PrismAgent: Illuminating Harm in Memes via a Zero-Shot Interpretable Multi-Agent Framework

Anonymous ACL submission

Abstract

The rapid spread of memes makes harmful content detection increasingly crucial, as effective identification can curb the circulation of misinformation. However, existing methods rely heavily on high-volume annotated data, which leads to substantial training costs and limited generalization. To address these challenges, we propose **PrismAgent**, a zero-shot, multi-agent, interpretable framework. PrismAgent conceptualizes this task as a criminal case investigation, employing four specialized agents responsible for the analysis, investigation, prosecution, and judgment stages within a structured collaborative workflow. In the first stage, the analyst agent paraphrases each meme under benevolent and malicious assumptions to probe its underlying intent. The investigator agent then retrieves supporting evidence from an unannotated dataset and constructs contextual interpretations for the meme and its variants. Next, the prosecutor agent performs three independent preliminary judgments by pairing the original meme with each of the three interpretations. Finally, the judge agent deliberates across all evidence to render a final verdict. Moreover, PrismAgent's explicit multi-stage reasoning chain makes the model inherently interpretable, since every intermediate step is explicitly explained rather than only producing a final detection result. Extensive experiments on three public datasets show that PrismAgent significantly outperforms existing zero-shot detection methods.

1 Introduction

With the rapid development of social media, memes have become a pervasive form of multimodal communication. While memes were initially created for humor expression, they are increasingly being misused to spread misinformation and harmful biases. This emerging trend poses serious risks to social media, highlighting the necessity of detecting harmful memes (Sharma et al., 2022).

Traditional harmful meme detection methods primarily rely on large-scale and carefully annotated dataset (Yang et al., 2024b; Lin et al., 2024). However, these methods tend to overfit to the training distribution, which compromises their robustness when applied to unseen or evolving memes. Given the rapidly changing nature of meme content, such approaches face substantial challenges when deployed in real-world scenarios (Cao et al., 2024; Huang et al., 2024).

Another challenge lies in the inherently implicit nature of memes. Their meaning is often conveyed through shared cultural knowledge, visual symbolism, or multimodal irony rather than explicit textual content. Such indirect expressions make it challenging for models to capture the underlying intent, as understanding memes typically requires commonsense reasoning and cultural context that extend beyond literal interpretation.

To address these challenges, we propose **PrismAgent**, a zero-shot, multi-agent, interpretable framework for harmful meme detection. Much like a prism that refracts a single beam of light into multiple revealing facets, PrismAgent decomposes each meme into diverse motivational and contextual perspectives, which makes its underlying meaning clearer and its judgment more interpretable. To achieve this, PrismAgent simulates the workflow of a real criminal case investigation, where different stages of the process are handled by distinct specialized roles. The investigation begins with analysts who infer potential motives and intentions behind a meme. Investigators then collect relevant contextual ev-

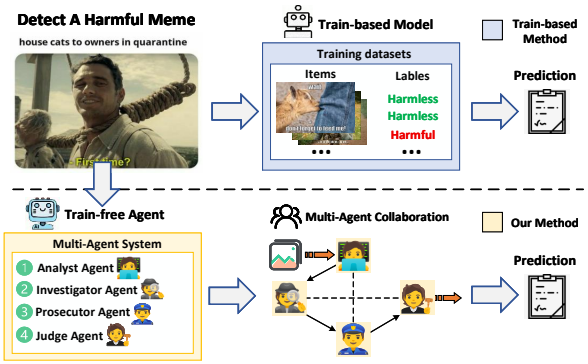

Figure 1: Pipeline comparison between training-based methods and our proposed approach.

idence to reason about the case and establish preliminary interpretations. Finally, a judge evaluates both the collected evidence and the investigators reasoning to deliver a fair and well-grounded verdict.

PrismAgent instantiates each stage of the investigation process as a dedicated agent, including the analyst, investigator, prosecutor, and judge. The analyst agent paraphrases each meme under two opposing disseminator intentions, benevolent and malicious, to reveal the range of plausible potential intent. Next, the investigator agent searches the unannotated dataset for evidence relevant to the original meme and its two rewritten variants, which are then used to construct contextual interpretations. The prosecutor agent then reasons over the original meme and each contextualized variant, producing three independent judgments. If these judgments are consistent, the prosecutor finalizes the decision. Otherwise, the case is forwarded to the judge agent, who integrates all available information and deliberates to render a coherent and interpretable final verdict.

Unlike previous methods that rely on a single-step prediction, we decompose harmful meme detection into a series of structured subtasks. This design induces an explicit reasoning chain and substantially improves the interpretability of the detection process. Moreover, by adopting a reasoning-based pipeline rather than training a task-specific classifier, PrismAgent eliminates the dependence on annotated data and naturally enables effective zero-shot detection. A comparison between previous methods and ours is illustrated in Fig. 1. Our main contributions are summarized as follows:

- We propose PrismAgent, a multi-agent framework that decomposes harmful meme detection into structured subtasks, which enables modular role-specific reasoning.
- We design a multi-stage reasoning chain that integrates the outputs of several specialized agents to enhance interpretability via explicit, stage-wise explanations.
- Extensive experiments on public meme datasets demonstrate the effectiveness of our method and show that it performs competitively against state-of-the-art approaches.

2 Related Works

Harmful Meme Detection The multimodal nature of memes necessitates multimodal strategies (Borakati, 2021; Beyer and Alexy, 2025), which consistently outperform unimodal methods in harmful meme detection (He et al., 2016; Devlin et al., 2019). Previous studies primarily follow two methodological directions. The first relies on classical two-stream architectures that perform harmful meme classification by fusing textual and visual modalities (Singh et al., 2022; Suryawanshi et al., 2020; Pramanick et al., 2021b; Kumar and Nandakumar, 2022). The second direction fine-tunes large pre-trained multimodal models to adapt them to domain-specific detection tasks (Velioglu and Rose, 2020; Hee et al., 2022, 2025; Hee and Lee, 2025). In addition, relevant studies have proposed various improved strategies (Cao et al., 2022; Ji et al., 2023; Cao et al., 2023; Garg et al., 2025; Kumari et al., 2025), as well as detection methods for few-shot annotation (Cao et al., 2024; Huang et al., 2024) and zero-shot annotation (Liu et al., 2025) scenarios to reduce reliance on labeled data.

VLM-based multi-agent framework When Vision Language models (VLMs) are deployed as agents across various domains, they demonstrate strong planning and reasoning capabilities in a wide range of scenarios (Yao et al., 2022; Sun et al., 2023; Bang et al., 2024; Zhao et al., 2024; Yang et al., 2024a; Chen et al., 2024; Gao et al., 2025). These advancements show that LLM-based methods can effectively handle complex tasks even

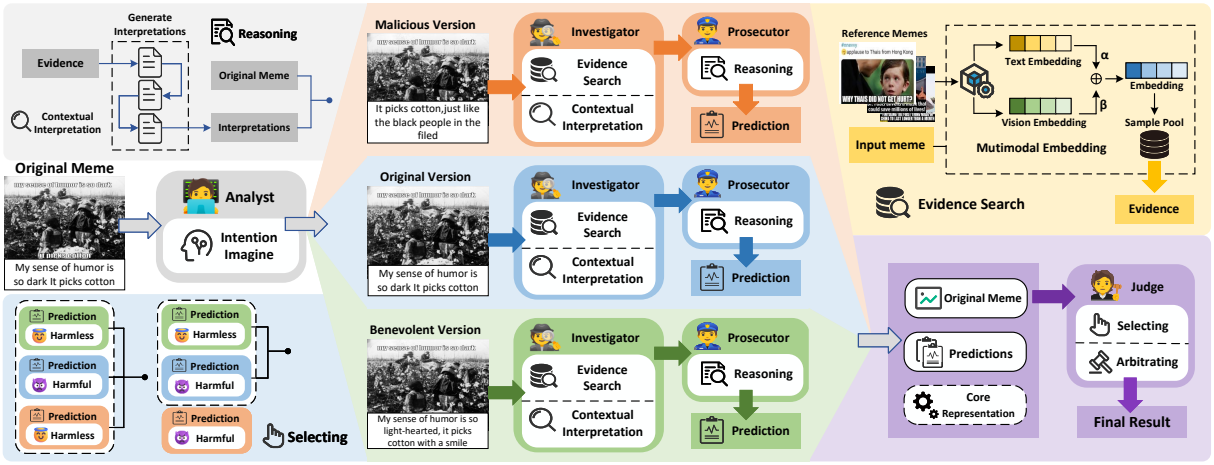

Figure 2: Overview of the proposed zero-shot, multi-agent, interpretable framework for harmful meme detection, PrismAgent.

under weakly supervised conditions. Building on the success of single-agent systems, multi-agent frameworks (Qian et al., 2024; Tao et al., 2024; Hong et al., 2024; Ma et al., 2024; Zahedifar et al., 2025; Yang et al., 2025) further enable interactive collaboration and collective problem-solving. However, research on multi-agent frameworks for harmful meme detection remains limited. Although Liu et al. (2025) proposed the MIND framework, it still fails to fully address the inherent challenge of interpreting meme metaphors under zero-shot settings.

3 Method

3.1 Overview

Detecting harmful memes presents a significant challenge because their meaning often relies on subtle cultural references, visual symbols, or humorous wordplay, rather than explicit textual or visual cues. Additionally, most existing methods depend on large-scale annotated training datasets, which often cause detectors to overfit to the training memes. As memes continually evolve, these methods struggle with poor generalization.

To alleviate these challenges, we propose PrismAgent, a zero-shot, multi-agent, interpretable framework for harmful meme detection. An overview of the framework is depicted in Figure 2. PrismAgent introduces an explicit reasoning chain and assigns distinct roles to four agents, including analyst, investigator, prosecutor, and judge. These agents collaborate to reveal the intentions behind memes. Specifically, the analyst agent imagines two opposing intentions for the meme's disseminator and paraphrases the meme accordingly. The investigator agent then searches an unannotated dataset for relevant supporting evidence for the original meme and its two rewrites, which construct contextual interpretations for each version. The prosecutor agent produces three independent judgments by reasoning over the original meme together with each of the contextual interpretations. If these judgments conflict, the judge agent integrates all evidence and delivers a final, robust, and interpretable decision.

PrismAgent requires no training and operates as a zero-shot detection framework. By leveraging a structured multi-agent reasoning process, it renders decisions that are highly interpretable and transparent.

3.2 Analyst Agent

Revealing the hidden intent of memes is a crucial yet challenge step in meme understanding. To make such implicit intentions more accessible, we introduce the analyst agent to explicitly amplify implicit intentions by assuming two opposing disseminator intentions, benevolent and malicious, to reveal the real intent.

Specifically, for each meme version, we first input it into the agent, and the agent is guided by the carefully designed prompts to imagine possible dissemination motives and paraphrase two corresponding rewrites: one reflecting benevolent and the other reflecting malicious intensions. In this process, the original linguistic style and contextual tone are preserved as much as possible, ensuring that the

rewritten memes remain semantically coherent while exposing their potential intentions. Consequently, the procedure produces two rewritten versions that make the original memes latent intentions more explicit from two opposing motives. This process can be expressed as:

$$M_b = Agent_{Ana}(M_{ori}, P_b), \tag{1}$$

$$M_m = Agent_{Ana}(M_{ori}, P_m), \tag{2}$$

where $M_{ori} = \{\mathcal{V}, \mathcal{T}\}$ denotes the original meme. $M_b = \{\mathcal{V}, \mathcal{T}_b\}$ and $M_m = \{\mathcal{V}, \mathcal{T}_m\}$ represent the benevolent and malicious memes obtained by our analyst agent, respectively. $\mathcal{V}$ and $\mathcal{T}$ denote the visual content and the accompanying text, respectively. $\mathcal{T}_b$ and $\mathcal{T}_m$ represent the benevolent and malicious rewrites, respectively. Besides, P_b and P_m denote the prompts used to rewrite the meme from benevolent and malicious intentions, respectively.

Finally, a single original meme M_{ori} is expanded into a set $\{M_{ori}, M_b, M_m\}$, which contains the original sample and its two rewrites. This process makes the latent intent more salient by forcing the meme to be expressed under contrasting motivational assumptions, thereby amplifying the subtle cues embedded in the original content. The detailed prompts are provided in Appendix A.3.

3.3 Investigator Agent

After the analyst agent paraphrases each meme, the investigator agent then searches the unannotated dataset for supporting evidence to the original meme and its two rewrites. This step provides the subsequent investigation stage with additional context and precedents, enabling it to gather supporting evidence that further explores the potential intentions behind the original meme and its rewrites. Moreover, this process offers a potential advantage: since memes are constantly evolving, the investigator agent can continuously expand the unannotated dataset in the deployment stage to incorporate emerging contexts. This allows our framework to adapt to emerging meme trends in real time, enhancing its robustness and generalization in the zero-shot scenario. The evidence search process can be formulated as:

$$\mathcal{Q}_{\text{evi}} = f(D_{ref}, M_{ori}) \tag{3}$$

where $\mathcal{Q}_{\text{evi}} = \{M^1_{\text{evi}}, ..., M^k_{\text{evi}}\}$ denotes the set of top-k relevant memes retrieved from the unannotated reference datase D_{ref}. M^k_{evi} denotes most k-th most similar meme retrieved from D_{ref}. $f(\cdot)$ represents the matching function that measures the similarity between the original meme and each sample in D_{ref} and returns the top-k closest memes. k is empirically set to 3 in this paper.

After retrieving the relevant cases, the investigator agent interprets them in descending order of similarity, from the most similar to the least similar. At each step, the current meme is interpreted together with the interpretations accumulated from previous steps, enabling the agent to integrate newly evidence with prior observations. This sequential accumulation of information helps the investigator agent to capture subtle contextual cues and latent intentions that may not be apparent from a single meme. This process can be formulated as follows:

$$\mathcal{O}_i = Agent_{Inv}(M^i_{\text{evi}}, \mathcal{O}_{i-1}, P_{int}), \tag{4}$$

where $\mathcal{O}_i$ and $\mathcal{O}_{i-1}$ represent the interpretations generated in the i-th and (i-1)-th steps, respectively; M^i_{evi} denotes the i-th most similar evidence retrieved from D_{ref}. P_{int} denotes the interpretation prompt, which is detailed in Appendix A.4.

3.4 Prosecutor Agent

After the investigator agent has gathered relevant evidence and constructed contextual interpretations for the original meme and its rewrites, the prosecutor agent takes over to perform harm assessment. The prosecutor agent evaluates the original meme with different interpretations individually and issues an independent judgment regarding its potential harmfulness. In this way, the prosecutor agent reasons the harmfulness based on different assumed intentions and detect whether harmful meaning persists across perspectives, thereby providing a more reliable basis for final judgment. The process can be formulated as:

$$\mathcal{R}_u = Agent_{Pro}(\mathcal{O}_u, M_{ori}, P_{\text{pro}}), \quad u \in \{ori, b, m\}. \tag{5}$$

where $\mathcal{R}$ denotes the prosecution result, denotes the prosecution result, which includes both the harmful judgment and the support-

ing rationale. P_{pro} denotes the prosecution prompt, which is detailed in Appendix A.5.

The interpretations captured by the investigator agent reflect how the memes harmfulness manifests under distinct dissemination intents. The prosecutor agent reasons over these interpretations to assess whether the harmful interpretation shifts when the meme is rewritten with benevolent or malicious motives. Accordingly, if the harmful judgments from $\mathcal{R}_{\text{ori}}, \mathcal{R}_{\text{b}}, \mathcal{R}_{\text{m}}$ converge to the same conclusion, then no further arbitration is required.

3.5 Judge Agent

During arbitration by the judge agent, we guide the model to concentrate on the core points of contention across different viewpoints. To enhance the objectivity and reliability of the analysis, we introduce auxiliary evidence into the arbitration process. Specifically, we reuse $\mathcal{Q}_{\text{evi}}$ as the input of the judge agent. Unlike the stepwise interpretation of similar samples mentioned in Sec. 3.4, here the original meme and its evidence are fed into the judge agent simultaneously. This enables the judge agent to conduct a joint analysis that takes both the original meme and its evidence into account to reason from a broader contextual basis, which can be formally expressed as:

$$\mathcal{R}_{\text{jud}} = Agent_{Jud}(M_{ori}, \mathcal{Q}_{\text{evi}}, P_{\text{core}}), \quad (6)$$

where $\mathcal{R}_{\text{jud}}$ represents the core representation, P_{core} denotes the related prompt.

$\mathcal{R}_{\text{jud}}$ summarizes the core characteristics shared by similar memes, such as common themes (e.g., pandemics, racial issues) and expressive techniques (e.g., exaggeration, irony). During arbitration, it serves as an anchor that guides the judge agent toward relevant reasoning paths, preventing irrelevant reasoning and improving the objectivity and robustness of the decision. Then, the judge agent integrates all information and complete arbitration process as:

$$\mathcal{J} = Agent_{Jud}(\mathcal{R}_{\text{ori}}, \mathcal{R}_{\text{amb}}, M_{\text{ori}}, \mathcal{R}_{\text{jud}}, P_{jud}), \quad (7)$$

where $\mathcal{R}_{\text{amb}}$ denotes the ambiguity prosecution result, which is inconsistent with the original one. $\mathcal{R}_{\text{amb}} = \{\mathcal{R}_{\text{x}} \mid \mathcal{R}_{\text{x}} \neq \mathcal{R}_{\text{ori}}, x \in \{b, m\}\}$. P_{jud} denotes the final judge prompt. $\mathcal{J}$ denotes the final judgment result and the related supporting rationale. The detailed prompts are provided in Appendix A.6.

In this way, we construct a reasoning chain that mirrors the practical workflow of a criminal case investigation. Hence, PrismAgent is an interpretable framework that provides a stage-wise reasoning process supported by explicit reasoning interpretations, rather than merely outputting a final detection result.

4 Experiment

4.1 Experiment Setup

Datasets. We use three publicly available meme datasets for evaluation: (1) HarM (Pramanick et al., 2021b), (2) FHM (Kiela et al., 2020), and (3) MAMI (Fersini et al., 2022).

Baselines. We compare PrismAgent with various advanced methods: 1) GPT-4o (OpenAI et al., 2023); 2) Gemini-2.0-Flash (Team et al., 2024) 3) Late Fusion (Pramanick et al., 2021a); 4) MOMENTA (Pramanick et al., 2021b) 5) LLaVA-1.5-7B (Liu et al., 2024); 6) InstructBLIP-7B (Dai et al., 2023); 7) MiniGPT-v2-7B (Chen et al., 2023); 8) OpenFlamingo-9B (Awadalla et al., 2023); 9) LLaVA-1.5-13B (Liu et al., 2024); 10) InstructBLIP-13B (Dai et al., 2023); 11) LLaVA-1.5-34B (Liu et al., 2024); 12) MIND (Liu et al., 2025).

Metrics. We use the accuracy and macro-averaged F1 scores as the evaluation metrics. The detail experimental setting and implementation details can be found in Appendix A.

4.2 Harmful Meme Detection Experiments

We evaluate PrismAgent against a range of existing methods on three public datasets, and the results are summarized in Table 1. These methods can be roughly divided into four groups, including the closed-source VLMs, the classical training-based methods, the open-source VLMs, and agent-based methods. It can be seen that PrismAgent consistently outperforms other baselines, demonstrating its strong detection performance. Even with a 13B backbone, PrismAgent achieves performance comparable to models with over 34B parameters, highlighting its efficiency and reasoning capability. Compared with the latest de-

Dataset	HarM		FHM		MAMI	
Model	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
Closed-Source VLMs						
GPT-4o (OpenAI et al., 2023)	67.51	60.29	68.80	68.25	81.00	81.00
Gemini-2.0-FLash (Team et al., 2024)	64.67	58.84	60.40	54.04	80.00	79.92
Training-Based Methods						
Late Fusion (Pramanick et al., 2021a)	73.24	70.25	59.14	44.81	63.20	59.76
MOMENTA (Pramanick et al., 2021b)	83.32	82.80	61.34	57.45	72.10	66.93
Open-Source VLMs						
LLaVA-1.5-7B (Liu et al., 2024)	59.23	49.44	53.80	45.51	52.90	41.53
InstructBLIP-7B (Dai et al., 2023)	51.53	50.99	52.00	48.85	53.10	46.93
MiniGPT-V-2B (Chen et al., 2023)	60.12	52.39	51.30	47.88	57.40	52.22
OpenFlamingo-9B (Awadalla et al., 2023)	63.42	54.36	50.50	49.52	54.70	49.88
LLaVA-1.5-13B (Liu et al., 2024)	62.28	50.45	55.20	53.01	60.10	55.52
InstructBLIP-13B (Dai et al., 2023)	64.92	49.61	55.40	51.89	60.00	57.97
LLaVA-1.6-34B (Liu et al., 2024)	67.51	61.59	64.00	63.51	71.30	71.28
Agent-Based Methods						
MIND (LLaVA-1.5-13B) (Liu et al., 2025)	68.93	65.19	60.80	60.71	68.90	68.84
PrismAgent (LLaVA-1.5-13B)	70.62	68.44	64.00	63.96	70.70	70.69
PrismAgent (LLaVA-1.6-34B)	**71.19**	**69.78**	**66.80**	**66.72**	**73.20**	**73.18**

Table 1: Zero-shot harmful meme detection results on three datasets. The accuracy and macro-averaged F1 scores(%) are reported as the metrics. The best results in **open-source** setting are in **bold**.

Dataset	HarM		FHM		MAMI	
Model	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
LLaVA-1.5-7B	59.23	49.44	53.80	45.51	52.90	41.53
MIND (LLaVA-1.5-7B)	62.71 (+3.48)	57.22 (+7.78)	54.00 (+0.20)	48.28 (+2.77)	53.90 (+1.00)	45.45 (+3.92)
PrismAgent (LLaVA-1.5-7B)	63.56 (+4.33)	58.59 (+9.15)	53.80 (+0.00)	49.29 (+3.78)	55.10 (+2.20)	47.48 (+5.95)
LLaVA-1.5-13B	62.28	50.45	55.20	53.01	60.10	55.52
MIND (LLaVA-1.5-13B)	68.93 (+6.65)	65.19 (+14.47)	60.80 (+5.60)	60.71 (+7.70)	68.90 (+8.80)	68.84 (+13.28)
PrismAgent (LLaVA-1.5-13B)	70.62 (+8.34)	68.44 (+17.99)	64.00 (+8.80)	63.96 (+10.95)	70.70 (+10.60)	70.69 (+15.17)
LLaVA-1.6-34B	67.51	61.59	64.00	63.51	71.30	71.28
MIND (LLaVA-1.6-34B)	69.49 (+1.98)	66.12 (+4.53)	66.40 (+2.40)	68.38 (+4.87)	73.60 (+2.30)	75.38 (+4.10)
PrismAgent (LLaVA-1.6-34B)	71.19 (+3.68)	69.78 (+8.19)	66.80 (+2.80)	66.72 (+3.21)	73.20 (+1.90)	73.18 (+1.90)
Gemini-2.0-FLash	64.67	58.44	60.40	54.04	80.00	79.92
MIND (Gemini-2.0-FLash)	64.84 (+0.17)	60.13 (+1.69)	67.20 (+6.80)	66.05 (+12.01)	80.10 (+0.10)	80.00 (+0.08)
PrismAgent (Gemini-2.0-FLash)	67.51 (+2.84)	63.05 (+4.61)	68.20 (+7.80)	67.51 (+13.47)	80.50 (+0.50)	80.46 (+0.54)

Table 2: Accuracy performance improvements of our proposed framework across different model scales and datasets for zero-shot harmful meme detection. Numbers in red indicate absolute improvements over original models.

tection framework MIND, it also achieves improvements of 2.23% in average accuracy and 2.78% in macro F1-score, fully demonstrating its performance advantages.

4.3 Generalization Experiments

To verify the generalization of the proposed PrismAgent, we integrate it with a variety of VLMs, including both open- and closed-source models. Besides, we compare our performance with the advanced zero-shot framework MIND. As shown in Table 2, PrismAgent consistently enhances performance across the majority of models and datasets, demonstrating its adaptability to diverse backbones. Among them, the performance improvement on LLaVA-1.5-13B is particularly remarkable; even for the more powerful LLaVA-1.6-34B model, the average accuracy and macro F1-score across three datasets are increased by 2.79% and 4.43%, respectively.

4.4 Ablation Study

To comprehensively evaluate the effectiveness of different agents in PrismAgent, we conduct ablation experiments, and the results are shown in Table 3. The experiment results demonstrate that all core strategies in the proposed framework play critical and complementary roles in the harmful meme detection, and omitting any part of them will lead to a degradation of the overall performance. The synergistic effect of these agents significantly improves the robustness of the framework in detecting harmful content, the performance degradation of the model after disabling any

Dataset	HarM		FHM		MAMI	
Model	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
Baseline	62.28	50.45	55.20	53.01	60.10	55.52
PrismAgent (LLaVA-1.5-13B)	70.62	68.44	64.00	63.96	70.70	70.69
w/o Analyst Agent	67.80	59.00	60.00	59.49	67.30	66.09
w/o Investigator Agent	66.67	60.49	58.20	57.65	63.90	62.53
w/o Prosecutor Agent	60.73	53.58	52.20	52.13	60.10	59.96
w/o Judge Agent	67.80	65.13	61.60	61.35	68.20	68.03

Table 3: Ablation studies on our proposed framework.

single strategy fully confirms their effectiveness.

4.5 Analyst Agent Experiments

To evaluate the effectiveness of our intention imagine strategy, we design three sets of comparative experiments: 1) The first setting corresponds to the full version of PrismAgent, including all agent components and functionalities; 2) The second set uses only one of the three memes, either the original or one of its two rewrites, as input, and does not involve the judge agent, as no arbitration is required when the judgments are consistent; 3) The third set includes the original meme and one of its variations, and the judge agent would work when there is a discrepancy between the decisions of the two.

The results are shown in Table 4. It can be observed that our analyst agent effectively enhances performance. Moreover, even when combining only the original meme with a single variation, the performance still shows notable improvement. This can be attributed to the agents ability to reveal latent intentions and amplify information from the original meme. Notably, in the HarM dataset, only treating M_b as input can achieve a higher accuracy. This phenomenon stems from the unbalanced distribution of the dataset, where harmless memes constitute a substantially larger proportion. By incorporating the original meme and its two invariations, our method can partially mitigate the impact of the imbalanced class distribution, providing a more balanced representation for effective harmful meme detection.

4.6 Prosecutor Agent Experiments

For the prosecutor agent, we investigate how the number of retrieved evidence affects the detection performance. The results are presented in Figure 3, we can find that a larger value of K does not necessarily lead to better performance. While increasing K allows the prosecutor agent to have more similar cases, it also tends to include instances with lower relevance, which may introduce noise rather than useful evidence. Our observations indicate that this trade-off between coverage and relevance is crucial for the prosecutor agents effectiveness. Across all datasets, we suggest setting K to 7, which achieves an optimal trade-off by capturing sufficient contextual evidence while minimizing irrelevant or misleading cases.

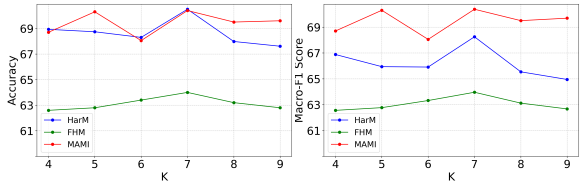

Figure 3: Effect of Top_k in Similar Sample Retrieval of generating Core Representation: (a) Accuracy; (b) Macro F1-score

4.7 Case Study

To thoroughly analyze the processing and evaluation mechanisms of the proposed framework for memes, we select and present two typical cases in Figure 4. If these two samples are directly detected solely based on their original content, both would lead to misjudgments. After intent-enhanced rewriting via our framework, diversified analytical perspectives are provided, creating conditions for the collision and analysis of different viewpoints. This ultimately corrects the initial misjudgments and achieves accurate predictions.[1]

Furthermore, to evaluate the impact of intent rewriting on similar sample retrieval and final judgment, as shown in Figure 5 we se-

[1]**Disclaimer:** This paper includes examples that may be disturbing, shown only for research purposes.

Dataset	HarM		FHM		MAMI	
	Accuracy	Macro-F_1	Accuracy	Macro-F_1	Accuracy	Macro-F_1
PrismAgent (LLaVA-1.5-13B)	70.62	68.44	64.00	63.96	70.70	70.69
w/ M_{ori}	65.25	63.02	63.40	63.22	69.40	69.46
w/ M_{m}	64.69	61.36	59.20	57.77	65.60	65.35
w/ M_{b}	70.62	67.33	62.20	62.17	67.20	67.11
w/ M_{ori} + M_{m}	66.67	64.14	62.80	62.48	70.20	70.19
w/ M_{ori} + M_{b}	67.80	65.12	63.00	62.76	69.20	69.16

Table 4: Experiments about the effectiveness of our Analyst Agent.

Figure 4: Examples of correctly predicted harmful memes: (a) The target meme is harmful, and the analysis result is harmless when analyzed solely based on the original meme, while the analysis result becomes harmful after using PrismAgent; (b) The target meme is harmless, and the analysis result is harmful when analyzed solely based on the original meme, while the analysis result turns harmless after using PrismAgent.

lected another typical correct prediction case. It can be observed that adequate rewriting of the text can effectively optimize the retrieval results, and the introduction of new evidence can generate new perspectives. This in turn helps the system obtain diversified analytical perspectives and ultimately achieve accurate judgments. This result fully demonstrates that our framework can effectively improve the recognition performance of hard-to-detect memes.

Figure 5: Examples of correctly predicted harmful memes with retrieve process: The target meme is harmless, and the analysis result is harmful when analyzed solely based on the original meme, while the analysis result becomes harmless after using PrismAgent.

5 Conclusion and Future Work

To address the reliance on labeled data in harmful meme detection tasks, this paper proposes the zero-shot interpretable detection framework PrismAgent. To tackle the inherent challenge of difficulty in deeply understanding the underlying intent of memes under zero-shot conditions, we have designed a multi-agent role-playing collaborative strategy. Through conducting comprehensive experiments and in-depth analyses on three public datasets, the effectiveness of the proposed framework and strategy has been fully verified. In future work, we plan to extend PrismAgent beyond harmful meme detection to other reasoning multimodal tasks, exploring its broader applicability across diverse content understanding scenarios.

6 Limitations

Although our proposed achieved satisfactory performance, there are several ways to further improve this work:

1) When revealing the hidden intentions of memes, we amplify their underlying meanings by rewriting the original meme in both benevolent and malicious directions. While this method has shown good results, it inevitably introduces some additional computational costs. We believe that predicting the rewriting direction can effectively reduce the computational cost.

2) Although PrismAgent has significantly reduced substantial training costs and data annotation costs, its relatively long inference time due to the need for collaborative invocation of multiple LLMs poses certain challenges for real-time deployment scenarios. However, our extensive experimental results demonstrate that the adoption of lightweight models with smaller parameter sizes or module pruning on existing models can significantly reduce the overall resource consumption.

3) PrismAgent incorporates a variety of strategies to uncover the metaphors behind memes, thereby helping the model gain a deeper understanding of their core semantics, but the final detection performance is still constrained by the capabilities of the underlying LLM. In the future, adopting models with stronger capabilities in capturing sociocultural differences or more superior performance is expected to further improve the accuracy of detection.

7 Ethics Statement

This study aims to combat harmful meme content through zero-shot detection methods, contributing to building a safer online space. The types of harmful content focused on in this study are core issues that have been fully verified in the field of social media research. Our work concentrates on detecting various forms of harmful content, including hate speech, misogynistic speech, and disinformation all of which may have negative impacts on individuals and communities. However, we also recognize that malicious users may use reverse engineering techniques to create memes, in order to evade detection by AI systems like PrismAgent or cause them to make misjudgments. We firmly condemn such behaviors and emphasize that this study is solely for the purposes of scientific research and harmful content prevention. The relevant framework and supporting resources are strictly prohibited from being used for commercial profit or malicious abuse. To ensure the responsible development and evaluation of the framework, we have implemented a number of protective measures: 1) All experiments use publicly available research datasets and fully comply with the usage agreements of each dataset; 2) No user personal data has been collected or used in this study. We believe that the benefits of improving harmful meme detection capabilities far outweigh the potential risks especially at a time when the challenge of governing harmful content on social media is becoming increasingly severe. It should also be noted that the views and content contained in the meme samples do not represent the positions of the studys authors. The framework we designed is intended to assist rather than replace human review work, maintaining a healthy online community environment through collaborative efforts.

References

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint*, arXiv:2308.01390.

Jihwan Bang, Sumyeong Ahn, and Jae-Gil Lee. 2024. Active prompt learning in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27004–27014.

Alexy Beyer and Luca Alexy. 2025. An image equals 16x16 words: Scaling image recogni:on with transformers.

Aditya Borakati. 2021. Evaluation of an international medical e-learning course with natural language processing and machine learning. *BMC Medical Education*, 21(1).

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-cap: Leveraging a frozen vision-language

model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 52445252. ACM.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM Web Conference 2024*, page 45754584. ACM.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26540–26550.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint*, arXiv:2310.09478.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint*, arXiv:2305.06500.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics. [link].

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. 2025. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. In *The Thirteenth International Conference on Learning Representations*.

Rahul Garg, Trilok Padhi, Hemang Jain, Ugur Kursuncu, and Ponnurangam Kumaraguru. 2025. Just KIDDIN' : Knowledge infusion and distillation for detection of INdecent memes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23067–23086, Vienna, Austria. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 770778. IEEE.

Ming Shan Hee, Zihan Gao, Yinglong Wang, Xiangxiang Chu, Roy Ka-Wei Lee, and Zengchang Qin. 2025. Contrastive instruction fine-tuning large multimodal model for hateful meme classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:760773.

Ming Shan Hee and Roy Ka-Wei Lee. 2025. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:774785.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, page 36513655. ACM.

Shixin Hong, Yu Liu, Zhi Li, Shaohui Li, and You He. 2024. Multi-agent collaborative perception via motion-aware robust communication network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15301–15310.

Jianzhao Huang, Hongzhan Lin, Liu Ziyan, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with lmm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 22692293. Association for Computational Linguistics.

Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying creative harmful memes via prompt based approach. In *Proceedings of the ACM Web Conference 2023*, page 38683872. ACM.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint*, arXiv:2005.04790.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint*, arXiv:2205.11916.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact*

(NLP4PI), pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gitanjali Kumari, Jitendra Solanki, and Asif Ekbal. 2025. MemeDetoxNet: Balancing toxicity reduction and context preservation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25076–25098, Vienna, Austria. Association for Computational Linguistics.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, page 23592370. ACM.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. *arXiv preprint*, arXiv:2312.05434.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2628626296. IEEE.

Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. 2025. MIND: A multi-agent framework for zero-shot harmful meme detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–947, Vienna, Austria. Association for Computational Linguistics.

Feipeng Ma, Yizhou Zhou, Yueyi Zhang, Siying Wu, Zheyu Zhang, Zilong He, Fengyun Rao, and Xiaoyan Sun. 2024. Task navigator: Decomposing complex tasks for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2248–2257.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. Gpt-4 technical report. *arXiv preprint*, arXiv:2303.08774.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. *arXiv preprint*, arXiv:2110.00413.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024. Scaling large language model-based multi-agent collaboration. *arXiv preprint*, arXiv:2406.07155.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint*, arXiv:2103.00020.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint*, arXiv:2205.04274.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1561715629. IEEE.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023. Adaplanner: Adaptive planning from feedback with language models. *arXiv preprint*, arXiv:2305.16653.

Shardul Suryawanshi, Mihael Arcan, and Paul Buitelaar. 2020. Nuig at semeval-2020 task 12: Pseudo labelling for offensive content classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, page 15981604. International Committee for Computational Linguistics.

Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. 2024. Magis: Llm-based multi-agent framework for github issue resolution. *arXiv preprint*, arXiv:2403.17927.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*, arXiv:2403.05530.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint*, arXiv:2012.12975.

Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. 2024a. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26275–26285.

Yunxiang Yang, Ningning Xu, and Jidong J. Yang. 2025. Multi-agent visual-language reasoning for comprehensive highway scene understanding. *Preprint*, arXiv:2508.17205.

Ziyuan Yang, Ming Yan, Yingyu Chen, Hui Wang, Zexin Lu, and Yi Zhang. 2024b. Trustworthy hate speech detection through visual augmentation. *arXiv preprint arXiv:2409.13557*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint*, arXiv:2210.03629.

Rasoul Zahedifar, Sayyed Ali Mirghasemi, Mahdieh Soleymani Baghshah, and Alireza Taheri. 2025. Llm-agent-controller: A universal multi-agent large language model system as a control engineer. *arXiv preprint*, arXiv:2505.19567.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):1963219642.

A Implementation Details

A.1 Datasets

We use three publicly available meme datasets for evaluation: (1) **HarM** (Pramanick et al., 2021a), (2) **FHM** (Kiela et al., 2020), and (3) **MAMI** (Fersini et al., 2022). **HarM** consists of memes related to COVID-19. **FHM** was released by Facebook as part of a challenge to crowd-source multimodal harmful meme detection in hate speechsolutions. **MAMI** contains memes that are predom-inantly derogatory towards women, exemplifyingtypical subjects of online vitriol. Different from FHM and MAMI, where each meme was labeled as harmful or harmless, HarM was originally labeled with three classes: very harmful, partially harmful,and harmless. For a fair comparison, we merge the very harmful and partially harmful memes into the harmful class, following the setting of recent work (Pramanick et al., 2021b; Cao et al., 2022; Lin et al., 2023; Huang et al., 2024). The detailed statistics for the original test splits of the three datasets are shown in Table 5.

Dataset	Test	
	#harmful	#harmless
HarM	124	230
FHM	250	250
MAMI	500	500

Table 5: Statistics of test sets.

A.2 Baselines

To benchmark the performance of PrismAgent in zero-shot harmful meme detection, we compare it against a series of state-of-the-art (SOTA) baseline methods: **GPT-4o** (OpenAI et al., 2023): A proprietary large-scale multimodal model developed by OpenAI, which excels in zero-shot visual-language task processing via in-context learning mechanisms; **Gemini-2.0-Flash** (Team et al., 2024): Googles up-to-date multimodal model, characterized by competitive performance in both reasoning tasks and visual comprehension tasks; **Late Fusion** (Pramanick et al., 2021a) and **MOMENTA** (Pramanick et al., 2021b): Two widely recognized earlier approaches, which are trained on manually annotated datasets for harmful meme detection; **LLaVA-1.5-7B** (Liu et al., 2024): A lightweight multimodal model constructed upon the Vicuna-7B foundation, trained on diverse visual instruction datasets to handle general vision-language tasks; **InstructBLIP-7B** (Dai et al., 2023): An instruction-finetuned vision-language model rooted in the BLIP-2 framework, utilizing Vicuna-7B as its language modeling component; **MiniGPTv2-7B** (Chen et al., 2023): A compact but high-performing multimodal model that integrates visual encoding modules with instruction-finetuned language generation capabilities; **OpenFlamingo-9B** (Awadalla et al., 2023): An open-source instantiation of Flamingo-style models, which allows frozen language models to parse visual inputs via cross-

attention mechanisms; **LLaVA-1.5-13B** (Liu et al., 2024): A medium-scale variant of the LLaVA series (built on Vicuna-13B), featuring improved visual grounding and reasoning capacities; **InstructBLIP-13B** (Dai et al., 2023): An upgraded iteration of InstructBLIP, adopting Vicuna-13B as its language model backbone; **LLaVA-1.6-34B** (Liu et al., 2024): The most recent and largest-scale LLaVA model to date, boasting advanced reasoning, OCR, and world knowledge competencies; **MIND**(LLaVA-1.5-13B) (Liu et al., 2025): A state-of-the-art zero-shot harmful meme detection framework available currently. We use the accuracy and macro-averaged F1 scores as the evaluation metrics.

For different baseline methods, we implement the following CoT (Kojima et al., 2022) prompt structure, which achieves better performance than direct classification: " *Given the meme, with the Text: "{}" embedded in the image {}, your task is to determine whether this meme is harmful or not, in order to maintain the harmlessness and integrity of information on the Internet. Please leverage your extensive knowledge to deeply analyze and understand this meme, and give your final judgment. Your output should stricly follow the format: "Thought: [Your analysis] Answer: [harmful/harmless]."* "

For our proposed framework, we conduct extensive experiments with various LMMs, including the LLaVA series (LLaVA-1.5-7B, LLaVA-1.5-13B, LLaVA-1.6-34B) and Gemini-2.0-Flash as our LMM agents. For the baseline of closed-source models, we use the gpt-4o-2024-11-20 and gemini-2.0-flash versions respectively. For pretrained models, we adopt the default parameters as specified in their original papers without modification. For meme rewriting tasks, we set the model temperature to 0.3 to introduce a moderate level of exploratory reasoning. In all subsequent modules, the model temperature is configured to 0 and no sampling mechanisms are enabled, thereby ensuring experimental reproducibility. For multimodal embedding computation, we employ CLIP (Radford et al., 2021) with the specific version ViT-L/14 as our frozen pretrained vision and text Transformer encoders. The modality fusion weights v and t are set to 0.8 and 0.2 respectively through grid search optimization within the [0,1] interval. In our implementation, to retrieve the Top_k most similar reference samples during the process for generating core representations, we empirically set K to 7. This setting achieves a balance between computational efficiency and sufficient contextual information.

In the actual experimental process, for the base model LLaVA-1.5-7b, we use one NVIDIA 4090 24GiB GPU; for the 13b model, we use two NVIDIA 4090 24GiB GPUs; and for the 34b model, we use one NVIDIA H800 80GiB GPU.

A.3 Analyst Agent

The detailed prompts we designed for the benevolent Analyst Agent and malicious Analyst Agent are as follows:

"You are looking at a meme image paired with the following text: {}. Task: Decide if the original text already reads as kind, lighthearted, wholesome, or supportive. - If YES, return the original text unchanged. - If NO, rewrite it to sound more positive, friendly, or uplifting like it's encouraging someone, making them smile, or gently teasing with warmth. Grounding requirements: - Make the rewrite feel tightly connected to the actual image (objects, expressions, actions, or implied context). - The rewrite can highlight anything relevant: people, behavior, stereotypes, or situations depicted but spin it in a kind or amusing way. - Implicitly leverage what's visible to soften the message but don't describe the image directly. Style constraints: - Keep the sentence structure and rhythm close to the original. - Make it short, meme-ready, and punchy. - Use a tone that feels like wholesome humor, gentle teasing, or online positivity but don't break character or explain. Output: Return ONLY the final text (original or rewritten) with NO additional content. Do not include any labels, explanations, brackets, quotes, or extra words. Just the text itself."

"You are looking at a meme image paired with the following text: {}. Task: Decide if the original text already reads as sarcastic, mocking, mean-spirited, or edgy. - If YES, return the original text unchanged. - If NO, rewrite it to sound more hostile, offensive, or mean like it's mocking someone, making fun of them, or taking a nasty jab. Grounding requirements: -

Make the rewrite feel tightly connected to the actual image (objects, expressions, actions, or implied context). - The rewrite can target anything relevant: people, behavior, stereotypes, or situations depicted. - Implicitly leverage what's visible to sharpen the insult but don't describe the image directly. Style constraints: - Keep the sentence structure and rhythm close to the original. - Make it short, meme-ready, and punchy. - Use a tone that feels like edgy humor, mockery, or online trolling but don't break character or explain. Output: Return ONLY the final text (original or rewritten) with NO additional content. Do not include any labels, explanations, brackets, quotes, or extra words. Just the text itself."

Notably, due to the inherent limitations of the models, even though we have designed relatively comprehensive prompts, the model outputs may still contain errors in a few cases. For example, to address this, while preserving the original model outputs as much as possible, we only perform minimal necessary corrections: for instance, obvious redundant expressions (e.g., "rewrite: xxx") are directly simplified to "xxx"; in cases of garbled output or empty output, the original meme content is retained unchanged.

A.4 Investigator Agent

The specific details of how to define the similarity degree when searching for relevant supporting evidence in an unannotated dataset are as follows: For a single meme sample $M = \{\mathcal{V}, \mathcal{T}\}$ we first generate its visual embedding and textual embedding respectively, then fuse them in a specific proportion to obtain the multimodal representation:

$$\mathbf{Emb} = \alpha \cdot \mathbf{VE}(\mathcal{V}) + \beta \cdot \mathbf{TE}(\mathcal{T}), \qquad (8)$$

where **Emb** denotes the multimodal embedding, **VE** and **TE** represent the visual encoder and textual encoder respectively, and α and β are fixed weights in the fusion process. Using the same method, all meme samples in the test set and the reference set D_{ref} are processed to obtain the embedding vectors of all samples. To retrieve the most similar samples, we use cosine similarity to measure the matching degree between the multimodal embeddings of the target meme and the reference samples, with the formula as follows:

$$d = \text{sim}(\mathbf{E}_{\text{target}}, \mathbf{E}_{\text{ref}}), \qquad (9)$$

where $\mathbf{E}_{\text{target}}$ and $\mathbf{E}_{\text{ref}}$ stand for the embedding vectors of the target sample M and the reference sample M_{ref} respectively, $\text{sim}(\cdot)$ denotes the cosine similarity function, $\mathbf{d}$ is the corresponding similarity score. And a higher similarity score $\mathbf{d}$ indicates a greater similarity between the two.

The detailed prompt we designed for the Investigator Agent is as follows: *"You are a sophisticated analytical agent tasked with formulating and refining insightful recommendations. Your ultimate objective is to assess whether a meme is harmful. However, the target meme will not be provided directly. Instead, you will receive a meme related to the target meme. Based on this related meme, you are to synthesize advanced insights to determine whether such memes are harmful, reflect on existing insights, and update these reflections into the current rules. Rules you update must be GENERAL and HIGH LEVEL. Here is the related meme: [Meme] Image: <image placeholder> Text embedded: "org_sent" [Meme] Here are the EXISTING RULES: [EXISTING RULES START] rules [EXISTING RULES END] If the existing rules are empty, it indicates that you are the first person to analyse on this related meme, and your output will be of great importance. Please be serious and careful, think step by step, and ultimately derive the first version of your updated rules. You can directly build upon the existing rules by adding new ones, editing them if there is room for enhancement, and removing those that are contradictory or duplicated. You should focus on distilling insights from the related meme to assess similar memes, rather than determining the harmfulness of the meme currently presented to you. Each rule should be concise and easy to follow. Make sure there are a maximum of five rules. Your output should stricly follow the format: "Thought: [Reflect on whether the current rules are helpful in assessing the harmfulness of this related meme.] Operations of updating existing rules: [Update your reflections on this meme into the existing rules. You can choose to add, edit, and remove or retain existing rules. You may perform up to two operations. Please specify in detail*

which rule you want to operate on and how, but do not mention any operations in the rules themselves; your operations regarding the rules should only be mentioned here.] Updated rules: [Do not mention the operations you have made and the related meme in the rules, directly output all the updated rules, only output the rules that have been determined, and do not leave spaces for rules that have not yet been generated. Rules you update must be GENERAL and HIGH LEVEL.]" "

A.5 Prosecutor Agent

The detailed prompt we designed for the Prosecutor Agent is as follows: *"Given the meme, with the Text: {} embedded in the image {}. Is it harmful or harmless? Note: [{note}] (Consider this note above with caution - it may contain biases or subjective interpretations that could influence judgment. Evaluate its reliability and relevance carefully) Your output should strictly follow the for mat: "Thought: [First, analyze the memes content independently. Then, carefully consider how the provided note may inform or bias your understanding. Weigh the notes credibility and relevance before incorporating it into your final assessment.] Answer: [harmful/harmless]."*

A.6 Judge Agent

The detailed prompt we designed for the process of generating Core Representation is as follows: *" Task: Analyze the target meme and its 7 most similar memes to identify their common characteristics. Target Meme: - Text: "{target_text}" - (Image is provided) Similar Memes:""" for i, sample in enumerate(similar_samples, 1): prompt += f"""Meme i: - Text: "sample['text']" - (Image is provided)""" Requirements: 1. Identify shared themes, topics, and content elements across all memes 2. Analyze common expressive techniques (humor, symbolism, irony, etc.) 3. Summarize recurring patterns in imagery, text style, or cultural references 4. Focus only on descriptive characteristics, not on whether they are harmful or harmless 5. Provide a concise, structured summary of these commonalities Output: A clear summary of the shared characteristics between these memes. "*

For the Judge Agent, the specific prompt we designed is as follows: *"Given the meme, with the Text: {orig_text} embedded in the image. Is it harmful or harmless? CORE PERSPECTIVES: Investigator A: Verdict: {investigator_a_verdict} Reasoning: {investigator_a_reasoning} DISSENTING PERSPECTIVES: {dissenting_investigators} ANALYSIS OF SIMILAR MEMES: {Core Representation} As a judge, your task is to: 1. First, identify the key contradictions between the opposing viewpoints. Analyze where and why the investigators disagree, focusing on their interpretations of the meme's text, imagery, and implied meaning. 2. Then, carefully evaluate the meme itself (considering its text and image) and the shared characteristics of similar memes provided above. 3. Determine which*

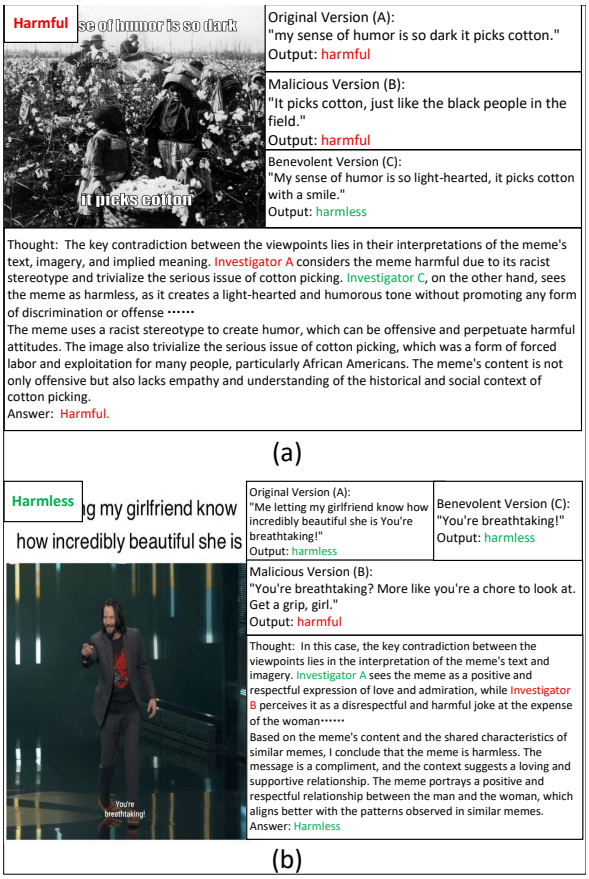

Figure 6: Examples of correctly predicted harmful memes: (a) The target meme is harmful, Although the analysis result of the benevolent rewritten meme is harmless, the final analysis result remains harmful after using PrismAgent; (b) The target meme is harmless, Although the analysis result of the malicious rewritten meme is harmful, the final analysis result remains harmless after using PrismAgent.

viewpoint is more consistent with the actual content of the meme and aligns better with the patterns observed in similar memes. 4. Explain your reasoning for resolving the contradiction. Your response should strictly adhere to this format: Thought: [First, analyze the key contradictions between the viewpoints. Then, explain how the meme's own content and the similar memes' characteristics inform your judgment. Finally, state which viewpoint is more appropriate and why.] Answer: [Your final judgment(harmful/harmless)]."

A.7 Case Study

In addition to the examples presented in the main text, we further provide two correct prediction cases of PrismAgent in Figure 6, to illustrate that our framework exhibits a certain degree of robustness against intentional rewriting. These two typical cases indicate that interfering noise is inevitably introduced during the process of intentional reasoning and rewriting, causing interference to memes that were correctly analyzed initially. However, PrismAgent exhibits excellent anti-interference performance and will not deviate from the final correct judgment due to such interference.