Generative Agents for Multimodal Controversy Detection

Tianjiao Xu[1] , **Jinfei Gao**[1] , **Keyi Kong**[1] , **Jianhua Yin**[1] , **Tian Gan**[1,*] and **Liqiang Nie**[2]
[1]Shandong University
[2]Harbin Institute of Technology (Shenzhen)
{xutianjiao1998, jinfei_gao, luxinyayaya}@mail.sdu.edu.cn, {gantian, jhyin}@sdu.edu.cn, nieliqiang@gmail.com

Abstract

Multimodal controversy detection, which involves determining whether a given video and its associated comments are controversial, plays a pivotal role in risk management on social video platforms. Existing methods typically provide only classification results, failing to identify what aspects are controversial and why, thereby lacking detailed explanations. To address this limitation, we propose a novel **Agent**-based **M**ultimodal **C**ontroversy **D**etection architecture, termed **AgentMCD**. This architecture leverages Large Language Models (LLMs) as generative agents to simulate human behavior and improve explainability. AgentMCD employs a multi-aspect reasoning process, where multiple judges conduct evaluations from diverse perspectives to derive a final decision. Furthermore, a multi-agent simulation process is incorporated, wherein agents act as audiences, offering opinions and engaging in free discussions after watching videos. This hybrid framework enables comprehensive controversy evaluation and significantly enhances explainability. Experiments conducted on the MMCD dataset demonstrate that our proposed architecture outperforms existing LLM-based baselines in both high-resource and low-resource comment scenarios, while maintaining superior explainability.

1 Introduction

Social video platforms serve as important channels for the dissemination of contemporary information, hosting a large amount of video content that spreads worldwide at unprecedented speed [Newman *et al.*, 2023; Gan *et al.*, 2023b]. This content profoundly influences public cognitive perceptions, emotional tendencies, and societal behavior patterns [Wang *et al.*, 2022; Gan *et al.*, 2023a]. The rapid spread and significant impact of large amounts of videos highlight the urgent need for early detection and effective management of dissemination risks [Hessel and Lee, 2019; Wang *et al.*, 2023c]. Particularly concerning are videos that can provoke widespread social controversies [Conover *et al.*, 2011], evoke negative emotions [Wang *et al.*, 2023a], or exacerbate social divisions [Tarrow, 2008]. The identification and management of controversial content, referred to as controversy detection [Martin, 2014], requires increased attention and robust methodologies.

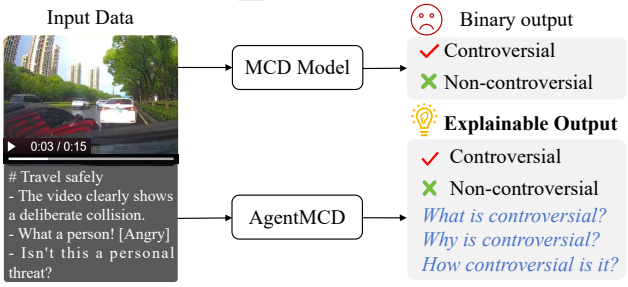

Figure 1: Comparison of AgentMCD with traditional MCD models.

Multimodal controversy detection is an emerging field that aims to assess whether a given video and its associated content are controversial [Xu *et al.*, 2024]. If not properly managed, controversial videos can spread widely, potentially damaging individual reputations, disrupting group dynamics, and threatening societal stability [Mejova *et al.*, 2014]. Effective controversy detection facilitates timely identification of risks, supports the implementation of appropriate intervention measures, and helps maintain social media platforms as constructive environments for information exchange [Garimella *et al.*, 2018]. Controversial videos on social media can be defined from three perspectives. First, it considers whether the video itself is controversial, such as through sensationalism or violence. Second, it examines conflicts between the video content and user comments, including opposing viewpoints, personal attacks, and criticisms. Third, it focuses on the controversy within the comments, looking for clear support or opposition.

A significant limitation of existing methods is that they only provide a binary output indicating whether a video and its associated content are controversial, without offering detailed explanations. Emphasizing the ethical implications and explainability of model predictions is essential, particularly in sensitive contexts such as controversy. As the adage goes,

*Corresponding author.

"There are a thousand Hamlets in a thousand people's eyes", reflecting the notion that each individual has their interpretation or perspective, which is particularly relevant in understanding controversy. To address this gap, we strive to explore more explainable approaches for this task.

Recent advancements in Large Language Models (LLMs) have significantly enhanced capabilities in natural language understanding and generation [Bang *et al.*, 2023; Jiao *et al.*, 2023]. Various training paradigms have also emerged, enabling LLMs to perform tasks in a zero-shot manner and adhere more closely to human-provided instructions [Sanh *et al.*, 2022]. Although a single powerful LLM is already capable of addressing a wide range of tasks, recent studies suggest that multiple LLMs can further enhance each other's performance through debate and collaboration [Du *et al.*, 2024; Liang *et al.*, 2024]. By integrating multiple LLMs into a cohesive group and designing specific interaction mechanisms, these models can propose and deliberate on unique responses and thought processes across several rounds. Additionally, these agents can simulate human behavior based on user-specified descriptions and profiles [Shanahan *et al.*, 2023]. Such advancements enable LLM-based agents to effectively handle complex scenarios on social media, simulate controversy formation, and improve the explainability of controversy detection. As illustrated in Figure 1, traditional Multimodal Controversy Detection (MCD) models can only output a binary indicator of controversy based on the input video and text. In contrast, incorporating LLM-based agents enhances the models' ability to address more explanatory questions, such as "*What is controversial?*", "*Why is it controversial?*", and "*How controversial is it?*".

In this light, we propose a framework called Agent-based Multimodal Controversy Detection (AgentMCD), which utilizes LLM-based agents to generate explainable text and simulate the formation of controversy. Our approach considers two scenarios: one with abundant comments and another with limited comments. In scenarios with abundant comments, AgentMCD primarily uses LLM-based agents as judges to conduct multi-aspect reasoning, delivering a comprehensive evaluation of video controversies. This includes assessing: (1) the intrinsic controversy of the video content, (2) the controversy between the video content and user comments, and (3) the controversy within the comments themselves. To emphasize the importance of early detection, we also investigate the user engagement simulation in the initial stages of video release, marked by a limited number of comments, where generative agents simulate audience comments and discussions. We evaluate our method using the MMCD dataset [Xu *et al.*, 2024] and compare it with various publicly available LLM-based methods. The experiments demonstrate that AgentMCD achieves notable reasoning performance and effectively simulates controversy with multi-agent mechanism. Ablation studies validate the effectiveness of the proposed multi-aspect reasoning and multi-agent simulation. Finally, we conduct comprehensive case studies and in-depth analyses to assess the effectiveness of the proposed method. This work tackles the critical societal challenge of risk control in short videos, aiming to provide early warnings and directional guidance for risks encountered during the dissemination of short videos.

Our main contributions are summarized as follows:

- We are the first to explore incorporating LLM-based agents into the task of multimodal controversy detection, thereby enhancing the explainability of the process.
- We introduce a novel framework called Agent-based Multimodal Controversy Detection (AgentMCD), which utilizes a multi-aspect reasoning process to systematically evaluate controversy. Furthermore, it incorporates a multi-agent simulation mechanism to model the formation of controversy in the early stages of video dissemination.
- Extensive experiments highlight the superiority of our approach compared to other LLM-based methods, while maintaining a good explainability. Our work is publicly available[1].

2 Related Work

2.1 Controversy Detection

Controversy detection has garnered significant attention in recent research due to its broad implications across various domains such as risk management, content moderation, and sentiment analysis [Linmans *et al.*, 2018; Zhong *et al.*, 2020; Hessel and Lee, 2019; Xu *et al.*, 2024]. The early studies in this area typically assumed that topics inherently possessed controversy and focused on identifying such controversial topics [Popescu and Pennacchiotti, 2010; Garimella *et al.*, 2018]. However, controversial topics cover a wide range of evolving subjects and they do not directly reveal the inherent properties of controversy. Recent studies have further deepened the understanding of controversy identification and analysis, considering various factors such as semantic [Linmans *et al.*, 2018], viewpoint consistency [Hessel and Lee, 2019; Zhong *et al.*, 2020], contextual content [Beelen *et al.*, 2017], and the influence of cultural and social backgrounds [Dori-Hacohen and Allan, 2015]. The application of controversy detection has expanded to a variety of platforms, from traditional web pages [Dori-Hacohen *et al.*, 2016; Linmans *et al.*, 2018] to more dynamic and socially interactive environments like social media [Garimella *et al.*, 2018; Koncar *et al.*, 2021]. Notably, there has been a growing interest in detecting controversy in multimodal contexts, particularly on social media platforms [Xu *et al.*, 2024], underscoring the importance of addressing the complexities of multimodal content.

To effectively address evolving applications, various methods for controversy detection have been developed. Traditional approaches have primarily relied on statistical techniques [Popescu and Pennacchiotti, 2010; Hamad *et al.*, 2018], which identify controversial content based on specific terms, phrases, or predefined rules. More recent research has shifted towards graph-based methods [Mendoza *et al.*, 2020; Benslimane *et al.*, 2023], which effectively capture structural relationships among users, topics, and comments [Benslimane *et al.*, 2023; Zhong *et al.*, 2020; Li *et al.*, 2023]. Early

[1]Codebase is available at https://github.com/skylie-xtj/AgentMCD_Release.

prediction of controversy has also gained increasing attention [Hessel and Lee, 2019]. For multimodal controversy detection, methods integrating neural network modules have been developed to identify controversial videos based on specific definitions of controversy [Xu *et al.*, 2024]. The advancement of pretrained language models offers new tools and methodologies for controversy detection [Calvo Figueras *et al.*, 2023; Canute *et al.*, 2023]. However, these methods typically provide only binary outputs and lack explanatory details. To address this limitation, our approach incorporates generative agents to deliver more explainable results.

2.2 Generative Agents

Recent advancements in Large Language Models (LLMs) have showcased their exceptional capabilities in understanding and reasoning in cross-disciplinary research [Bang *et al.*, 2023; Jiao *et al.*, 2023]. A notable development in this field is the emergence of generative agents, which are designed to role-play specific characters or perform targeted tasks [Park *et al.*, 2023; Shanahan *et al.*, 2023]. Researchers have invested substantial efforts in exploring how to utilize these models more effectively to solve various complex problems, thereby advancing the exploration of the application boundaries and potential of LLMs [Kojima *et al.*, 2022; Wang *et al.*, 2023b; Sun *et al.*, 2023; Shinn *et al.*, 2023]. In the field of multi-agent systems, current research primarily focuses on simulating human behavior and tackling intricate tasks through collaborative approaches [Park *et al.*, 2023; Zhu *et al.*, 2023; Wu *et al.*, 2023; Hong *et al.*, 2024]. For instance, generative agents [Park *et al.*, 2023] create a sandbox environment that facilitates reliable human behavior simulation, allowing intelligent agents to interact with one another. Ghost in the Minecraft [Zhu *et al.*, 2023] integrates LLMs with text-based knowledge and memory to develop generally capable agents within the Minecraft environment. AutoAgent [Chen *et al.*, 2024] serves as a cooperative agent framework that enables role-playing, collaboration, and the resolution of complex tasks. These methodologies significantly enhance decision-making efficiency and accuracy [Zhang *et al.*, 2024].

Significant progress has also been made in simulating social dynamics using LLMs [Cheng *et al.*, 2023; Liu *et al.*, 2023; Liu *et al.*, 2024]. CoMPosTCA [Cheng *et al.*, 2023] highlights LLMs' unique advantages in portraying human traits and managing comic material with personalization and exaggeration. The stable alignment method [Liu *et al.*, 2023] allows LLMs to learn from simulated social interactions, while [Liu *et al.*, 2024] used LLMs to simulate the dynamics of fake news propagation in social media environments and explore its impact on public attitude changes. These studies demonstrate the potential of using LLM agents to emulate human social behaviors, offering new insights into understanding and predicting human actions.

To the best of our knowledge, our work is the first to propose the use of generative agents specifically for multimodal controversy detection.

3 Methodology

In this section, we introduce AgentMCD, a hybrid framework that integrates a controversy simulation mechanism with an explainable reasoning process. This framework is designed to model audience behavior on social video platforms effectively and to predict the long-term engagement impact of videos reliably, even in early-stage scenarios with limited user comments. This is achieved by leveraging the human-like capabilities of LLM-powered generative agents, providing interpretable outputs for multimodal controversy detection. To realize these objectives, two core components are emphasized: (1) *User Engagement Simulation*: This component faithfully replicates users' personalized preferences and simulates the process of browsing social videos and commenting. (2) *Multi-Aspect Reasoning*: This component evaluates controversy scores through a comprehensive process while ensuring good explainability. An overview of the AgentMCD framework is presented in Figure 2, and the complete algorithmic process is described in Algorithm 1.

3.1 Task Formulation

Multimodal controversy detection task aims to detect whether a given video and its associated content are controversial. For an input sample that includes video $\mathcal{V}$ and additional context information $\mathcal{C}$, let $\hat{y} \in \{0, 1\}$ indicate the predicted result of whether the input is controversial. Our objective is to design an explainable multi-modal controversy detection framework F:

$$\hat{y}, \mathcal{O} = \mathrm{F}\left(\mathcal{V}, \mathcal{C}\right), \tag{1}$$

where $\mathcal{O}$ denotes the explainable output report, including controversy scores and judgment bases.

3.2 Initialization

The Initialization module outlines the formation of the video profile and user profile within AgentMCD, both of which are fundamental to the subsequent analysis.

Video Profile

In the subsequent module, the video profile—especially the generated video description—plays a pivotal role. Each input dataset includes raw video, title, keywords, metadata about the publishers, comments, and recognized Automatic Speech Recognition (ASR) content. These components, along with the video description generated by the LLM, collectively form the video profile. In general, video descriptions are generated by encoding the input video using a video model aligned with a language model, and subsequently deriving outputs from the language model. Given that the video content is in Chinese, we utilize mPLUG-Video [Xu *et al.*, 2023], fine-tuned on a 10-million Chinese large-scale video-text dataset, to generate descriptions from raw video.

However, this approach occasionally produces irrelevant descriptions. For instance, when publishers express personal opinions on certain topics while appearing on camera, the generated description might merely state, "a person speaking to the camera, occasionally covering their nose or pacing back and forth." Such descriptions are insufficient for our purposes, as we require more detailed and contextually relevant content. Specifically, we are interested in understanding what the person is discussing, the opinions they express, and the conclusions they draw. To address this issue, we further

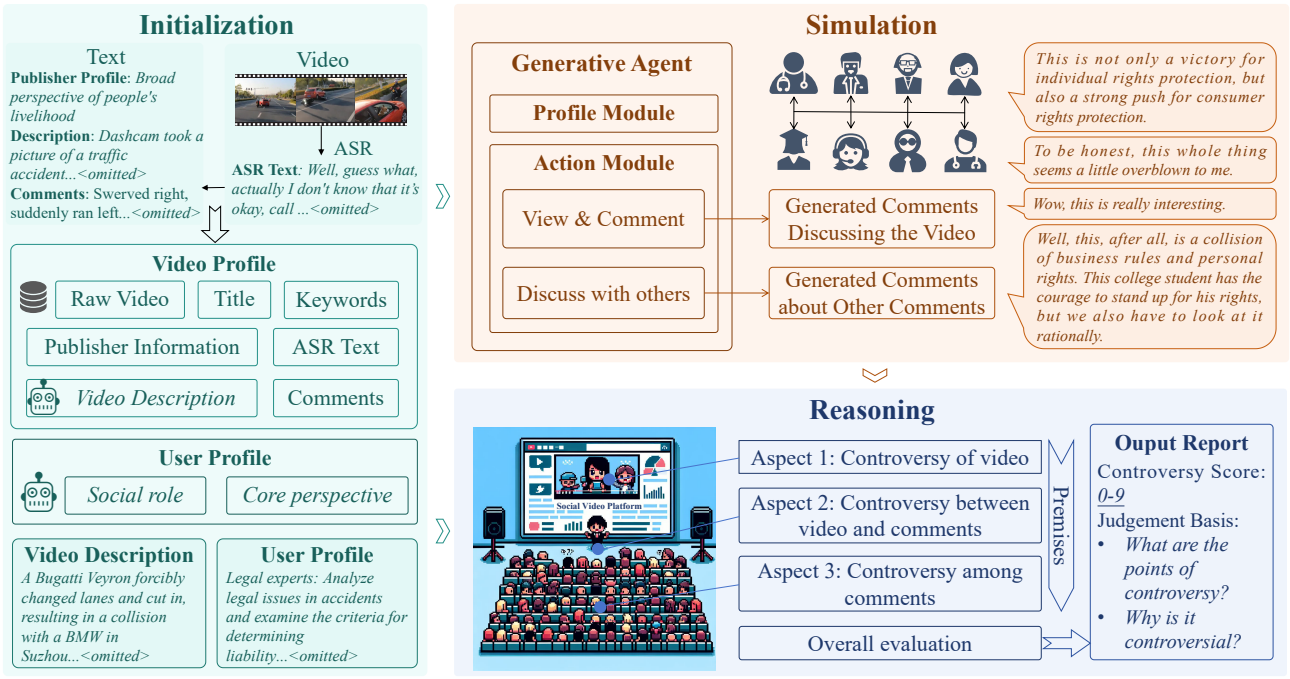

Figure 2: Architecture of Agent-based Multimodal Controversy Detection (AgentMCD) framework.

refine the video description through prompt engineering, incorporating additional information from the video profile, including title, keywords, publisher information, and ASR text, to enrich the level of detail in these descriptions.

User Profile

In the subsequent simulation process, the user profile module plays a pivotal role in aligning agent behavior with genuine human responses. Traditionally, agent descriptions are either predefined or randomly selected from an established dataset [Zhang *et al.*, 2024; Park *et al.*, 2023; Liu *et al.*, 2024]. However, we observed during experiments that such initialization often led to repetitive responses and a lack of diverse perspectives.

To overcome this limitation, we explored alternative initialization strategies. Drawing inspiration from the personalized recommendation mechanisms commonly used on video platforms, we began generating audience profiles based on video descriptions. We also refined the prompting process to encourage a broader range of opinions.

Each agent's profile consists of two key components: social role and core perspective. The social role includes characteristics such as profession or personal interests—for example, a student preparing for a teaching certification. The core perspective captures the agent's main concerns or viewpoints; for instance, a media commentator may approach a case from a news angle, focusing on its societal impact and media coverage. This design significantly improved the diversity of generated opinions, better capturing the distinct social dynamics associated with different videos.

3.3 Simulation

The Simulation module employs a multi-agent mechanism to replicate the user engagement process, enabling the observation of controversy emergence in complex social video platform scenarios and offering valuable insights into the dynamics of controversy. This module is particularly crucial in contexts where commentary is sparse.

Generative agents in AgentMCD, built on an LLM-based foundational architecture, are enhanced with two specialized modules designed specifically for social video platform scenarios: the *profile* module and the *action* module. To emulate personalized and authentic human behavior, each agent incorporates a user profile module that combines social roles and core perspectives, effectively representing the types of users likely to view the video and the aspects they might focus on. Additionally, inspired by the processes humans undergo while interacting with social video platforms, the agents are equipped with action modules that enable them to comment and express emotions in a coherent and contextually appropriate manner. They comment after viewing videos, observe other audiences' reactions and comment, simulating the reactions of viewers encountering the video on a social platform. During these processes, agents' responses may be positive or negative, aiming to replicate the natural flow of interaction between videos and their audiences. This approach effectively simulates the potential conflicts and controversies that may arise from these interactions.

3.4 Reasoning

The Reasoning module serves as the core component of our architecture, where we implement a multi-aspect evaluation

Algorithm 1 AgentMCD: Agent-based Multimodal Controversy Detection

Input: Video title T, comments C, publisher information P, ASR text R, threshold t
Output: Predicted controversy label $\hat{y}$, output report $\mathcal{O}$
Generate detailed video descriptions D based on T, P, R

if C is None **then**
  Simulation Process:
  Generate adaptive agents' descriptions A based on D
  for each agent description a in A **do**
    Initialize Agent i based on a
    Generate comments on D and add to C
    Generate more comments based on C and add to C
  end for
end if
Multi-aspect Reasoning Process:
Aspect 1: Evaluate controversy of video content based on D
Aspect 2: Evaluate controversy between video D and comments C
Aspect 3: Evaluate controversy within comments C
Judge the final controversy score s based on the evaluations of the above three aspects
Final controversy score s and judgement basis b are added into $\mathcal{O}$
$\hat{y} \leftarrow \begin{cases} 0, & \text{if s <t,} \\ 1, & \text{otherwise.} \end{cases}$
Return: Predicted controversy label $\hat{y}$, output report $\mathcal{O}$

process to comprehensively assess controversy from three dimensions: (1) *controversy inherent in the video itself*, such as the presence of sensationalism, violent content, or other risky elements; (2) *controversy between the video and its associated comments*, where videos are considered controversial if comments reflect opposing viewpoints, include personal attacks on the video creator, critique the phenomena depicted, or question the individuals or objects featured; and (3) *controversy among comments*, characterized by disagreements, such as debates or clear expressions of opposing viewpoints (e.g., support versus opposition).

In scenarios with limited comments, we incorporate generated comments from the Simulation module to supplement the evaluation. The final prediction is derived by synthesizing the assessments across these three dimensions. Each aspect of evaluation includes two key components: (1) *controversy score*, ranging from 0 to 9 (with 0 indicating no controversy and 9 representing high controversy), and (2) *judgement basis*, which explains the assigned score by identifying the specific sources and reasons for the controversy.

4 Experiments

In this section, we present experiments to compare our proposed models with various baseline models. Specifically, we aim to address the following evaluation questions:

EQ1: Is our proposed AgentMCD framework more effective than other LLM-based methods in scenarios?

EQ2: Does our proposed framework effectively utilize multi-aspect reasoning, and does multi-agent demonstrate effectiveness?

EQ3: How does the quality of comments generated by the simulation module compare to real comments?

EQ4: How does the multi-aspect reasoning module enhance the interpretability of multimodal controversy detection tasks?

4.1 Dataset

We evaluate the MMCD dataset [Xu *et al.*, 2024], a large-scale Multimodal Controversial Dataset consisting of over 10,000 Chinese videos, accompanied by extensive social context information. The MMCD was sourced from Douyin[2], a popular Chinese social video platform with a substantial user base comprising millions of active participants. For our experiments with LLM-based methods, we utilize only the validation and test data, excluding the training data. Specifically, the validation set contains 1,130 samples, while the test set includes 1,132 samples.

4.2 Implementation Details

Our proposed framework is implemented using a Python script, utilizing GLM4-9B [GLM, 2024] as the backbone LLM with a temperature setting of 0 to ensure reproducibility. GLM4-9B, from the latest GLM-4 series by Zhipu AI, supports long-text reasoning up to 128K tokens. All experiments are conducted in one-shot or few-shot settings without additional training or fine-tuning of the language models. Each method first determines the optimal threshold based on the validation set, which is then used to evaluate the test set results.

4.3 Baselines

To validate the effectiveness of our methods, we implemented several representative LLM-based methods on the MMCD dataset. The following elaborates on them:

Standard Prompting employs the GLM4-9B [GLM, 2024] backbone model with a standard zero-shot prompt. Zero-shot Chain of Thought (CoT) [Kojima *et al.*, 2022] uses "Let's think step by step" as a prompt for LLMs. Plan-and-Solve (PS) [Wang *et al.*, 2023b] involves devising and executing a task plan in two steps. Self-Consistency [Wang *et al.*, 2023d] samples multiple LLM responses and uses majority voting to determine the final answer. Self-Reflect [Shinn *et al.*, 2023] involves LLMs refining their outputs until satisfactory. Self-Refine [Madaan *et al.*, 2023] improves initial LLM outputs through iterative feedback. Tree of Thoughts (ToT) [Yao *et al.*, 2024] enables decision-making by evaluating reasoning paths and adjusting actions. Cumulative Reasoning (CR) [Zhang *et al.*, 2023] uses iterative, cumulative reasoning to solve problems. RECITE [Sun *et al.*, 2023] retrieves relevant passages from LLM memory before producing answers.

We also incorporate several multi-agent methods: AutoAgents [Chen *et al.*, 2024] is a framework for building and coordinating AI agents. Multi-Agent (Debate) [Du *et al.*, 2024]

[2]https://www.douyin.com/

Method	with abundant comments				with limited comments			
	F1	Rec.	Prec.	Acc.	F1	Rec.	Prec.	Acc.
Standard Prompting	67.91	67.93	67.98	67.93	64.39	64.40	64.42	64.40
Zero-shot CoT [Kojima *et al.*, 2022]	69.15	69.17	69.22	69.17	64.02	64.05	64.09	64.05
Plan-and-Solve [Wang *et al.*, 2023b]	68.11	68.11	68.12	68.11	62.87	63.25	63.81	63.25
Self-Consistency [Wang *et al.*, 2023d]	65.76	65.83	65.95	65.83	61.75	61.75	61.75	61.78
Self-Reflect [Shinn *et al.*, 2023]	69.13	69.26	69.58	69.26	63.96	63.96	63.96	63.96
Self-Refine [Madaan *et al.*, 2023]	62.08	63.96	67.41	63.96	61.78	63.25	65.66	63.25
Cumulative Reasoning (CR) [Zhang *et al.*, 2023]	50.16	57.42	67.78	57.42	58.22	59.72	61.34	59.72
RECITE [Sun *et al.*, 2023]	61.89	63.03	64.82	63.04	53.14	57.42	61.69	57.42
Tree of Thoughts (ToT) [Yao *et al.*, 2024]	68.41	68.46	68.58	68.46	60.81	61.61	62.66	61.63
AutoAgents [Chen *et al.*, 2024]	64.17	65.2	65.93	64.39	60.14	60.28	60.42	60.27
Multi-Agent (Debate) [Du *et al.*, 2024]	63.44	64.27	65.68	64.25	60.46	60.75	61.07	60.74
MAD [Liang *et al.*, 2024]	61.84	61.84	61.85	61.85	62.58	62.63	62.71	62.63
AgentMCD (Ours)	**69.98**	**70.14**	**70.60**	**70.14**	**66.29**	**66.45**	**66.78**	**66.46**

Table 1: Performance (%) comparison among different methods on MMCD in terms of F1-score, recall, precision, and accuracy.

Method		F1	Rec.	Prec.	Acc.
with abundant comments	Aspect1-only	63.65	63.69	63.77	63.69
	Aspect2-only	64.74	65.99	68.63	65.99
	Aspect3-only	65.72	65.99	66.52	65.99
	w/o Aspect1	67.04	67.05	67.07	67.05
	w/o Aspect2	67.93	67.93	67.95	67.93
	w/o Aspect3	68.94	68.99	69.13	68.99
AgentMCD		**69.98**	**70.14**	**70.60**	**70.14**
with limited comments	Aspect1-only	63.36	63.79	64.48	63.81
	Aspect2-only	63.35	63.96	64.99	63.98
	Aspect3-only	63.80	64.32	65.20	64.34
	w/o Aspect1	65.86	66.08	66.50	66.08
	w/o Aspect2	64.78	65.02	65.44	65.02
	w/o Aspect3	65.19	65.28	65.44	65.28
	AgentMCD (SA†)	65.09	65.28	65.64	65.28
AgentMCD (MA)		**66.29**	**66.45**	**66.78**	**66.46**

† SA: Single Agent; MA: Multi-Agent.

Table 2: Experimental results of the ablation study, including F1-score, recall, precision, and accuracy.

involves multiple LLM instances proposing, debating, and refining answers. MAD [Liang *et al.*, 2024] uses a debate process with multiple agents and a judge to reach a final solution.

4.4 Main Results (EQ1)

Table 1 presents the quantitative evaluation, where AgentMCD consistently outperforms other LLM-based approaches in both abundant and limited comment scenarios, demonstrating its superior effectiveness and simultaneously achieving explainability.

In scenarios with abundant comments, LLM-based methods such as Zero-shot CoT [Kojima *et al.*, 2022], Plan-and-Solve [Wang *et al.*, 2023b], Self-Reflect [Shinn *et al.*, 2023], and Tree of Thoughts [Yao *et al.*, 2024] exceed the performance of the Standard Prompting approach. However, some previous multi-agent methods, such as AutoAgents [Chen *et al.*, 2024], Multi-Agent (Debate) [Du *et al.*, 2024], and MAD [Liang *et al.*, 2024], have not performed well. These methods rely on cooperative mechanisms to achieve objectives or debate mechanisms to refute opposing arguments and reach a correct answer. In contrast, AgentMCD introduces an innovative framework incorporating a multi-agent simulation mechanism that specifically models user engagement within the video propagation process and simulates the formation of controversies. In scenarios with limited comments, not only do multi-agent methods but also certain LLM-based methods that perform well in abundant environments fail to enhance the performance of Standard Prompting. Thus, AgentMCD proves to be particularly effective in such environments by simulating real-world controversy formation.

4.5 Ablation Studies (EQ2)

We conducted a series of ablation experiments to evaluate the significance of various components in the proposed AgentMCD framework, with results presented in Table **??**. Specifically, we examined the contribution of each aspect within the reasoning module by independently evaluating controversy within video content (Aspect 1), controversy between video content and comments (Aspect 2), and controversy among comments (Aspect 3). The results demonstrate that each aspect plays a vital role in the overall performance of AgentMCD, with Aspect 3 consistently achieving the best performance across all scenarios. Furthermore, the ablation study revealed that AgentMCD with a single agent underperforms compared to its multi-agent counterpart, underscoring the effectiveness of the multi-agent mechanism.

4.6 Simulation Evaluation (EQ3)

To evaluate the quality of generated comments in the simulation module, we employed an additional LLM (the Qianwen series) alongside human evaluators to score both real and generated comments across multiple dimensions. The results are presented in Table **??**. The evaluation criteria include: (1) *Relevance*: examining topic alignment; (2) *Consistency*: evaluating sentence coherence; (3) *Logicality*: measuring the logical flow; (4) *Sentiment Polarity*: determining emotional

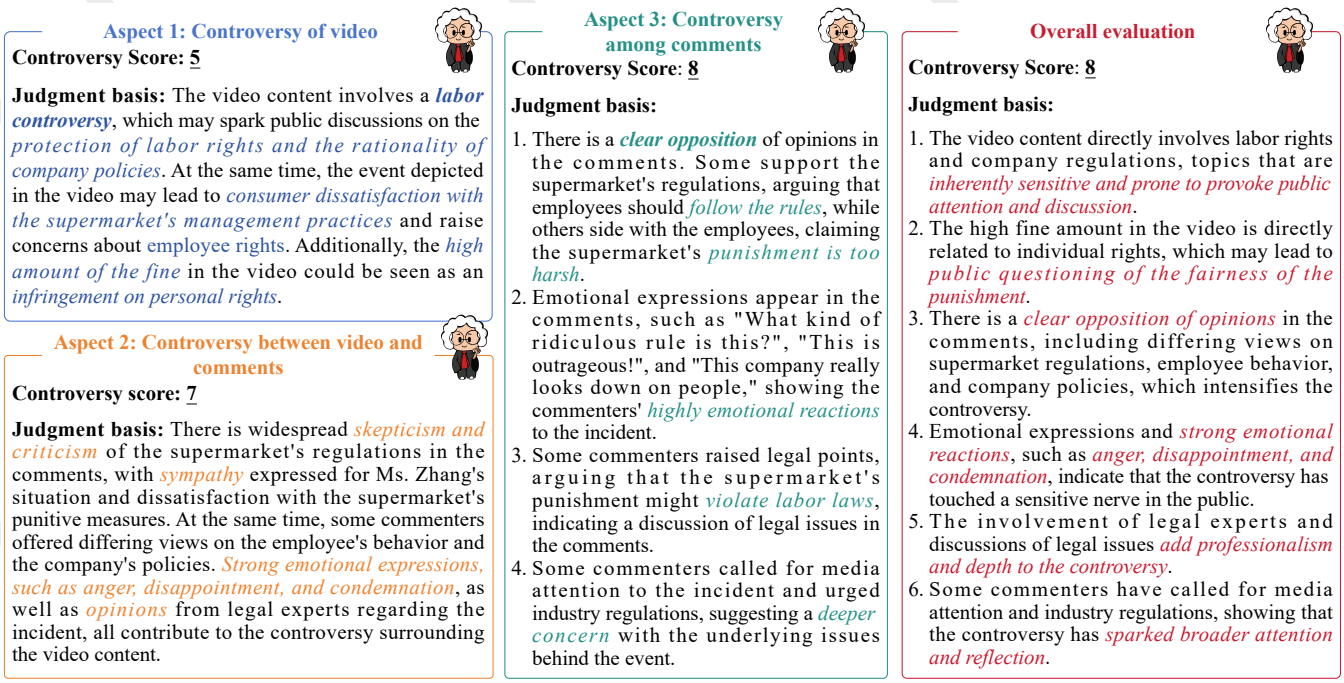

Figure 3: Case study of a controversial video demonstrating the results of the multi-aspect reasoning module.

Aspect	LLM Score		Human Score	
	RC†	GC	RC	GC
Relevance	7.97	**8.42**	7.57	**8.44**
Consistency	7.84	**8.28**	7.35	**7.78**
Logicality	4.99	**6.42**	6.80	**8.40**
Sentiment Polarity	8.21	**8.45**	7.05	**7.80**
Diversity	7.27	**7.28**	7.01	**7.54**

† RC: Real Comments; GC: Generated Comments.

Table 3: Comparison between real and generated comments across various aspects, including relevance, consistency, logicality, sentiment polarity, and diversity, presented through both LLM-based evaluations and human assessments.

intensity; and (5) *Diversity*: assessing the richness of perspectives. Results suggest that generated comments generally outperform real ones, especially in logicality. Case analysis revealed that real comments often scored lower due to noise, topic deviation, brevity, incomplete arguments, insufficient analysis, and a lack of causal reasoning. Conversely, while generated comments excelled in all dimensions, they lacked the authenticity found in the imperfections of real comments.

4.7 A Case Study (EQ4)

Figure 3 presents a case study of a controversial video, detailing the evaluation results in the reasoning module. For the *controversy of video* aspect, the content potentially involves a labor controversy. In the *controversy between video and comments* aspect, it highlights critical comments questioning the supermarket's regulations or expressing sympathy towards customers, accompanied by strong emotions. Regarding *controversy among comments*, it detects heated exchanges with clear support and opposition. A comprehensive summary of all three aspects is also provided.

The comprehensive output report enhances the interpretability of multimodal controversy detection tasks and addresses the three key questions posed in Figure 1: (1)*"What is controversial?"*: The evaluation reveals that the controversy scores for Aspect 2 and Aspect 3 are 7 and 8, respectively, indicating that the controversy arises from the interaction between video content and comments, as well as among the comments themselves. (2)*"Why is it controversial?"*: This is explained through the judgment basis provided in the evaluation results. (3)*"How controversial is it?"*: The degree of controversy is quantified by the controversy scores in the evaluation results.

5 Conclusion

We propose AgentMCD, a comprehensive framework that leverages a multi-agent mechanism to simulate user engagement and the formation of controversies, along with a multi-aspect reasoning process to assess and interpret these controversies. This study explores the application of LLMs to understand the emergence of controversies during video dissemination, enhancing the explainability of traditional multimodal controversy detection tasks. Empirical evaluations reveal that AgentMCD significantly outperforms existing LLM-based approaches. Moreover, the approach of role-playing to enhance interpretability, as discussed in this paper, can be applied to other domains, such as fake news detection, as well as scenarios involving collaboration and competition.

Ethical Statement

As discussed, LLMs may occasionally produce irrelevant or harmful outputs, necessitating caution when interpreting their results. In our approach, LLM-based multi-agent systems are employed solely to enhance the simulation of controversy formation. However, additional research is required for language models intended for practical applications to refine prediction accuracy and bolster the model's authenticity and safety, thereby mitigating potential user risks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China, No. 62176137; SMP-IDATA Open Youth Fund, SMP2023-iData-007; Fundamental Research Funds for the Central Universities, NO. NJ2024029.

References

[Bang *et al.*, 2023] Yejin Bang, Samuel Cahyawijaya, and Nayeon Lee et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *IJCNLP-AACL*, pages 675–718. ACL, 2023.

[Beelen *et al.*, 2017] Kaspar Beelen, Evangelos Kanoulas, and Bob van de Velde. Detecting controversies in online news media. In *SIGIR*, pages 1069–1072. ACM, 2017.

[Benslimane *et al.*, 2023] Samy Benslimane, Jérôme Azé, and Sandra et al. Bringay. A text and GNN based controversy detection method on social media. *World Wide Web*, 26(2):799–825, 2023.

[Calvo Figueras *et al.*, 2023] Blanca Calvo Figueras, Asier Gutiérrez-Fandiño, and Marta Villegas. Anticipating the debate: Predicting controversy in news with transformer-based NLP. *Procesamiento del Lenguaje Natural*, pages 123–133, 2023.

[Canute *et al.*, 2023] Matt Canute, Mali Jin, and holtzclaw et al. Dimensions of online conflict: Towards modeling agonism. In *EMNLP*, pages 12194–12209. ACL, 2023.

[Chen *et al.*, 2024] Guangyao Chen, Siwei Dong, and Yu Shu et al. Autoagents: A framework for automatic agent generation. In *IJCAI*, pages 22–30. ACM, 2024.

[Cheng *et al.*, 2023] Myra Cheng, Tiziano Piccardi, and Diyi Yang. Compost: Characterizing and evaluating caricature in LLM simulations. In *EMNLP*, pages 10853–10875. ACL, 2023.

[Conover *et al.*, 2011] Michael D. Conover, Jacob Ratkiewicz, Matthew R. Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *ICWSM*. The AAAI Press, 2011.

[Dori-Hacohen and Allan, 2015] Shiri Dori-Hacohen and James Allan. Automated controversy detection on the web. In *European Conference on IR Research*, pages 423–434. Springer, 2015.

[Dori-Hacohen *et al.*, 2016] Shiri Dori-Hacohen, David D. Jensen, and James Allan. Controversy detection in wikipedia using collective classification. In *SIGIR*, pages 797–800. ACM, 2016.

[Du *et al.*, 2024] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *ICML*, 2024.

[Gan *et al.*, 2023a] Tian Gan, Qing Wang, Xingning Dong, Xiangyuan Ren, Liqiang Nie, and Qingpei Guo. Cnvid-3.5m: Build, filter, and pre-train the large-scale public chinese video-text dataset. In *CVPR*, pages 14815–14824. IEEE, 2023.

[Gan *et al.*, 2023b] Tian Gan, Xiao Wang, Yan Sun, Jianlong Wu, Qingpei Guo, and Liqiang Nie. Temporal sentence grounding in streaming videos. In *ACM MM*, pages 4637–4646. ACM, 2023.

[Garimella *et al.*, 2018] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACL Trans. Soc. Comput.*, 1(1):1–27, 2018.

[GLM, 2024] Team GLM. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *ArXiv*, 2024.

[Hamad *et al.*, 2018] Mhd Mousa Hamad, Marcin Skowron, and Markus Schedl. Regressing controversy of music artists from microblogs. In *ICTAI*, pages 548–555. IEEE, 2018.

[Hessel and Lee, 2019] Jack Hessel and Lillian Lee. Something's brewing! early prediction of controversy-causing posts from discussion features. In *NAACL*, pages 1648–1659. ACL, 2019.

[Hong *et al.*, 2024] Sirui Hong, Mingchen Zhuge, and Jonathan Chen et al. Metagpt: Meta programming for A multi-agent collaborative framework. In *ICLR*, 2024.

[Jiao *et al.*, 2023] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt A good translator? A preliminary study. *ArXiv*, 2023.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

[Koncar *et al.*, 2021] Philipp Koncar, Simon Walk, and Denis Helic. Analysis and prediction of multilingual controversy on reddit. In *Proceedings of the ACM Web Science Conference*, page 215–224. ACL, 2021.

[Li *et al.*, 2023] Zihan Li, Jian Zhang, Qi Xuan, Xiang Qiu, and Yong Min. A novel method detecting controversial interaction in the multiplex social comment network. *Frontiers in Physics*, 10:1107338, 2023.

[Liang *et al.*, 2024] Tian Liang, Zhiwei He, and Wenxiang Jiao et al. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*, 2024.

[Linmans *et al.*, 2018] Jasper Linmans, Bob van de Velde, and Evangelos Kanoulas. Improved and robust controversy detection in general web pages using semantic approaches under large scale conditions. In *CIKM*, pages 1647–1650. ACM, 2018.

[Liu *et al.*, 2023] Ruibo Liu, Ruixin Yang, and Chenyan Jia et al. Training socially aligned language models in simulated human society. *ArXiv*, 2023.

[Liu *et al.*, 2024] Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. In *IJCAI*, pages 7886–7894. ACM, 2024.

[Madaan *et al.*, 2023] Aman Madaan, Niket Tandon, and Prakhar Gupta et al. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

[Martin, 2014] Brian Martin. *The controversy manual*. Irene Publishing, 2014.

[Mejova *et al.*, 2014] Yelena Mejova, Amy X. Zhang, Nicholas Diakopoulos, and Carlos Castillo. Controversy and sentiment in online news. *CoRR*, 2014.

[Mendoza *et al.*, 2020] Marcelo Mendoza, Denis Parra, and Álvaro Soto. GENE: Graph generation conditioned on named entities for polarity and controversy detection in social media. *Information Processing & Management*, 57(6):102366, 2020.

[Newman *et al.*, 2023] Nic Newman, Richard Fletcher, Kirsten Eddy, Craig T Robertson, and Rasmus Kleis Nielsen. Reuters institute digital news report 2023. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023, 2023. Accessed: 2025-05-20.

[Park *et al.*, 2023] Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *UIST*, pages 2:1–2:22. ACM, 2023.

[Popescu and Pennacchiotti, 2010] Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. In *CIKM*, pages 1873–1876. ACM, 2010.

[Sanh *et al.*, 2022] Victor Sanh, Albert Webson, and Colin Raffel et al. Multitask prompted training enables zero-shot task generalization. In *ICLR*, 2022.

[Shanahan *et al.*, 2023] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

[Shinn *et al.*, 2023] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *NeurIPS*, 2023.

[Sun *et al.*, 2023] Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. Recitation-augmented language models. In *ICLR*, 2023.

[Tarrow, 2008] Sidney Tarrow. Polarization and convergence in academic controversies. *Theory and Society*, 37:513–536, 2008.

[Wang *et al.*, 2022] Xiao Wang, Tian Gan, Yinwei Wei, Jianlong Wu, Dai Meng, and Liqiang Nie. Micro-video tagging via jointly modeling social influence and tag relation. In *ACM MM*, pages 4478–4486. ACM, 2022.

[Wang *et al.*, 2023a] Haiyang Wang, Ye Wang, Xin Song, Bin Zhou, Xuechen Zhao, and Feng Xie. Quantifying controversy from stance, sentiment, offensiveness and sarcasm: a fine-grained controversy intensity measurement framework on a chinese dataset. *World Wide Web*, 26(5):3607–3632, 2023.

[Wang *et al.*, 2023b] Lei Wang, Wanyu Xu, and Yihuai Lan et al. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *ACL*, pages 2609–2634. ACL, 2023.

[Wang *et al.*, 2023c] Xiao Wang, Yaoyu Li, and Tian et al. Gan. RTQ: Rethinking Video-language Understanding Based on Image-text Model. In *ACM MM*, pages 557–566, 2023.

[Wang *et al.*, 2023d] Xuezhi Wang, Jason Wei, and Dale Schuurmans et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

[Wu *et al.*, 2023] Qingyun Wu, Gagan Bansal, and Jieyu Zhang et al. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *ArXiv*, 2023.

[Xu *et al.*, 2023] Haiyang Xu, Qinghao Ye, and Xuan et al. Wu. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *ArXiv*, 2023.

[Xu *et al.*, 2024] Tianjiao Xu, Aoxuan Chen, Yuxi Zhao, Jinfei Gao, and Tian Gan. A chinese multimodal social video dataset for controversy detection. In *ACM MM*. ACM, 2024.

[Yao *et al.*, 2024] Shunyu Yao, Dian Yu, and Jeffrey Zhao et al. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Zhang *et al.*, 2023] Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. Cumulative reasoning with large language models. *ArXiv*, 2023.

[Zhang *et al.*, 2024] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *SIGIR*, pages 1807–1817. ACM, 2024.

[Zhong *et al.*, 2020] Lei Zhong, Juan Cao, Qiang Sheng, Junbo Guo, and Ziang Wang. Integrating semantic and structural information with graph convolutional network for controversy detection. In *ACL*, pages 515–526. ACL, 2020.

[Zhu *et al.*, 2023] Xizhou Zhu, Yuntao Chen, and Hao Tian et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *ArXiv*, 2023.