VMD-FACT: A New Video Dataset and MLLM-based method for Detecting Realistic AI-Generated Video Misinformation

Supplementary Material

Overview

- Section 1 presents general LLMs' performance on claim manipulation detection in RAVM.
- Section 2 presents general MLLMs' performance on cross-modal manipulation detection in RAVM.
- Section 3 presents more visualization results.
- Section 4 provides the narrative-driven template.
- Section 5 shows the detailed prompt contents.

1. Experiments on Claim Manipulation

Since RAVM incorporates intent polarity when manipulating claims, it allows us to evaluate state-of-the-art general LLMs on both claim authenticity and intent polarity detection. As shown in Table S1, existing LLMs perform poorly on both tasks. Authenticity detection is particularly challenging, indicating that the manipulated claims are highly realistic and difficult for current LLMs to recognize directly. Even the advanced model DeepSeek-V3.2-Exp [1] achieves only 57.28% and 67.94% accuracy on authenticity and intent polarity, respectively, further demonstrating the high deceptiveness and substantial challenge posed by claims in RAVM. Since the intent polarity of a claim does not correspond one-to-one with the authenticity of a claim–video pair, even if a claim–video pair is labeled as false, its intent polarity may still be unharmful. Therefore, although InternVL3-9B [7], Qwen2.5-7B [4], and Qwen3-8B [5] achieve an intent polarity detection accuracy of around 71%, this capability cannot be directly used to determine the authenticity of claim–video pairs.

2. Experiments on Cross-Modal Manipulation

In the RAVM dataset, most samples are highly realistic. For the vast majority of samples—except those with semantic inconsistencies between claim and video introduced via data augmentation or existing datasets—the claim and video exhibit strong semantic consistency. In such cases, relying solely on multimodal analysis makes it difficult to identify key evidence for determining the authenticity of a claim–video pair. Therefore, attribution analysis, *i.e.*, *multimodal analysis* and *fact-checking*, is required. We further provide attribution annotations for each claim–video pair. Table S2 presents the performance of general MLLMs on attribution and intent polarity detection for claim–video pairs. The Qwen3-VL [5] series models generally perform poorly on attribution detection. The strongest closed-source MLLM, Gemini 2.0 [2], achieves only 74.25% accuracy in attribution detection. However, as shown in Table 2 of the main paper, its accuracy on claim–video authenticity detection in the RAVM dataset is only 68.07%. Meanwhile, for intent polarity detection of claim–video pairs, the highest accuracy achieved by existing general MLLMs is merely 68.89%. These results indicate that the RAVM dataset is highly challenging and that determining the authenticity of claim–video pairs cannot rely on a single metric or aspect alone. Moreover, individual tasks such as attribution or intent polarity detection remain inherently difficult.

Methods	Authenticity	Intent Polarity
InternVL3-1B [7]	49.38	65.57
InternVL3-8B [7]	59.93	69.70
InternVL3-9B [7]	63.25	71.53
Qwen2.5-7B [4]	64.56	71.54
Qwen3-8B [5]	64.72	71.36
DeepSeek-V3.2-Exp [1]	57.28	67.94

Table S1. Performance of general LLMs on claim manipulation detection in RAVM.

Methods	Attribution	Intent Polarity
Qwen3-VL-2B-Instruct [5]	19.21	47.13
Qwen3-VL-4B-Instruct [5]	22.68	67.80
Qwen3-VL-8B-Instruct [5]	33.97	66.69
InternVL3.5-8B [3]	22.99	68.89
VideoLLaMA3-7B [6]	64.37	45.61
Gemini 2.0 [2]	74.25	62.86

Table S2. Performance of general MLLMs on cross-modal manipulation detection in RAVM.

3. More Visualizations

We present additional claim-video pairs from the RAVM dataset, as shown in Figure S1 and Figure S2.

Generated Claim-Video Pairs. We first show claim–video

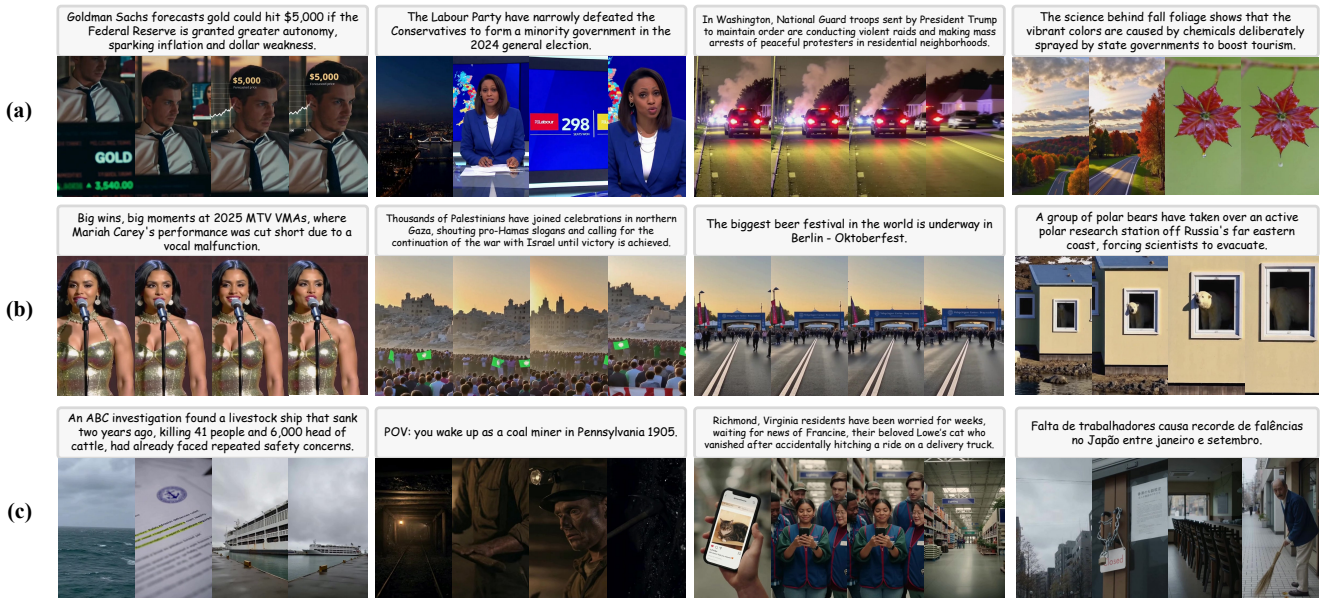

Figure S1. Visualization of more cases from the RAVM dataset. The claim–video pairs in rows (a) and (b) are labeled as **fake**, while the claim-video pairs in row (c) are labeled as **real**. Moreover, the videos in rows (a), (b), and (c) are all **generated**.

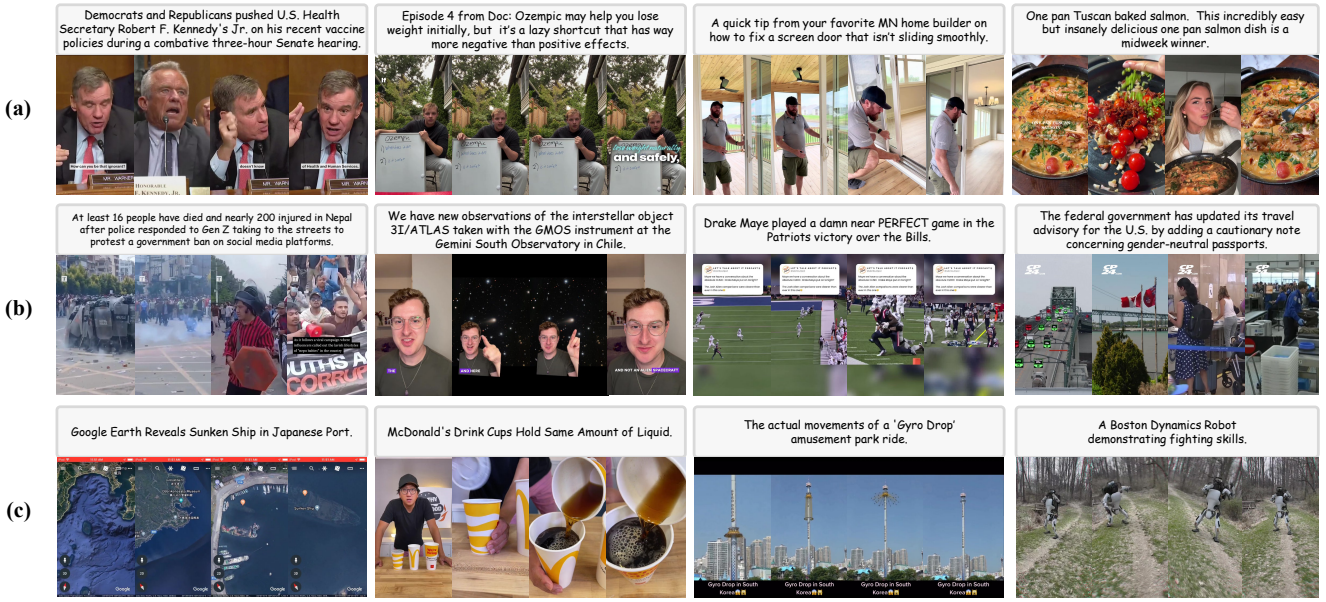

Figure S2. Visualization of more cases from the RAVM dataset. The claim–video pairs in rows (a) and (b) are labeled as **real**, while the claim-video pairs in row (c) are labeled as **fake**. Moreover, all the claim–video pairs in rows (a), (b), and (c) are **originally collected**.

pairs that contain video manipulation, as illustrated in Figure S1. Specifically, the claim–video pairs in rows (a) and (b) are labeled as **fake**, while those in row (c) are labeled as **real**. All videos in rows (a), (b), and (c) are generated.

Originally Collected Claim-Video Pairs. We show the originally collected claim–video pairs in Figure S2, where the samples in rows (a) and (b) are labeled as **real**, and the samples in row (c) are labeled as **fake**.

4. Narrative-Driven Template

To guide the Prompt Generator in producing semantically aligned manipulation prompts, which in turn direct the Video Generation Model Library to generate realistic initial videos, we introduce a narrative-driven template, whose content is shown below.

Narrative-Driven Template

<Perspective>
First-person perspective / Third-person perspective
<Scene>
(1) Spatiotemporal context (Where and when the event takes place)
(2) Time: daytime / night / dusk / dawn
(3) Location: street / city hall / disaster zone / border checkpoint / hospital / studio ...
(4) Event background: rescue operation / protest site / live broadcast / war front ...
(5) Environmental attributes (Visual and physical characteristics of the scene)
(6) Physical state: chaotic / in ruins / calm / crowded / ablaze / filled with smoke ...
(7) Spatial features: skyscrapers / collapsed buildings ...
(8) Background elements: police cars / banners / smoke / debris / barricades / crowds ...
(9) Social context / atmosphere (The emotions or social tensions conveyed by the scene)
(10) Emotional atmosphere: tense / panicked / angry / chaotic / solemn / peaceful ...
(11) Social state: confrontation / turmoil / orderly / public anxiety ...
(12) Details: (Additional descriptive elements of the scene)
<Subject>
(1) Attributes: age / gender / ethnicity / appearance / clothing / attire . accessories ...
(2) Behavior: interactions with the environment or other subjects (dialogue, actions, etc.)
(3) Emotion: anger / fear / pain / calmness / anticipation / sarcasm / sadness / shock ...
(4) Stance (Direct expression of the creator's intention): support / oppose / neutral / question / affirm / mock / warn / appeal / claim ...
(5) Details: (Additional descriptive elements of the subject)
<Narrative>
Pay attention to the overall logic of the narrative.
<Cinematography>
Camera movements and transitions: pan left / pan right / tilt up / tilt down / zoom in / follow shot / Speed of movement ...

5. Prompts

In this section, we present some of the prompts used in the AI-generative framework.

Prompt for Claim Manipulation

You will receive a <claim> and a <description> . Your task consists of two steps:
Step 1 — Forge the <claim> (must strictly choose only one strategy) From the list of forgery strategies below, choose only one and use only that one strategy to modify the original claim. The modification must introduce falsification, inaccuracy, or a counterfactual element, rather than merely replacing words with synonyms. Do not combine multiple strategies. Additionally, adaptively select the single strategy that is most appropriate for the specific input claim and will produce the strongest forgery effect.
Possible strategies:
(1) Replace or modify entities in the claim (people, places, organizations, dates, etc.) in a way that alters factual truth;
(2) Change the narrative structure or event logic of the claim (e.g., alter causality, sequence, or motives) to create a false or misleading statement;
(3) Fine-tune the claim to create counterfactual statements (making statements that are opposite or inconsistent with reality);
(4) Fine-tune or rewrite the claim to create specific emotions, feelings, or biases (e.g., exaggeration, provocation, fear, or undue optimism) that manipulate perception or plausibility;
(5) Adjust the claim for a specific purpose (e.g., to influence public opinion, create panic, downplay responsibility, or mislead) in a way that introduces falsehood.
(6) Important constraints: Each forgery must choose only one of the strategies above; do not mix multiple strategies;
The modified claim should maintain some similarity with the original claim (theme or structure), but must clearly introduce falsification or counterfactual content; Avoid mere synonym replacement or minor stylistic changes; the goal is to produce a claim that would be considered untrue or misleading;
Do not include explanations about which strategy you chose in the output (unless otherwise requested).
Based on the above analysis, determine the explicit manipulation intent (harmful/unharmful) and generate the final manipulated claim accordingly.
Step 2 — Reconstruct the <description> Based on the modified <claim> and referencing the style, information density, and structure of the original <description> , generate a new <description> that sup-

ports the forged <new claim> and makes it appear coherent and plausible. The output <claim> and <description> must completely remove all tags (e.g., #XXX) and must not include any words starting with # under any circumstances.
The output <claim> and <description> should not contain any tags, such as #XXX Output format (must strictly follow, only include the following two lines):
<new claim> XXX
<new description> XXX
The claim is: {claim}
The description is: {description}

Prompt for Alignment Evaluator

Here is a video, and a claim: {claim}
You are a multimodal semantic alignment evaluator. Please assess the core semantic consistency between the following claim and video, noting the following:
The claim is a rewritten statement that may not include specific visual details present in the video (e.g., clothing, background color, brand of objects).
The video was generated based on this claim and its associated description, so it may contain additional visual elaborations.
Your task is not to check whether every detail matches exactly, but rather to determine: Does the video reasonably and plausibly depict the core event or state described in the claim?
Please consider the following aspects:
Is the key action, state, or event in the claim visually presented or demonstrated in the video?
Are there any clear visual contradictions that directly violate the claim?
Even if the presentation differs (e.g., different setting or characters), can the video still be interpreted as supporting or illustrating the claim?
Output format:
<alignment_score>
(an integer from 0 to 10, where 10 means highly consistent and 0 means completely contradictory)
<reason>
(Briefly explain your judgment, focusing on whether the core event is substantiated)

Prompt for Quality Evaluator

What visual quality issues does this video have? Please be specific about the entities in the video where you think problems exist.
Output format:
<score> (Video quality score, minimum 0, maximum 10, can keep 1 decimal place) </score>
<reasoning> (The issues you think are present) </reasoning>

Prompt for Adversarial Evaluator

You are an experienced short video news analyst, skilled in assessing the authenticity of news videos using multimodal evidence. Given a short video and its claim, your task is to carefully analyze the visual, textual, and audio modalities, extract key evidence from each, and reason step by step to determine whether the video conveys real or fake news.
Use both unimodal and cross-modal reasoning to assess the internal consistency of the evidence, the alignment across modalities, the plausibility of the content in relation to real-world knowledge, and any signs of emotional or intentional manipulation.
Conclude your analysis with a clear justification and a final verdict.
The claim is: {claim}.
Avoid analyzing the claim in isolation or making one-sided judgments based solely on the claim.
You must give a clear answer.

Prompt for Optimizer

This is a text-to-video generated fake video, created to align with a manipulated claim and mislead viewers.
The manipulated claim is: {claim}
The description, which explains the claim briefly: {description}
The original prompt guiding the text-to-video model to generate a video highly consistent with the claim: {original_prompt}
Video evaluation results:
<Quality score> : {quality_score}
<Quality problem> : {quality_problem}
<Alignment score> : {alignment_score} (Semantic alignment score between the claim and the video's main argument (narrative), ranging from 0 (lowest) to 10 (highest).)
<Alignment problem> : {alignment_problem}
<MIS score> : {mis_score} (Whether the Multimodal Misinformation Detection model classifies the above claim+video as fake (1 = yes, 0 = no), with the focus of the model being on visual/textual evidence and inconsistencies or contradictions within and across modalities.)
<Prompt optimization history> : {history}

Your task:
Analyze the claim, description, video details, evaluation results, and prompt optimization history carefully, and refine the original prompt to guide the model in generating a new video that addresses the <Quality problem> while maximizing <Quality score> , <Alignment score> and <MIS score> .
Output format:
Optimized prompt only (Be consistent with the above original prompt format);
No additional text. You must output according to the specified format!

Prompt for Perceiver

These are two images, intended as the first and last frames for an image-to-video generation. <start> <end>
Here is a claim: {claim}
Here is a detailed description of the claim: {description}
Both the claim and description are manipulated.
<your task> Analyze the claim and description in detail, and provide instructions on how to modify image1 and image2 to be used as the first and last frames of the video so that the generated video is semantically highly aligned with the claim and description. Output format:
<start> XXX (editing instructions for the first frame)
<end> XXX (editing instructions for the last frame)
The editing instructions should be concise but cover the key points.
Only essential elements of start and end frames should be edited, such as background replacement, characters, slogans, fonts, objects, removal, or substitution.

Prompt for Semantic Perceiver

This is a claim: {claim}
The following is a text-to-video generated clip without audio.
<Your task>
(1) Analyze the claim and the video to decide whether background music is needed.
(2) Determine whether speech should be added. If yes, provide the speech content ($\leq$ 40 words) and specify the emotion of delivery. Available emotions: neutral, happy, sad, angry, excited, fearful, disgusted, surprised, calm, serious.
(3) The speech should align with the core meaning of the claim and support the video context, either as narration or as a character's line.
<Output format>
<music> YES/NO </music>
<speech> NO </speech> or (if speech should be added)
<speech>
(Speaker's information: Gender: male/female; Emotion: neutral/happy/sad/...; Speed: fast/slow) [Speech content here]
</speech>
Output example:
<music> YES </music>
<speech>
(Speaker's information: Gender: female; Emotion: sad; Speaking speed: fast)
The evening sun cast long shadows across the tech blogger's desk, illuminating the glow of his monitor. He stared at the leaked internal documents, the words significant security flaw and unauthorized access burning into his retinas.
</speech>
You must strictly follow the output format and produce no additional text.

References

[1] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025. 1

[2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[3] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1

[4] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1

[5] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan

Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1

[6] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 1

[7] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1