evidence and truthfulness labels. Thus, we obtain two debiased results $Y_1 + Y_2$, $Y_1 - Y_3$ and the result Y_1 before debiasing. Note that each of the above results may be best in individual scenarios (e.g., $Y_1 + Y_2$ is the best result in Figure 4). However, due to the varying quality of system evidence, employing a fusion strategy to integrate the above three results is necessary and beneficial, such opinion is verified in ablation experiments. Specifically, we employ an averaging fusion strategy to integrate the above three results. Besides, we explore more fusion strategies in the experimental section.

Data	Train	Val	Test
# Claims	11,669	1,490	2,440
# Refuted Labels	4,542	488	825
# Supported Labels	3,826	501	817
# NEI Labels	3,301	501	800
# Text evidence	23,545	4,067	6,268
# Image evidence	8,927	1,178	2,007

Table 1: Statistics of the MOCHEG dataset.

5 Experiments

In this section, we conducted experiments for quantitative and qualitative analysis to validate the effectiveness of our proposed method.

5.1 Experimental Settings

5.1.1 Dataset

We conducted experiments on the only existing end-to-end multimodal fact-checking dataset:

MOCHEG: a large-scale dataset consisting of 15,601 claims where each claim is annotated with a truthfulness label and a ruling statement, and 33,880 textual paragraphs and 12,112 images in total as evidence. We preprocess and divide the dataset in the same way as in (Yao et al., 2023). The dataset statistic is shown in Table 1. Following prior works, we adopt Macro F1 as evaluation metric to assess the performance of our model.

5.1.2 Implementation Details

Regarding evidence retrieval, we use the pretrained retrieval model from (Yao et al., 2023) to retrieve top-5 text and image evidence respectively for claim verification. We use frozen CLIP-ViT-B/32 as our backbone. For hyperparameter settings, the training batch size is 128, the training epoch is 50, and the Adam optimizer with a learning rate 1e-5 is used to update the parameters. Besides, the trade-off hyperparameters λ_1, λ_2, λ_3 are set to 1.0. According to the early-stopping strategy, the

Methods	F-score (%)
Majority Label	33.78
Average Similarity (Gold)	32.72
SpotFakePlus (Gold)	44.11
Pre-CoFactv2 (Gold)	47.17
Mocheg (Gold) †	51.64
Ours (Gold)	**54.48**
Mocheg (System) †	42.44
Ours (System)	**49.61**

Table 2: Main results comparing with the SOTA methods. Note that *Gold* denotes gold multi-modal evidence while *System* means system-retrieved evidence. † represents our re-implemented results.

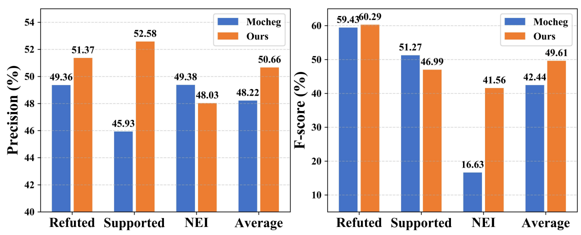


Figure 5: Performance comparison between Mocheg and our method in different truthfulness labels.

	Methods	Acc.	F-score (%)
Gold	Full Model	**55.57**	**54.48**
	w/o C_2	55.16	53.83
	w/o C_3	54.79	53.19
	w/o C_2+C_3	54.38	51.64
	w/o CI	54.99	53.04
	w/o CT	55.04	52.81
System	Full Model	**50.86**	**49.61**
	w/o C_2	50.49	48.93
	w/o C_3	49.67	47.41
	w/o C_2+C_3	47.91	42.44
	w/o CI	48.40	44.70
	w/o CT	47.83	43.81

Table 3: Evaluation results for ablation study.

training process ends when the Accuracy on the validation set does not increase within 10 epochs. We evaluate the best model on the test set. To show the superiority of our method in eliminating confirmation bias, for the factual $checker1$, we choose the same model as mocheg and train our proposed two counterfactual checkers separately. We conduct the experiments in Ubuntu 18.04.5 with a single NVIDIA A6000 GPU with 48GB of RAM.

5.2 Compared Methods

Due to the scarcity of end-to-end multimodal fact-checking, we followed previous work (Yao et al., 2023) and selected the current SOTA methods:

- **SpotFakePlus** (Singhal et al., 2020) focus on capturing text and image's semantic and contextual information. (Yao et al., 2023) adapts it to the multi-modal fact-checking task.
- **Pre-CoFactv2** (Du et al., 2023) is a novel framework with parameter-efficient foundation models that achieves SOTA results at the Factify 2 challenge (Suryavardan et al., 2023).
- **Mocheg** (Yao et al., 2023) first propose end-to-end MFC and introduce stance representation to help fact verification, achieving SOTA performance on the challenging Mocheg.

5.3 Performance Comparison

Table 2 shows the experimental results of our proposed framework compared with SOTA baselines under *Gold* and *System* settings, respectively. Note that the system-retrieved evidence used in different methods is the same. From Table 2, we observe that our method achieves the best performance. Specifically, our method improves the average F-score by 5.5% and 16.9% compared to the second-best method (i.e., Mecheg) under the *Gold* and *System* settings respectively, highlighting the superiority of our proposed method.

Notably, the performance improvement under the *System* setting is larger than that under the *Gold* setting (16.9% *vs* 5.5%). Moreover, our method under the *System* setting outperforms most baselines (e.g., SpotFakePlus, Pre-CoFactv2) under the *Gold* setting. This indicates that our method has a significant advantage in real-world MFC. We believe that our method benefits from the evidence rethink and the claim recheck via our proposed multi-check process.

We further compare the performance of our method with Mocheg in detail truthfulness labels under real-world *System* setting. Figure 5 shows the precision and F-score in different labels. Specifically, our method is superior in the majority of cases and falls slightly short in a few cases (precision in *NEI*, F-score in *Supported*). Overall, considering all types of labels, our method outperforms Mocheg, exhibiting more stable performance across various labels and higher model robustness.

5.4 Ablation Study

To study the impact of each component of our proposed method, we conduct ablation experiments by defining the following variants:

w/o C_2 or C_3: Remove $checker2$ or $checker3$.

Methods		Acc.	F-score (%)
Gold	Average	**55.57**	**54.48**
	Max	54.83	53.98
	Voting	55.45	54.34
System	Average	**50.86**	49.61
	Max	50.66	**49.87**
	Voting	50.75	49.80

Table 4: Results of different reasoning strategies.

Methods		Acc.	F-score(%)
Gold	GPT-3.5	53.64	45.76
	GPT-4o	**58.52**	50.63
	Ours	55.57	**54.48**
System	GPT-3.5	46.15	39.44
	GPT-4o	**53.32**	47.74
	Ours	50.86	**49.61**

Table 5: Comparison results with LLMs.

Methods	# Refuted	# Supported	# NEI
Raw Distribution	825	817	800
GPT-3.5 (Gold)	628	1,723	91
GPT-4o (Gold)	1,318	962	162
Ours (Gold)	1,176	614	652
GPT-3.5 (System)	437	1,833	172
GPT-4o (System)	1,267	935	240
Ours (System)	1,172	660	610

Table 6: Statistics on the results of different methods.

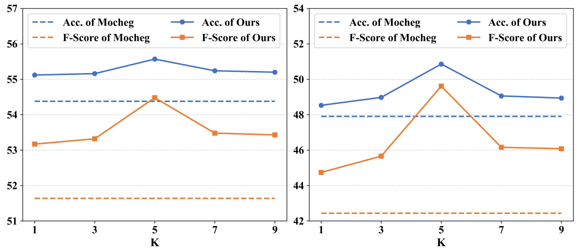


Figure 6: Impact of different values of K. Note that the left and right sub-figures represent the results under *Gold* and *System* settings, respectively.

w/o CI: Construct counterfactual instances with only text changes, leaving images unchanged.

w/o CT: Construct counterfactual instances with only image changes, leaving textual unchanged.

The ablation results in Table 3 show that all proposed components are beneficial. Specifically, when we remove $checker2$ (w/o C_2) or $checker3$ (w/o C_3), the performance drops. When we remove both $checker2$ and $checker3$ (w/o C_2+C_3), the performance further drops, demonstrating the effectiveness of the multi-check process. Besides, we also perform the ablation study on the specific construction of counterfactual instances. When we construct counterfactual instances by changing only the unimodal evidence ($w/o\ CI$ or $w/o\ CT$), the performance drops, indicating the superiority of our counterfactual instance construction.

As shown in Table 3, the results show that our proposed modules are more effective in the $System$ setting. This is consistent with our hypothesis that our approach can mitigate confirmation bias, and the harm of confirmation bias is more pronounced in the $System$ setting.

5.5 Impact of Different Reasoning Strategies

We investigated the impact of different fusion strategies during multi-check reasoning. The $Average$ strategy refers to averaging the outputs from three checkers while the Max strategy aims to select the output with the highest probability. The $Voting$ strategy refers to predicting the label with the most votes from all checkers. Note that if no consensus in the $Voting$ strategy, the Max strategy will be used. From Table 4, we find that the $Average$ strategy achieves the best performance in the accuracy metric. This suggests that integrating all checkers is most effective, indicating the effectiveness of our multi-check approach. We believe introducing different checkers based on actual conditions and applying various strategies is worth exploring.

5.6 Impact of the Value of K

We tried different values of K, i.e., the number of evidence selected to construct the counterfactual instance. Figure 6 shows that our method always outperforms Mocheg, and $K = 5$ leads to the best performance. We analyze the reason accounting for the results is that a small amount of evidence may not be sufficient for multi-check training, while too much irrelevant evidence may lead to biases in model training. This indicates the effectiveness of our approach and emphasizes the importance of selecting an appropriate quantity of noise evidence.

5.7 Comparision with LLMs

We apply the OpenAI-API [1] (gpt-3.5-turbo-0125[2] and gpt-4o[3]) to the end-to-end MFC using the prompt template. The implementation details are described in Appendix A.1. From Table 5, we can observe that our method outperforms GPT-3.5 in both accuracy and F-score, demonstrating the effectiveness of our approach. Further, compared to the current state-of-the-art GPT-4o, our model is

[1] https://platform.openai.com/docs/api-reference
[2] https://platform.openai.com/docs/models/gpt-3-5-turbo
[3] https://platform.openai.com/docs/models/gpt-4o

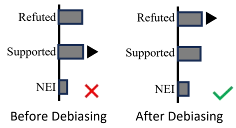
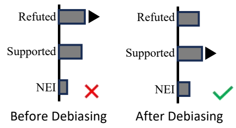
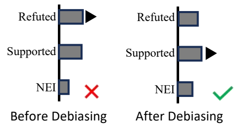
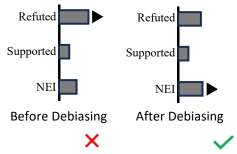
Claim	Textual Evidence	Image Evidence	Multi-view Debiasing
Says William Barr tweeted, 'BREAKING NEWS Senator Mitt Romney is the only Republican Senator who voted to remove President Trump from office... **Refuted**	• And a similar plea came from Sen. Mitt Romney, R-Utah, the only Republican senator who voted to to remove Trump from office. • Senator Mitt Romney, Republican of Utah, was the only member to break with his party, voting to remove Mr. Trump from office. • Romney votes to convict Trump of abuse of power, the only Republican to support removing the president. • Nevertheless, in a statement after Mr. Trump's tweet, Attorney General William P. Barr said the F.B.I. ...		
Disney is replacing the 'Tower of Terror' attraction with a 'Guardians of the Galaxy' themed ride at their Disney California Adventure Park. **Supported**	• The popular Twilight Zone Tower of Terror attraction at Disney California Adventure Park will be transformed into a 'Guardians of the Galaxy' ... • The Guardians of the Galaxy ride is a much newer addition... • 'Tower of Terror is a classic Disney ride and ... Guardians of the Galaxy? • Disney is replacing the vertigo-inducing 'Twilight Zone'-themed elevator ride at its California theme park with ... space super heroes. • Tower of Terror to be Removed, Replaced With Elsa's Ice Castle Disney announced ... replaced by Elsa's Ice Castle, featured in the movie, Frozen.		
Ellen DeGeneres has decided to end her long-running daytime talk show in 2022. **Supported**	• Let's start with the decision to end the show in 2022: ... • In June 2016, an ... disguised ... reporting that Ellen DeGeneres would be leaving her popular daytime television talk show to sell skin care products. • Ellen Degeneres recently announced she will be leaving The Ellen Show in November to promote a new skincare line that was recently voted... • In fact, NBCUniversal Owned Television Stations announced in January 2016 that the Ellen DeGeneres Show had been renewed through 2020.		
The Biden administration had to start from scratch with a comprehensive COVID-19 vaccine distribution plan because the Trump administration had no working plan. **Not Enough Information (NEI)**	• ... the underlying claim was whether the Harris-Biden administration ... 'start from scratch' with ... because their predecessors had no working plan. • ... its so-called 'scoop' that Biden inherited 'no vaccine distribution plan from the Trump administration' and had to 'start from scratch.' • The Trump administration has released no comprehensive plan to combat COVID-19, except ... the development and distribution of vaccines. • it is false to claim that it was literally 'starting from scratch,' or that the Trump administration had done nothing .... • Biden administration officials were reportedly ... President Joe Biden's predecessor didn't have a plan ... COVID-19 vaccine.		

Figure 7: Some representative cases, where green font indicates support for the claim, red indicates refutation, and blue indicates insufficient information. Note that only some key evidence is shown.

lagging in accuracy. However, our method outperforms both LLMs in the F-score. We analyze the reason accounting for the results is that ChatGPT tends to answer with "support" or "refuted". The statistics in Table 6 show that both LLMs exhibit significant classification bias, especially GPT-3.5 (628/825, 1723/817 and 91/800 under gold setting). GPT-4o outperforms GPT-3.5 but still exhibits noticeable bias (1318/825, 162/800). In contrast, our model demonstrates smaller classification bias, indicating that our approach is more robust than current LLMs in the MFC. Overall, our method is more feasible for the end-to-end MFC.

5.8 Case Study

Figure 7 shows some representative cases of our approach. Some key information is highlighted in different colors and the results before and after multi-check debiasing are illustrated. For the refuted example (first one), before debiasing, the model supports the claim based on partial evidence (green), yet ignores conflicting information that contradicts the claim (red). However, our multi-check method can capture such conflicting information and then make correct predictions after debiasing. For the NEI examples (last one), the model also ignores conflicting information in the evidence and relies on some piece of evidence. For the supported examples (second and third ones), we can see that the model is misled by the retrieved unreliable evidence (e.g., “Replaced With Elsa’s Ice Castle”, “June 2016 ... leaving The Ellen Show”) and makes incorrect predictions. Our multi-check process can rethink the evidence, and find reliable evidence (e.g., “be transformed into a ‘Guardians of the Galaxy’", “the decision to end the show in 2022...") to recheck the claims. These cases show the superiority of our proposed framework, which eliminates the confirmation bias by introducing counterfactual checkers to rethink the evidence.

6 Conclusion

In this work, we observe the confirmation bias in real-world end-to-end MFC. To eliminate this bias, we propose a novel causal intervention and counterfactual reasoning based multi-check framework for end-to-end MFC. We formulate the end-to-end MFC as a causal graph and reduce the confirmation bias by multi-check learning. Specifically, we imagine a counterfactual world and construct two types of counterfactual instances via causal intervention for multi-check training. The outputs of all checkers are fused to verify claims during reasoning. Eventually, experiments on a public large-scale dataset and some cases are given, showing the excellent performance of our proposed method.

7 Limitations

We recognize the following limitations in our approach: (1) While employing random sampling to construct counterfactual training examples for $checker2$ and $checker3$ is efficient, it may not always yield suitable counterfactual examples for every case. (2) This paper does not thoroughly investigate explanation generation. From Table 8 in Appendix A.3, given the same evidence, more accurate prediction results (ours) do not significantly improve the performance of explanation generation. This suggests that current explanation generation models do not fully leverage the information from verification results, relying instead on summarizing the provided evidence. Moreover, from Table 3, it is evident that counterfactual construction significantly impacts real-world MFC (system setting), especially image counterfactual instances construction. This indicates the low performance of current multimodal evidence retrieval (especially image evidence retrieval, see Table 8 in the appendix for details). In future work, we plan to explore more appropriate methods for counterfactual instances construction and to delve deeper into the study of explanation generation and evidence retrieval.

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14920–14929. IEEE.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of EMNLP 2023, Singapore, December 6-10, 2023*, pages 5430–5448. Association for Computational Linguistics.

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *ACL 2023*, pages 627–638. Association for Computational Linguistics.

Wei-Wei Du, Hong-Wei Wu, Wei-Yao Wang, and Wen-Chih Peng. 2023. Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. In *Proceedings of De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023*, volume 3555 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, and Iryna Gurevych. 2024. Multimodal large language models to support real-world fact-checking. *CoRR*, abs/2403.03627.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206.

Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. 2020. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American journal of tropical medicine and hygiene*, 103(4):1621.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL 2020*, pages 7871–7880. Association for Computational Linguistics.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. In *EMNLP 2021*, pages 6801–6817. Association for Computational Linguistics.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

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 *CVPR 2021*, pages 12700–12710. Computer Vision Foundation / IEEE.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2023. RED-DOT: multimodal fact-checking via relevant evidence detection. *CoRR*, abs/2311.09939.

Judea Pearl. 2009. Causal inference in statistics: An overview.

Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep multimodal image-repurposing detection. In *MM 2018*, pages 1337–1345. ACM.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *AAAI 2020*, pages 13915–13916. AAAI Press.

S. Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Naresh Reganti, Aman Chadha, Amitava Das, Amit P. Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Factify 2: A multimodal fake news and satire news dataset. In *Proceedings of De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023, Washington DC, USA, February 14, 2023*, volume 3555 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30: Annual Conference on NeuralIPS Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1195–1204.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing NLU models via causal intervention and counterfactual reasoning. In *AAAI 2022*, pages 11376–11384. AAAI Press.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *SIGIR 2023*, pages 2733–2743. ACM.

Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 2023. Support or refute: Analyzing the stance of evidence to detect out-of-context mis- and disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4268–4280. Association for Computational Linguistics.

Fanrui Zhang, Jiawei Liu, Qiang Zhang, Esther Sun, Jingyi Xie, and Zheng-Jun Zha. 2023. Ecenet: Explainable and context-enhanced network for mutimodal fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 1231–1240. ACM.

Baohang Zhou, Ying Zhang, Kehui Song, Hongru Wang, Yu Zhao, Xuhui Sui, and Xiaojie Yuan. 2024. MCIL: multimodal counterfactual instance learning for low-resource entity-based multimodal information extraction. In *LREC/COLING 2024*, pages 11101–11110. ELRA and ICCL.

A Appendix

A.1 Prompt Template

Prompt:
Given the following claim and relevant evidence, please determine the label of the claim. You can only answer (support, refuted, or not enough information).
Claim: { }
Evidence: { }
Label: support, refuted, or not enough information?

Note that when using GPT-4o, we did not provide image evidence. This is because uploading images via the API is very expensive now.

A.2 Results of Multimodal Evidence Retrieval

Media	N	Rec@N	Pre@N	NDCG@N	MAP@N
Image	5	17.84	4.87	14.39	12.49
	10	23.20	3.17	16.22	13.30
Text	5	18.35	14.26	22.49	16.27
	10	23.00	9.57	23.01	15.51

Table 7: Performance of multimodal evidence retrieval.

Following (Yao et al., 2023), we retrieve the top-5 text and image evidence for every claim, the performance of multimodal evidence retrieval is shown in Table 7.

A.3 Results of Explanation Generation

Evidence	Truthfulness	ROUGE-1	ROUGE-2	ROUGE-L
Gold	Mocheg	45.80	26.89	35.33
	Ours	45.84	26.90	35.34
System	Mocheg	35.71	16.44	25.22
	Ours	35.81	16.39	25.15

Table 8: Performance of explanation generation.

We used the pre-trained BART-large model (Lewis et al., 2020) as a generator for our explanation generation experiments. Specifically, we provided the generator with the same evidence and fact-checking results obtained from different methods. The results are shown in Table 8,