Insight-A: Exploring Attribution with MLLM Insights for Multimodal Misinformation Detection

Anonymous ACL submission

Abstract

AI-generated content (AIGC) technology has emerged as a prevalent alternative to create multimodal misinformation on social media platforms, posing unprecedented threats to societal safety. However, standard prompting leverages multimodal large language models (MLLMs) to identify the emerging misinformation, which ignores the misinformation attribution. To this end, we present Insight-A, exploring attribution with MLLM insights for detecting multimodal misinformation. Insight-A makes two efforts: I) attribute misinformation to forgery sources, and II) an effective pipeline with hierarchical reasoning that detects distortions across modalities. Specifically, to attribute misinformation to forgery traces based on generation patterns, we devise cross-attribution prompting (CAP) to model the sophisticated correlations between perception and reasoning. Meanwhile, to reduce the subjectivity of human-annotated prompts, automatic attribution-debiased prompting (ADP) is used for task adaptation on MLLMs. Additionally, we design image captioning (IC) to achieve visual details for enhancing cross-modal consistency checking. Extensive experiments demonstrate the superiority of our proposal and provide a new paradigm for multimodal misinformation detection in the era of AIGC.

1 Introduction

The proliferation of AI-generated content (AIGC) technology has lowered the barrier for malicious fabricators to produce multimodal misinformation (Goodfellow et al., 2014; Wang et al., 2025). This has flooded social media platforms, posing threats to societal safety (Zhang et al., 2024b; Guo et al., 2025). Traditional methods to the task of multimodal misinformation detection have focused on AI-generated benchmarks, where they leverage supervised machine learning frameworks to identify the authenticity of news captions, news images, or image-text pairs (Wang et al., 2024b; Huang et al., 2024). This paradigm has now been disrupted by the emergence of both advanced multimodal large language models (MLLMs) and mixed-source multimodal misinformation (Wang et al., 2024a; Liu et al., 2025c). Effective multimodal misinformation detection is thus crucial to the credibility of multimodal content for the general public.

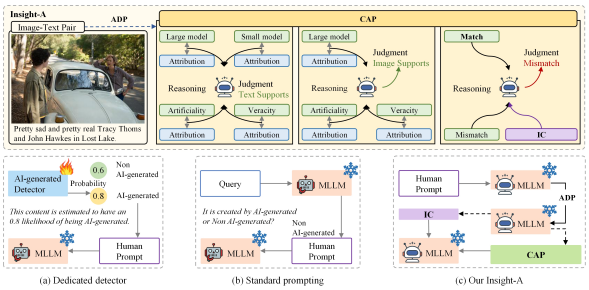

Figure 1: Misinformation detection methods (a) and (b) vs our Insight-A (c). Previous methods either train dedicated detectors, or use standard prompting to combat misinformation. In contrast, our Insight-A attributes multimodal content to generation patterns, thus making accurate decisions.

In the era of AIGC, multimodal misinformation evolves to multiple and random sources, including textual veracity distortion (TVD), visual veracity distortion (VVD), and cross-modal consistency distortion (CCD). While many studies exist for probing surface features to detect mixed-source multimodal misinformation (Beigi et al., 2024; Ali et al., 2025), research on the inherent attribution of each distortion remains under-explored. Previous work treats generation patterns as a critical signal to capture the distinction between different distortions, thereby accurately distinguishing multimodal misinformation. Firstly, as illustrated in Figure 1, a dedicated detector is used to judge the generation pattern, with its probability incorporated into a human-crafted prompt for the final judgment (Xu et al., 2025a). However, such proposal depends

primarily on large-scale datasets annotated with high-quality labels, which are time-consuming and expensive. On the other hand, a low-resource approach is to utilize standard prompting (SP) for directly querying MLLMs in a zero-shot setting (Huang et al., 2024). Nonetheless, this direct query method struggles to make correct decisions due to the lack of mimicking human-like reasoning. Hence, it is essential to explore generation patterns with effective reasoning for better understanding the characteristics of distinct distortions in the context of multimodal misinformation detection.

In this paper, our goal is to track attribution for multimodal misinformation detection. However, identifying multimodal misinformation attribution is non-trivial. Firstly, the subjectivity of human-crafted queries around multimodal misinformation attribution is inevitable to encounter language biases, such as rare words or grammatical mistakes. Secondly, the choice of the generation pattern introduces challenges in bridging the sophisticated correlations between perception and reasoning. Finally, the detailed description from news images is essential for compensation, enhancing consistency checking across modalities.

To this end, we present a novel zero-shot framework, dubbed Insight-A, exploring Attribution with MLLM Insights for effective multimodal misinformation detection. Specifically, we first design automatic attribution-debiased prompting (ADP) to eliminate language biases from human-crafted queries. Subsequently, a cross-attribution prompting (CAP) strategy achieves the reasoning paths across different generation patterns, and allocates alignment scores to each generation pattern and its reasoning path for accurate reasoning answers. Moreover, image captioning (IC) is designed to provide visual details, thus integrating clues into decision-making for debunking multimodal misinformation.

We conduct experiments to compare our Insight-A with advancing state-of-the-art (SOTA) methods on the mixed-sourced multimodal misinformation benchmark dataset MMFakeBench. Experimental results illustrate the superiority of Insight-A and further provide an in-depth analysis of different methods for future works in the era of AIGC.

In summary, our major contributions are as follows:

- We present a novel method, dubbed Insight-A, exploring attribution with MLLM insights for detecting multimodal misinformation in a zero-shot setting.
- We introduce a cross-attribution prompting method that facilitates the accurate decision-making more heavily from reasoning paths with higher credibility.
- Extensive experiments on the MMFakeBench dataset demonstrate the effectiveness and superiority of Insight-A, and our method of identifying distinct veracity distortions related to multimodal misinformation.

2 Related work

2.1 Multimodal Misinformation Detection

Most of existing works on the task of multimodal misinformation detection use deep models with supervised learning, which can be divided into the following three types: 1) some methods focus on textual veracity in human feedback (Min et al., 2022), pre-trained language models (PLMs) (Zhang and Gao, 2023), and style (Wang et al., 2024b). 2) some researchers are dedicated to visual veracity by extracting features based on sentiment (Shao et al., 2024), background (Wang et al., 2024b), and pre-trained visual models (Liu et al., 2025b). 3) multimodal misinformation strengthens the deception for the general public. Extracting multimodal features, and then fusing them captures the inconsistencies between both news captions and images. Some studies design various attention-based modules for cross-modal fusion (Ying et al., 2023; Chen et al., 2023; Xu et al., 2025b). In this paper, we rely on the reasoning and generation capabilities of MLLMs, thus achieving accurate judgments regarding multimodal misinformation.

2.2 MLLMs for Multimodal Misinformation Detection

With the rise and application of large language models (LLMs), numerous works have paved the way for enhancing the perceptual capacities of LLMs regarding visual content (Baichuan, 2023; Jiang et al., 2024; Wang et al., 2025). This promotes the development of MLLMs in multimodal and visual domains (Beigi et al., 2024; Restrepo et al., 2025; Fang et al., 2025; Liu et al., 2025a; Xu et al., 2025b). In the realm of multimodal misinformation detection, recent methods leverage diverse series of MLLMs to analyze news without post-training (Liu et al., 2025c). More importantly, in order to

identify mixed-source multimodal misinformation in the era of AIGC, a forerunner uses a hierarchical decomposition paradigm to achieve reasoning answers in a zero-shot setting (Liu et al., 2025c). Despite their impressive performance, they do not consider multimodal misinformation attribution. In contrast, our work is to explore attribution with MLLM insights on multimodal misinformation.

3 Methodology

3.1 Problem Definition

Multimodal misinformation detection is the task of identifying the authenticity of a given image-text pair. Formally, let $\mathcal{D} = \{(x_i^t, x_i^v, y_i) | (x_i^t, x_i^v) \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^N$ represent the set of image-text pairs $\mathcal{X}$ and their labels $\mathcal{Y}$. The symbols x_i^t and x_i^v denote the i-th news caption and its attached news image. N refers to the number of all image-text pairs in $\mathcal{D}$. Given a pair of image-text (x_i^t, x_i^v), the goal of a multimodal misinformation detector is to provide a prediction $y \in \{0, 1, 2, 3\}$, where $y = 0$ denotes the real item, $y = 1$ denotes the textual veracity distortion (TVD), $y = 2$ denotes the visual veracity distortion (VVD), and $y = 3$ denotes the cross-modal consistency distortion (CCD). The latter three are classified as the fake item.

The task of multimodal misinformation detection is regarded as a multiclass classification issue, where the model f learns a function: $f(\mathcal{X}) \rightarrow \mathcal{Y}$.

3.2 Attribution-Debiased Prompting

Empirical studies have shown that MLLMs exhibit the predicted consistency and instruction-following capabilities (Cai et al., 2025; Yao et al., 2025; Xu et al., 2025b). In a zero-shot setting, an effective approach is to use human-crafted prompts, for guiding MLLMs to probe multimodal misinformation detection and attribution. However, this typically includes language biases, such as rare words and grammatical mistakes, due to the subjectivity of human. This constrains MLLMs for task adaptation, thereby reducing the performance of multimodal misinformation detection (Zhang et al., 2024a).

To address this limitation, we introduce automatic attribution-debiased prompting (ADP). It leverages a MLLM to simulate human-like behaviors (Ye et al., 2025), thereby generating debiased prompts with the same semantics as manually annotated prompts for task adaptation of the MLLM. Based on the intention, we craft the query to the MLLM for natural language processing as follows:

> **Query:** ... your task is to generate a sentence, please follow the instructions below:
> 1. Eliminate language biases including rare words and grammatical mistakes in the Raw Sentence.
> 2. Only return a new prompt while ensuring that it aligns with the unchanged semantics of the Raw Sentence.
>
> **Input:** [Raw Sentence $\mathcal{S}_{ADP}$]

Specifically, the ADP directs MLLMs to perform language bias elimination and semantic alignment by analyzing the given query $\mathcal{Q}_{init}$. Formally, the generated sentence can be obtained by:

$$\mathcal{R}_{ADP} = \arg\max P(\mathcal{R} \mid \mathcal{Q}_{init}, S_{ADP}), \quad (1)$$

where $\mathcal{R}_{ADP}$ denotes the generated sentence based on the query $\mathcal{Q}_{init}$ and raw sentence S_{ADP}. From the results of Table 5, the ADP can effectively enhance the model to probe multimodal misinformation attribution for improving detection.

3.3 Cross-Attribution Prompting

After generating the query from the ADP, as shown in Figure 2, we further design cross-attribution prompting (CAP) to make rigorous and correct decisions more heavily from attribution reasoning with higher credibility. In contrast to illogical repetitive answers from MLLMs (Barez et al., 2025), the CAP integrates generation patterns with hierarchical attribution reasoning. This focuses on capturing the attribution of multimodal misinformation in a high-quality reasoning and high-confidence classification manner. In response to such ideas, the designed CAP conducts the following solution from the perspectives of generation pattern, attribution reasoning, and cross-attribution scoring.

Generation Pattern. Following the definition of multimodal misinformation in the era of AIGC (Huang et al., 2024; Liu et al., 2025c), we investigate generation categories of textual veracity distortion and visual veracity distortion. In textual veracity distortion, forgery captions are created by $\mathcal{P}^T$ = *{Largemodel, Smallmodel, Artificiality}*. In visual veracity distortion, forgery images are made from $\mathcal{P}^V$ = *{Largemodel, Artificiality}*. In our initial experiments, we have observed the fact that large models struggle to understand these specific words related to generation categories. To improve

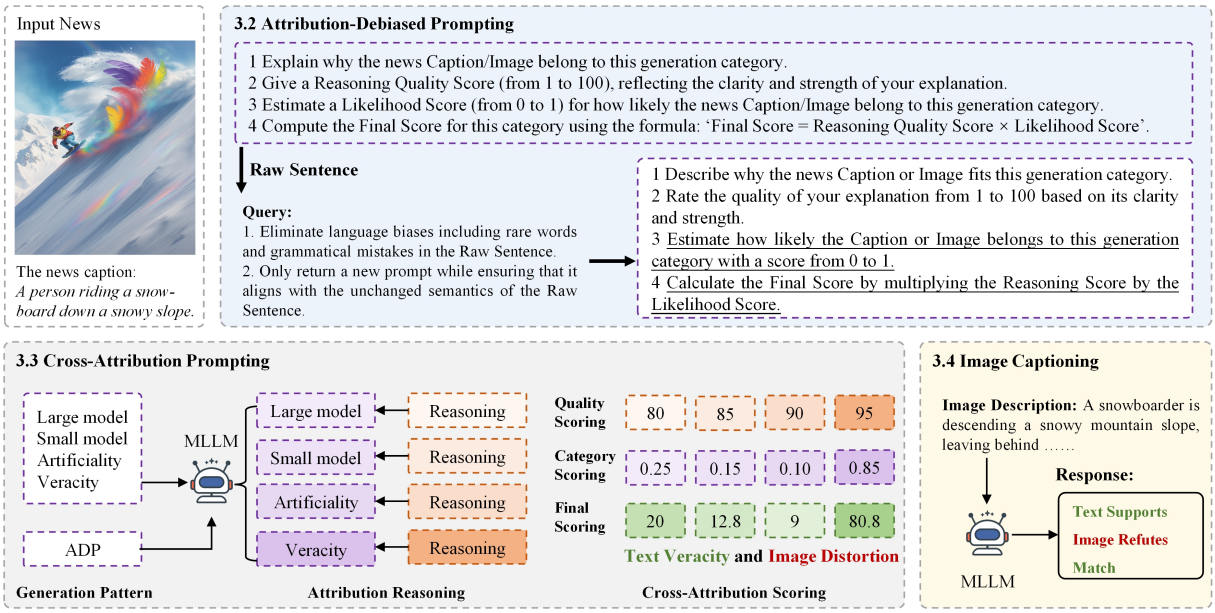

Figure 2: The overall architecture of Insight-A, which consists of automatic attribution-debiased prompting, cross-attribution prompting, and image captioning.

the MLLM's understanding, basic definitions are essential to $\{\mathcal{P}^T, \mathcal{P}^V\}$ for enhancing MLLM's reasoning beyond words, as shown in Figure 2.

Attribution Reasoning. Previous studies have shown that, although existing MLLMs have robust reasoning and generation capabilities in extensive domains, they may generate hallucinations for human preferences (Li et al., 2024; Weinzierl and Harabagiu, 2024; Jiang et al., 2024). In contrast to standard prompting for reasoning a certain generation category, we conduct fine-grained MLLM's reasoning for the news towards each potential generation category.

Specifically, depending on one generation category p_i^t from $\mathcal{P}^T$, the language of the ADP is regarded as a query $\mathcal{Q}_{ADP}$, which is introduced into the MLLM for attribution reasoning. The news caption x_i^t is also inserted into the query. Following this, the MLLM yields the response as:

$$\mathcal{R}_{p_i^t} = \arg\max P(\mathcal{R} \mid \mathcal{Q}_{ADP}, x_i^t, p_i^t), \quad (2)$$

where, $\mathcal{R}_{p_i^t}$ denotes the generated reasoning path with multiple steps in the news caption x_i^t and a certain generation category p_i^t. Furthermore, the attribution reasoning is performed in other generation categories from $\mathcal{P}^T$. Subsequently, as for the news image x_i^v, the MLLM produces the result by:

$$\mathcal{R}_{p_i^v} = \arg\max P(\mathcal{R} \mid \mathcal{Q}_{ADP}, x_i^v, p_i^v), \quad (3)$$

where, $\mathcal{R}_{p_i^v}$ denotes the yielded reasoning path with multiple steps in the news image x_i^v and one generation category p_i^v from the set $\mathcal{P}^V$.

Cross-Attribution Scoring. After performing attribution reasoning in both news captions and images, we further introduce cross-attribution scoring to assign an alignment score for the attribution of the news. Specifically, the first step is to guide the MLLM to allocate respective scores to each attribution reasoning R_p based on its quality evaluation, denoted as $s^r = \{s_1^r, s_2^r, ..., s_M^r\}$. The second step is to query the MLLM for the likelihood of each generation category $\mathcal{P}$ related to the news, denoted as $s^p = \{s_1^p, s_2^p, ..., s_M^p\}$. Eventually, we integrate these two scores for consistency across attributions. The whole process is formulated as follows:

$$\mathcal{R}_{CAP} = \arg\max_{i \in M} \; s_i^r \cdot s_i^p, \quad (4)$$

where s^r and s^p are independent during the process of cross-attribution scoring. This is beneficial to the model for comprehensive evaluation, rather than one-path reasoning.

3.4 Image Captioning

By selecting a possible generation category towards both news captions and news images, as shown in Figure 2, we incorporate these clues into the model for judging the authenticity of text or image. However, there may be semantic inconsistencies between them. To enhance cross-modal alignment, we conduct cross-modal consistency

checking within the final decision-making. Considering that the news image typically contains visual redundancy, leading to the disturbance in the sense (Wu et al., 2022; Xu et al., 2025b).

Inspired by previous work, visual signals in the form of linguistic descriptions can enable the model to effectively comprehend informative contents related to the news image (Barrett et al., 2016; Zhang et al., 2025). In response to this characteristic, we perform image captioning (IC) to give the textual content of the news image for improving cross-modal consistency checking. As shown in Figure 2, the descriptive words (e.g., "snowboarder" and "snowy mountain"), which correspond well with the news. Finally, we incorporate this description alongside the news into the MLLM for facilitating informed decision-making:

$$\mathcal{R}_{end} = \arg\max P(\mathcal{R} \mid \mathcal{Q}_{end}, x_i^t, x_i^v, x_i^{v'}), \quad (5)$$

where, $\mathcal{Q}_{end}$ and $\mathcal{R}_{end}$ denote the query and the final answer, and $x_i^{v'}$ represents the textual content corresponding to the news image.

4 Experiments

4.1 Experimental Setups

Datasets To evaluate the efficiency of our Insight-A, we conduct extensive experiments on the widely used benchmark dataset MMFakeBench (Liu et al., 2025c). This dataset comprises 11,000 image-text pairs, which intertwine mixed-source and multiple-type multimodal content. Moreover, it is split into validation and test sets with a ratio of 1:10.

Evaluation Metrics To ensure fairness, we are consistent with previous studies (Chen et al., 2023; Beigi et al., 2024; Liu et al., 2025c), employing the following evaluation metrics: F1, Pre, Rec, and ACC. Expressly, F1 is the widely adopted macro-F1 score, balancing precision and recall through their harmonic mean. Likewise, Pre, Rec, and ACC represent precision, recall, and accuracy as supplementary evaluation metrics.

Comparison Methods To demonstrate the effectiveness and generalizability of Insight-A against SOTA methods in the setting of zero-shot, we select the following representative MLLMs as foundational models for comprehensive comparison: VILA (Lin et al., 2024), InstructBLIP (Dai et al., 2023), BLIP-2 (Li et al., 2023), and LLaVA-1.6 (Liu et al., 2024b). These large models range from 13B parameters to 34B parameters. In MLLMs with 13B parameters, VILA, InstructBLIP, and BLIP-2 include LLaMA2 (Touvron et al., 2023), Vicuna (Chiang et al., 2023), and FlanT5-XXL (Chung et al., 2024). In addition, LLaVA-1.6 contains two types: Vicuna with 13B parameters, and Vicuna with 34B parameters.

Moreover, all experimental results are obtained by utilizing one NVIDIA A800 GPU (80GB) with PyTorch (Paszke et al., 2019). For consistency, we align hyper-parameters and setups with the SOTA MMD-Agent (Liu et al., 2025c).

5 Overall Results

In this part, we apply our Insight-A to the two baseline models, and illustrate the effectiveness and generalizability of Insight-A in the following settings: 1) multiclass evaluation, 2) binary evaluation, and 3) fine-grained evaluation.

Multiclass Evaluation. In Table 1, the experimental results emphasize the effectiveness and generalizability of Insight-A for multiclass classification in the realm of multimodal misinformation detection. Applied to LLaVA-1.6$_{\text{Vicuna-34B}}$, Insight-A substantially boosts performance across all evaluation metrics, achieving the best results against the MMD-Agent. Typically, in the model LLaVA-1.6$_{\text{Vicuna-13B}}$, Insight-A achieves 7.5% and 6.7% improvements over our baseline method in F1 score on the validation and test sets, respectively. In the model LLaVA-1.6$_{\text{Vicuna-34B}}$, Insight-A still attains significant increases, underscoring our method's effectiveness. Moreover, Insight-A presents a consistent tendency on both validation and test sets, highlighting the Insight-A's strength in the generalization capabilities. In a nutshell, this strong performance across all evaluation metrics confirms the efficacy of Insight-A and its potential to advance multimodal misinformation detection capabilities.

Binary Evaluation. From Table 2, we can observe that our Insight-A achieves the highest F1 score, and outperforms other compared methods regarding binary overall results. In the 3-th row and the 6-th row of Table 2, we provide the results when removing both MMD-Agent and Insight-A. The comparison between them shows that our Insight-A surpasses other methods by a large margin across different foundational models. Moreover, one can see that the MLLM with more parameters indeed introduces significant improvements in overall performance, which is consistent with the scaling law

Method	Model	Validation (1000)				Test (10000)			
		F1↑	Pre↑	Rec↑	ACC↑	F1↑	Pre↑	Rec↑	ACC↑
MMD-Agent	$\text{VILA}_{\text{LLaMA2-13B}}$	22.7	27.3	24.4	28.7	24.0	30.4	25.5	29.4
	$\text{InstructBLIP}_{\text{Vicuna-13B}}$	26.0	33.3	30.1	29.5	24.5	32.1	28.8	27.3
	$\text{BLIP2}_{\text{FlanT5-XXL}}$	31.6	39.8	32.2	34.4	28.8	39.0	30.4	32.1
	$\text{LLaVA-1.6}_{\text{Vicuna-13B}}$	38.0	44.5	41.0	40.6	34.5	42.7	37.5	37.4
	$\text{LLaVA-1.6}_{\text{Vicuna-34B}}$	49.9	54.4	52.9	48.7	47.7	52.1	49.6	46.6
$\text{Insight}_{\text{Baseline}}$	$\text{LLaVA-1.6}_{\text{Vicuna-13B}}$	34.6	43.1	38.9	40.1	34.5	43.6	38.2	39.5
	$\text{LLaVA-1.6}_{\text{Vicuna-34B}}$	53.7	53.7	56.4	54.9	53.2	52.3	55.5	55.0
Insight-A	$\text{LLaVA-1.6}_{\text{Vicuna-13B}}$	42.1	50.9	50.0	43.0	41.2	51.0	48.5	41.9
	Imp.(%)	*+7.5*	*+7.8*	*+11.1*	*+2.9*	*+6.7*	*+7.4*	*+10.3*	*+2.4*
	$\text{LLaVA-1.6}_{\text{Vicuna-34B}}$	**59.1**	**60.9**	**59.2**	**57.0**	**58.6**	**62.0**	**57.6**	**56.1**
	Imp.(%)	*+5.4*	*+7.2*	*+2.8*	*+2.1*	*+5.4*	*+9.7*	*+2.1*	*+1.1*

Table 1: Multiclass results (%) of different detection methods with different MLLMs on the benchmark dataset MMFakeBench. Best results are in bold.

Model	LLM	Size	MMD-Agent	Insight-A	Validation	Test
LLaVA-1.6	Vicuna	13B	✓	✗	51.8	50.2
LLaVA-1.6	Vicuna	13B	✗	✓	**56.1**	**57.7**
LLaVA-1.6	Vicuna	13B	✗	✗	50.6	52.5
LLaVA-1.6	Vicuna	34B	✓	✗	67.2	68.1
LLaVA-1.6	Vicuna	34B	✗	✓	**68.3**	**72.5**
LLaVA-1.6	Vicuna	34B	✗	✗	67.6	70.2

Table 2: Binary overall results (F1 score (%)) of different detection methods under MLLMs on both validation and test sets of MMFakeBench. Best results are in bold.

Method	Model	Real↑	TVD↑	VVD↑	CCD↑	Overall↑
MMD-Agent	$\text{VILA}_{\text{LLaMA2-13B}}$	32.4	13.4	4.3	37.6	21.9
	$\text{InstructBLIP}_{\text{Vicuna-13B}}$	41.9	18.8	19.6	23.8	26.0
	$\text{BLIP2}_{\text{FlanT5-XXL}}$	41.5	39.2	13.1	32.6	31.6
	$\text{LLaVA-1.6}_{\text{Vicuna-13B}}$	17.9	51.5	30.9	42.3	35.7
	$\text{LLaVA-1.6}_{\text{Vicuna-34B}}$	51.1	37.6	61.7	49.2	49.9
$\text{Insight}_{\text{Baseline}}$	$\text{LLaVA-1.6}_{\text{Vicuna-13B}}$	19.4	58.2	20.6	40.3	34.6
	$\text{LLaVA-1.6}_{\text{Vicuna-34B}}$	55.1	**65.3**	47.4	47.0	53.7
Insight-A	$\text{LLaVA-1.6}_{\text{Vicuna-13B}}$	42.4	45.4	36.4	44.0	42.1
	$\text{LLaVA-1.6}_{\text{Vicuna-34B}}$	**57.1**	58.9	**68.6**	**51.9**	**59.1**

Table 3: Performance (F1 score (%)) of different methods on different forgery sources of multimodal misinformation. Best results are in bold.

Category	Multiclass Cls.				Binary Cls.			
Method	F1↑	Pre↑	Rec↑	ACC↑	F1↑	Pre↑	Rec↑	ACC↑
$\text{GPT-4V}_{\text{ChatGPT}}$	51.0	66.8	49.7	54.0	72.3	72.1	72.8	75.6
+ MMD-Agent	61.6	**67.8**	59.3	62.1	74.0	73.4	75.5	76.8
+ $\text{Insight}_{\text{Baseline}}$	57.3	58.3	56.8	65.6	81.6	82.1	81.3	83.3
+ **Insight-A**	**61.8**	61.5	**62.3**	**67.5**	**82.2**	**82.3**	**82.1**	**83.7**

Table 4: Performance (%) comparison of different methods with the proprietary MLLM on both binary and multiclass classification (**Cls.**). Best results are in bold.

(Kaplan et al., 2020). Along with the $\text{LLaVA-1.6}_{\text{Vicuna-34B}}$, additional improvements of 1.1% and 4.4% are observed, showing the value of Insight-A even in larger models. Consistent results confirm the Insight-A's distinguishing capabilities.

Fine-Grained Evaluation. In Table 3, we compare Insight-A with other methods across different forgery sources, including TVD, VVD, and CCD. From the results, we can conclude that the majority of open-source models perform well in certain scenarios, but experience performance degradation in others. However, our Insight-A effectively reduces the performance discrepancies between different scenarios by a huge margin. Compared to the MMD-Agent with $\text{LLaVA-1.6}_{\text{Vicuna-34B}}$ across different forgery scenarios, our proposed method provides gains of 6%, 21.3%, 6.9%, and 2.7%, respectively. The results across different forgery sources effectively illustrate Insight-A's robustness in complex environments.

In Table 4, we also provide a detailed comparison between Insight-A and MMD-Agent within the proprietary model (e.g., GPT-4V (OpenAI, 2023)). Insight-A achieves the best performance on both multiclass and binary classification, reaching the new SOTA. Typically, although GPT-4V with standard prompting reaches a relatively high performance (see 1-th row in Table 4), the lack of inherent reasoning prevents it from debunking erroneous information in the sense. Each method can enhance the model to understand and capture synthetic multimodal misinformation (see 2-th row to 4-th row in Table 4), thus achieving a gain in performance. Furthermore, MMD-Agent with GPT-4V presents a high precision and a low recall, which indicates that it tends to make decisions for misinformation with identifiable points of falsification. However, Insight-A presents a good balance between precision and recall, highlighting its strength in debunking misinformation with both ambiguous contents and identifiable points of falsification.

To sum up, the experiments above confirm the

Category	Model	Real↑	TVD↑	VVD↑	CCD↑	Overall↑
Human		56.7	58.6	63.5	51.3	57.5
Language	LLaMA-3	43.5	41.5	56.8	44.5	46.6
	Qwen-3	49.2	58.4	58.4	44.7	52.7
	DeepSeek-V3	54.8	44.8	67.0	51.7	54.6
Multimodal	GPT-4o	55.4	57.9	67.3	50.5	57.8
	Gemma-3	55.9	57.4	68.3	51.5	58.3
	LLaVA-1.6	**57.1**	**58.9**	**68.6**	**51.9**	**59.1**

Table 5: Performance (F1 score (%)) on different large models for the ADP. Best results are in bold.

Model	+ CAP	+ IC	F1↑	Pre↑	Rec↑	ACC↑
LLaVA-1.6	✗	✗	53.7	53.7	56.4	54.9
	✗	✓	54.4 (↑0.7)	53.6 (↓0.1)	57.1 (↑0.7)	55.7 (↑0.8)
	✓	✗	58.7 (↑5.0)	60.5 (↑6.8)	58.7 (↑2.3)	56.4 (↑1.5)
	✓	✓	59.1 (↑5.4)	60.9 (↑7.2)	59.2 (↑2.8)	57.0 (↑2.1)

Table 6: Results (%) of ablations for the effect of CAP and IC on multiclass classification.

Category	Multiclass Cls.				Binary Cls.	Attribution
Model	F1↑	Pre↑	Rec↑	ACC↑	F1↑	F1↑
	53.7	53.7	56.4	54.9	67.6	42.1
Insight-A with predicted attributions						
LLaVA-1.6	59.1	60.9	59.2	57.0	68.3	52.8
Insight-A with ground truth attributions						
	68.6	67.3	71.2	67.9	74.0	-

Table 7: Performance (%) comparison on both classification (**Cls.**) and attribution.

Model	LLM	Size	+ ARS	+ PPS	F1↑
LLaVA-1.6	Vicuna	34B	✗	✗	54.4
LLaVA-1.6	Vicuna	34B	✓	✗	57.3 (↑2.9)
LLaVA-1.6	Vicuna	34B	✗	✓	57.8 (↑3.4)
LLaVA-1.6	Vicuna	34B	✓	✓	59.1 (↑4.7)

Table 8: Results (%) of ablations for cross-attribution scoring in Insight-A. ARS denotes attribution reasoning scoring, and PPS denotes generation pattern scoring.

effectiveness and generalizability of our Insight-A to the task of multimodal misinformation detection in a zero-shot setting.

6 Ablation Study

Impact of Models on ADP. In Table 5, we present the results of ADP by adopting different large models. LLMs include LLaMA-3 (AI@Meta, 2024), Qwen-3 (Yang et al., 2025), and DeepSeek-V3 (Liu et al., 2024a). MLLMs include GPT-4o (Hurst et al., 2024), Gemma-3 (Team, 2025), and LLaVA-1.6 (Liu et al., 2024b). From the results, we can achieve the following findings: The expert with domain knowledge performs the task adaptation prompting, resulting in accurate detections. The comparison between LLMs and MLLMs shows that MLLMs typically possess superior performance, which is attributed to their fine-tuning across multimodal tasks and datasets. On the other hand, the ADP on LLaVA-1.6 is consistent with the foundational model in Insight-A, thus leading to a significant gain in performance. Hence, leveraging the inherent capability of large models to the ADP can effectively enhance the detection performance of Insight-A. This promotes us to explore multimodal misinformation attribution from the perspective of MLLMs insight.

CAP and IC. As shown in Table 6, we further evaluate the effectiveness of CAP and IC in Insight-A with LLaVA-1.6$_{\text{Vicuna-34B}}$. In 1-th row, we remove both CAP and IC components to achieve the results of our baseline method. The 2-th row and 3-th row refer to CAP and IC, respectively. Contrasting 1-th row with 2-th row verifies that deriving visual details via IC can incorporate informative cues into the final decision-making, thereby yielding performance gains. The comparison between 1-th row and 3-th row reveals that utilizing the CAP in our baseline method provides a large improvement in performance. In 4-th row, working with all modules further enhances the detection capabilities. The results clearly demonstrate the significance of both CAP and IC in our Insight-A.

Effect of Attribution. To investigate the significance of multimodal misinformation attribution, we conduct extensive experiments as follows:

- In Table 7, we report the results on classification and attribution. Here, we leverage LLaVA-1.6 with 34B parameters as the foundational model. The F1 score is used as an evaluation metric for estimating attribution performance. The 1-th row denotes the results of our baseline method. Integrating attribution into our Insight-A significantly improves its performance (see 2-th row and 3-th row in Table 7). Furthermore, we illustrate that our Insight-A results using 70% of the ground truth attribution lead to the best performance on both multiclass and binary classification. This is expected because the ground truth attribution is achieved by decomposing the essence of misinformation production. The results confirm that predicting accurate attributions can ensure the high performance for multimodal misinformation detection.

- As shown in Figure 3, we exhibit the detection results of different methods on two preva-

Metric	MMD-Agent	$\text{Insight}_{\text{Baseline}}$	Insight-A
Average Inference Time (s)	49.04s	45.14s	50.13s
GPU Memory Usage (GB)	82G	72G	72G
Performance (F1)	49.9	53.7	59.1

Table 9: Efficiency comparison of different methods on the LLaVA-1.6 model with 34B parameters.

Figure 3: Performance (detection success rate) of different methods on two generation categories.

lent generation categories. In text, the clear definition of AI-generated contents enables models to capture their boundary. Non AI-generated texts include compound generation categories, leading to a more substantial challenge. In image, synthetic images are created on real-life scenarios and AIGC technologies, limiting the perceptual capabilities of models for detecting them. Another type of image is edited by humans, which typically contains explicit forgery traces. This facilitates models to identify them. In both texts and images, our Insight-A leverages the inherent capabilities of MLLMs for reasoning answers, resulting in significant gains in performance. This underscores Insight-A's superiority in the task of multimodal misinformation detection.

- Finally, in Table 8, we perform experiments for the effect of cross-attribution scoring in the CAP. From the results, we can note that each scoring way guides the model to catch forgery traces within multimodal contents, resulting in performance gains of 2.9% and 3.4%, respectively. This further confirms the significance of multimodal misinformation attribution. On the other hand, the mutual synergy between scoring ways (see 4-th row in Table 8) brings a 4.7% improvement, highlighting the strength of multiple reasoning paths in tasks demanding both perception and reasoning.

In summary, each designed module in Insight-A can promote it to effectively and accurately catch forgery sources and attributions, thereby leading to superior performance in the realm of multimodal misinformation detection.

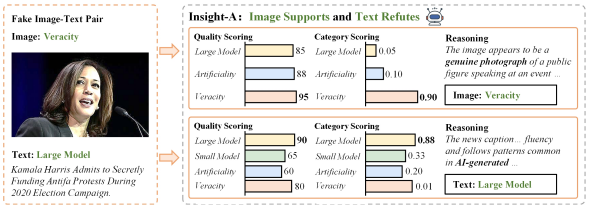

Figure 4: The qualitative results of Insight-A.

7 Further Analysis

Analysis on Efficiency. In Table 9, we compare the inference time and GPU memory usage of different methods for measuring the efficiency of our Insight-A. From the results, we can observe that while Insight-A shows a slight increase in the inference time due to multiple attribution reasoning paths, its performance surpasses the SOTA by a significant margin. In addition, our Insight-A utilizes fewer computational resources, and provides higher performance. This confirms the efficiency of Insight-A in the deployment stage.

Analysis on Cases. As shown in Figure 4, we present a challenging instance for analyzing the ability of Insight-A. Specifically, in text, Insight-A provides strong confidence levels of both perception and reasoning for a specific generation category, ensuring the news caption attribution. In image, Insight-A also provides high confidence levels for each attribution reasoning. However, the category scoring presents significant discrepancies between different generation categories. Therefore, such independent cross-attribution scoring effectively guides Insight-A to make correct decisions.

8 Conclusion

This paper introduces Insight-A, a novel framework that explores attribution with MLLM insights for debunking misinformation in the era of AIGC. We emphasize the significance of multimodal misinformation attribution using multiple reasoning paths in Insight-A, underscoring it as a pivotal factor for detecting veracity distortion. Experiments across various evaluation settings demonstrate the effectiveness and generalizability of Insight-A, paving the way for combating deceptive misinformation created by AIGC technology nowadays.

Limitations

Despite the effectiveness and generalizability of our proposed Insight-A, there are some limitations. Firstly, in this paper, we do not provide experimental results with other foundational models, such as series of Qwen and InternVL. This is because that LLaVA can be consistent with previous work, and this model has achieved satisfactory performance. Second, we evaluate the model on a mixed-source multimodal misinformation detection benchmark dataset. This is because that this dataset includes mixed-source and multiple-type multimodal content. In future work, we aim to expand Insight-A with more foundational models on more benchmark datasets for multimodal misinformation detection in the era of AIGC.

References

AI@Meta. 2024. Llama 3 model card.

Muhammad Zain Ali, Yuxia Wang, Bernhard Pfahringer, and Tony C Smith. 2025. Detection of human and machine-authored fake news in Urdu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3419–3428, Vienna, Austria. Association for Computational Linguistics.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. 2025. Chain-of-thought is not explainability.

Daniel Paul Barrett, Andrei Barbu, N. Siddharth, and Jeffrey Mark Siskind. 2016. Saying what you're looking for: Linguistics meets video search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2069–2081.

Alimohammad Beigi, Bohan Jiang, Dawei Li, Zhen Tan, Pouya Shaeri, Tharindu Kumarage, Amrita Bhattacharjee, and Huan Liu. 2024. Can llms improve multimodal fact-checking by asking relevant questions? *arXiv preprint arXiv:2410.04616*.

Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. 2025. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 18434–18443.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. 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, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267. Curran Associates, Inc.

Wenlong Fang, Qiaofeng Wu, Jing Chen, and Yun Xue. 2025. Notes-guided mllm reasoning: Enhancing mllm with knowledge and visual notes for visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19597–19607.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Hao Guo, Zihan Ma, Zhi Zeng, Minnan Luo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2025. Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):228–236.

Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. MiRAGeNews: Multimodal realistic AI-generated news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16436–16448, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1

others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27036–27046.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Dongfang Li, Zetian Sun, Baotian Hu, Zhenyu Liu, Xinshuo Hu, Xuebo Liu, and Min Zhang. 2024. Improving attributed text generation of large language models via preference learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5079–5101, Bangkok, Thailand. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Jiapeng Liu, Liang Li, Shihao Rao, Xiyan Gao, Weixin Guan, Bing Li, and Can Ma. 2025a. Union is strength! unite the power of llms and mllms for chart question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):5487–5495.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2025b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pages 38–55, Cham. Springer Nature Switzerland.

Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2025c. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *ICLR*.

Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 1148–1158, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023. Gpt-4v(ision) system card.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

David Restrepo, Chenwei Wu, Zhengxu Tang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Cong-Tinh Dao, Jack Gallifant, Robyn Gayle Dychiao, Jose Carlo Artiaga, André Hiroshi Bando, Carolina Pelegrini Barbosa Gracitelli, Vincenz Ferrer, Leo Anthony Celi, Danielle Bitterman, Michael G Morley, and Luis Filipe Nakayama. 2025. Multi-ophthalingua: A multilingual benchmark for assessing and debiasing llm ophthalmological qa in lmics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28321–28330.

Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. 2024. Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5556–5574.

Gemma Team. 2025. Gemma 3.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jin Wang, Chenghui Lv, Xian Li, Shichao Dong, Huadong Li, Kelu Yao, Chao Li, Wenqi Shao, and Ping Luo. 2025. Forensics-bench: A comprehensive forgery detection benchmark suite for large vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 4233–4245.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a.

Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shengkang Wang, Hongzhan Lin, Ziyang Luo, Zhen Ye, Guang Chen, and Jing Ma. 2024b. Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. *arXiv preprint arXiv:2406.11288*.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880, Bangkok, Thailand. Association for Computational Linguistics.

Xing Wu, Chaochen Gao, Zijia Lin, Zhongyuan Wang, Jizhong Han, and Songlin Hu. 2022. RaP: Redundancy-aware video-language pre-training for text-video retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3036–3047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qingzheng Xu, Heming Du, Szymon Łukasik, Tianqing Zhu, Sen Wang, and Xin Yu. 2025a. Mdam3: A misinformation detection and analysis framework for multitype multimodal media. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 5285–5296, New York, NY, USA. Association for Computing Machinery.

Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Ming Li, Wenxin Liang, Yang Li, and Sidan Du. 2025b. Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In *AAAI*, pages 8978–8986.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jixun Yao, Yang Yuguang, Yu Pan, Ziqian Ning, Jianhao Ye, Hongbin Zhou, and Lei Xie. 2025. Stablevc: Style controllable zero-shot voice conversion with conditional flow matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25669–25677.

Haoran Ye, Tianze Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. 2025. Generative psycho-lexical approach for constructing value systems in large language models. *arXiv preprint arXiv:2502.02444*.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5384–5392.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. 2024a. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12924–12933.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. 2024b. Reinforced adaptive knowledge learning for multimodal fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16777–16785.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Zhenxing Zhang, Yaxiong Wang, Lechao Cheng, Zhun Zhong, Dan Guo, and Meng Wang. 2025. Asap: Advancing semantic alignment promotes multi-modal manipulation detecting and grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 4005–4014.