Explainable AIGC Misinformation Video Identification Via Vision-based Language Models With Human-initiated Insights

Shuning Zhang[1] , **Yiqun Xu**[1] , **Ruokai Zhao**[2] , **Yuanyuan Wu**[3] , **Yabo Wang**[1] , **Linzhi Wang**[1] , **Simin Li**[4] , **Xin Yi**[1*] , **Hewu Li**[1]
[1]Tsinghua University
[2]Oxford University
[3]Shanghai Jiao Tong University
[4]Beihang University
zhang.sn314@gmail.com, xujixyq@163.com, ruokai.zhao@st-hughs.ox.ac.uk,
buddy.yuan@sjtu.edu.cn, {yb-wang22, wang-lz22}@mails.tsinghua.edu.cn,
lisiminsimon@buaa.edu.cn, yixin@tsinghua.edu.cn, lihewu@cernet.edu.cn

Abstract

Identifying AIGC-generated misinformation videos on social media remains a formidable challenge for purely automated methods. In contrast, human observers contribute distinctive insights into the specific regions of a video that exhibit telltale artifacts. Moreover, Vision-Language Models (VLMs) have demonstrated considerable promise in capturing the semantic content of videos. In this work, we present an initial investigation of how to leverage human-elicited insights, in conjunction with VLMs, to detect misinformation videos on social platforms. First, we conducted in-depth interviews to elicit the aspects of video content that users find most suspicious. Building on these findings, we introduce a structured prompting framework that integrates multiple perspectives derived from human insights. Experimental results across diverse misinformation video benchmarks and VLM architectures show that our approach consistently outperforms baseline prompting strategies. For instance, on the UADFV dataset, our method yields a 31.64 % increase in accuracy when applied to the Gemini model. Ablation studies further validate the individual contribution of each category of human insight.

1 Introduction

Detecting misinformation videos generated by AI remains a formidable challenge, owing to the increasing sophistication of content synthesis techniques [Papadopoulos *et al.*, 2023], the rapid proliferation and evolution of such videos across platforms [Xu *et al.*, 2023], and the heterogeneous nature of misinformation types [Micallef *et al.*, 2022]. Conventional solutions have largely depended on dedicated forensic video identification models [Xu *et al.*, 2023], which nonetheless frequently exhibit suboptimal accuracy and offer limited interpretability [Xu *et al.*, 2023]. In light of these limitations and obstacles, Vision-Language Models (VLMs) have emerged as a compelling alternative for discerning semantic inconsistencies in video data [Tahmasebi *et al.*, 2024], although the question of how best to construct prompts for VLM-based detection remains unresolved.

Humans exhibit a remarkable and surprising sensitivity to AIGC misinformation videos, often detecting them through the recognition of distinct visual and contextual artifacts [Joslin *et al.*, 2024]. These innate abilities provide valuable insights for the design of robust AI-driven detection systems, particularly by enhancing interpretability. In contrast, fully automated approaches frequently struggle to capture such subtle cues, leading to limited explainability. This motivates our core research question:*How can human-derived artifact identification insights be systematically integrated to improve both the accuracy and interpretability of VLMs for AIGC misinformation video detection?*

Our research methodology comprises two sequential stages. First, we carried out theory-driven user interviews to systematically elicit and analyze the heuristics and cues by which individuals distinguish misinformation videos. Second, we designed, implemented, and evaluated an explainable prompting framework for VLMs grounded in the human-derived insights obtained from these interviews.

Through our foundational analysis, we distilled three main principal dimensions that users employ when assessing video authenticity, each corresponding to a distinct cognitive heuristic: *Content and narrative coherence* (e.g., storyline plausibility, narrative complexity), *Visual detail and realism* (e.g., physical consistency, rendering artifacts), *Image integration and composition* (e.g., element fusion, artificial layering). These three dimensions serve as the pillars of our systematic prompting framework.

Building on these dimensions, we propose a systematic prompting framework that decomposes detection into three targeted subqueries corresponding to our identified human insights. By explicitly aligning each prompt with human-derived heuristics, our framework yields interpretable diagnostic axes.

*Corresponding author.

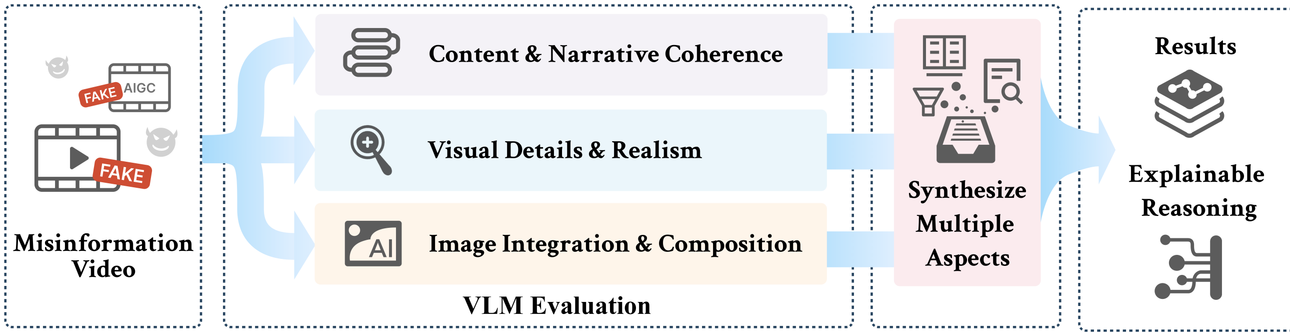

Figure 1: The algorithm framework for this paper.

We evaluated our human-insight-driven prompting methods across several benchmark datasets spanning from deepfakes to text-to-video generations, including GenVidBench [Ni *et al.*, 2025], T2VQA-DB [Kou *et al.*, 2024], Celeb-DF-v2 [Li *et al.*, 2020], FaceForensics++ [Rossler *et al.*, 2019] and UADFV [Li *et al.*, 2018], utilizing state-of-the-art VLMs including qwen-vl-max, gpt-4o, claude-3-7-sonnet and qvq. Our result demonstrated a significant improvement in misinformation video identification accuracy. For instance, with the Gemini model, accuracy on the UADFV dataset improved by 31.64% (from a baseline of 0.2857 to 0.6021 with our method). Furthermore, our ablation studies underscored the general effectiveness of human insight categories. On the FaceForensics++ dataset, for example, where a strong synergistic effect was observed, each of the three core insight categories (Content & Narrative Coherence, Visual Details & Realism, and Image Integration & Composition) was found to contribute positively. The removal of any single category degraded accuracy by at least 19.17%. To sum up, the primary contributions of this paper are:

- Derived key human insights, specifically narrative coherence, visual realism, and image composition that facilitates discerning misinformation videos.
- Proposed an explainable prompting framework that hierarchically integrates these human-derived insights to guide video misinformation analysis.
- Empirically established the superiority of this insight-driven prompting over naive methods.

2 Background & Related Work

The proliferation of manipulated videos necessitates advanced identification strategies. We survey key research concerning human cognitive and perceptual responses to manipulated visual content, the evolution of automated detection, particularly via LLMs, and the development of explainable AI for these visual systems.

2.1 Human Reception and Evaluation of Manipulated Video Content

The human visual system exhibits a notable specialization for face recognition [Morrisey *et al.*, 2019; Allison *et al.*, 1994], enabling extraction of facial traits and identification of differences crucial for social interaction [Morrisey *et al.*, 2019]. This inherent sensitivity to facial information, combined with the strong attentional draw of faces [Walther *et al.*, 2001], significantly influences online decision-making in contexts ranging from social connections [Rashtian *et al.*, 2014; Bakhshi *et al.*, 2014] to commercial transactions [Dai *et al.*, 2018; Banerjee *et al.*, 2022]. Consequently, facial imagery frequently becomes a prime target for manipulation in misinformation campaigns designed to illicitly gain user trust.

Despite specialized visual skills, humans often struggle to detect subtle video manipulations without guidance, as judgments are shaped by cognitive biases [Shahid *et al.*, 2022] and deceptive contexts like fake social profiles [Ruffin *et al.*, 2024]. Yet, human evaluations uniquely draw on narrative plausibility, source credibility, and visual inconsistencies. Integrating these holistic strategies [Seng *et al.*, 2020] into forensic tools is key to supporting human judgment.

2.2 Identifying Multimodal Misinformation through Large Models

Recent advances in generalized face attack detection have introduced several promising approaches. FA^3 -CLIP [Li *et al.*, 2025] demonstrates effective contrastive learning frameworks for cross-domain face anti-spoofing, while mixture-of-experts (MoE) based methods [Kong *et al.*, 2024] have shown superior performance by specializing different expert networks for various manipulation types. Particularly relevant is the work on interpretable face anti-spoofing that enhances generalization with multimodal large language models [Zhang *et al.*, 2025], which shares our goal of combining human interpretability with advanced model architectures.

Hybrid architectures have also been prominent, combing VLMs with LLMs through knowledge-guided adaptation and multi-turn prompting for explainable detection [Yu *et al.*, 2025], or integrating small and large language models to leverage their complementary strengths [Wang *et al.*, 2024]. Further innovations include zero-shot detection pipelines incorporating evidence re-ranking and fact verification [Tahmasebi *et al.*, 2024], and methodologies like MDAM that scrutinize intra-modal features and inter-modal inconsistencies [Xu *et al.*, 2025], with comprehensive reviews mapping this dynamic landscape and future directions [Abdali *et al.*, 2024]. Different from these approaches, our work proposed a structured, insight-driven prompting strategy to elicit consistent, informative misinformation identification, moving towards explainable detection.

2.3 Explainable Methods in Misinformation Videos' Identification

Enhancing the explainability of models for identifying misinformation videos has become a growing focus in recent research, as it plays a crucial role in fostering user trust and supporting model diagnostics. One line of work centers on prototype-based explanations, where model decisions are linked to representative visual exemplars. For instance, ProtoExplorer [de Leeuw den Bouter *et al.*, 2023] allows interactive exploration and refinement of a deepfake model's prototypes, while other methods leverage learned multimodal prototypes to improve robustness against unseen manipulations [Pellicer *et al.*, 2024]. Another direction involves feature attribution and localization, employing post-hoc techniques like LIME to highlight influential image regions [Tsigos *et al.*, 2024], applying network dissection to map CNN units to semantic facial concepts [Mansoor and Iliev, 2025], or extracting localized forensic features from global video representations [Soltandoost *et al.*, 2025]. A third category focuses on generating natural language justifications, such as the TruthLens framework [Kundu *et al.*, 2025], which uses a vision backbone and a multimodal language model to answer fine-grained queries and provide detailed textual reasoning. While diverse approaches aim to enhance the transparency of complex detection methods, our insight-guided prompting offers a direct route to explainability by employing a hierarchical, human-centric approach.

3 Developing Identification Strategies from Human Insights

To gather insights on what users cared about and from which perspective would users discern misinformation videos, we conducted a deep user interview.

3.1 Experiment Setup and Analysis

To examine user perceptions and detection strategies for AIGC-generated misinformation videos, we conducted one-on-one, semi-structured interviews with fourteen participants (5 male, 9 female; M = 22.9 years, SD = 2.9), recruited via online chat groups until thematic saturation [Aldiabat and Le Navenec, 2018]. Thematic saturation was determined when no new insights or themes emerged from consecutive interviews, specifically when the final interviews (participants 12-14) yielded no novel categories beyond the three primary dimensions identified from earlier participants. Participants varied in background: four held Ph.D.s, nine bachelor's degrees, and one completed high school. Each session involved discussing influential factors for detecting misinformation and demonstrating their application, supported by example videos (Section 5.2) and reflections on prior encounters with AIGC content. Interviews were audio-recorded, transcribed verbatim, and thematically analyzed [Clarke and Braun, 2017]. Two researchers independently coded transcripts, resolving differences through discussion, yielding a final framework with an inter-coder reliability of 0.90.

3.2 Results

Participants predominantly cited three dimensions: *content and narrative coherence*, *visual detail and realism*, and *image integration and composition.*

Content & Narrative Coherence: A substantial majority of participants (9/14) emphasized the structural and semantic integrity of video narratives when suspecting AI generation, noting examples of simplistic, formulaic, or superficial content that lacked expected depth (P1, P2, P7). Moreover, 7 of 14 participants specifically evaluated logical and temporal coherence: they assessed whether the content was *"reasonable and logical"* (P3), probed for internal inconsistencies (P7), or flagged instances of *"chaotic logic"* (P6, P9, P12) or *"absurd"* plot developments (P7, P8, P11).

Video length emerged as a salient factor for six of fourteen participants, eliciting mixed perspectives. Three participants (P3, P5, P11) reported that longer videos appeared more suspicious, whereas another three (P6, P9, P12) argued that extended duration can facilitate detection when narrative coherence deteriorates.

Visual Details & Realism: Participants placed substantial emphasis on detailed examination of visual artifacts, identifying inconsistencies in physical realism, material rendering, and object appearance as hallmarks of AI-generated manipulations. Notably, seven of fourteen participants explicitly pointed to visual or technical anomalies—such as irregular textures or implausible lighting—as key indicators. Illustrative feedback ranged from general remarks regarding AI's *"technical bottleneck"* leading to visible flaws (P1) and instances where *"relevant details are not handled well"* (P13), to precise mentions of *"unnaturalness in the background, facial expressions, or other elements"* (P2).

Moreover, six of fourteen participants (P5, P6, P8) reported relying on distinct technical anomalies as diagnostic cues, such as audio–visual desynchronization (P6), misaligned lip movements (P5), and unnatural biomechanical motion like atypical facial muscle activation—collectively described as "bugs in details" (P5). Another six participants commented more broadly on video quality, AI tool maturity, and viewer experience: less advanced outputs were easier to detect (P3); discernment improved with familiarity via AI editing tools (P4); poor visual consistency reduced credibility (P9); AI could mimic flaws yet still show weaknesses (P5, P10); and hyper-realistic clarity could also raise suspicion (P14). Finally, contextual factors—such as subject matter (P2, P11) or overall video sharpness (P7)—influenced the depth of scrutiny.

Image Integration & Composition: Participants' awareness of AI's use in augmenting footage and generating scenes critically shaped their authenticity assessments, with perceived integration seamlessness being pivotal. A notable group (7/14) reported observing flaws in visual element combination, including issues with blending, seams, manipulation artifacts, or compositional integrity. These users looked for anomalies in how elements like swapped faces or artificial backgrounds blended (P2); applied familiarity with AI

Table 1: Human insights which facilitated the hierarchical prompting strategy.

Category	Detailed Insights	VLM Evaluation Directive	User Study Basis
1. Content & Narrative Coherence	Lack of Complexity / Storyline Blandness	Evaluate if the storyline appears unusually bland or lacks expected complexity.	9/14 noted issues with overall narrative quality, like concerns about simplicity, formulaic structure, or lack of depth.
2. Visual Details & Realism	Material/Texture Unnaturalness	Assess if object materials or textures appear overly uniform, lack natural gradients, or are inconsistent with their real-world counterparts.	7/14 reported various visual or technical inconsistencies.
	Illogical Perspective Relationships	Examine perspective relationships for unscientific or illogical inconsistencies.	7/14 noted visual/technical inconsistencies.
	Computer-Generated Object Flexions	Inspect bends or curves in objects for hard-edged, computer-generated characteristics rather than natural flexions.	7/14 observed visual/technical flaws.
3. Image Integration & Composition	Artificial Element Separation / Boundaries	Analyze transitions and boundaries between foreground and background elements for overly clean, sharp or artificial-looking separations.	7/14 reported flaws in image or element integration.
	Modular or "Copy-Paste" Effects	Identify if figures or elements within the video exhibit significant modular characteristics, resembling a 'collage' or 'copy-paste' effect.	7/14 reported flaws in image or element integration.

techniques (e.g., face swaps, video splicing) to detect artifacts (P3, P5); or identified *"technical problems"* (P8, P11) and inconsistent behaviors (P9) indicating integration issues.

Perceived production value and the technical execution of compositional elements also significantly influenced trust for 6/14 participants (including P7, P8, P10, P14). Some, like P7, felt *"AI-generates fake videos seem cheaper,"* while P8 noted *"low production quality"* in element combination could foster suspicion. This sometimes linked to a perceived temporal limitation of current AI in maintaining coherent compositions, with 2 users (P9, P12) suggesting AI struggles beyond a few seconds. P10 cited lack of AI disclosure as decreasing trust, while P14 reported resorting to external validation (e.g., peers, reviews, platform labels) when intrinsic visual cues about integration were ambiguous.

4 Explainable Prompting Method

Built on the user study which identified key factors to discern AIGC misinformation videos, we developed an explainable prompting method. This method aims to systematically incorporate human-understandable insights into VLMs to enhance identification accuracy and provide interpretable results.

4.1 Prompt Strategy Based on Human Insights

We distilled the findings from user surveys and interviews into distinct categories of visual and contextual cues. These categories form the basis of our prompting strategy, where each category addresses a specific facet of potential AI-generated artifacts. The selected insights are:

Category 1: Content & Narrative Coherence This category assesses the plausibility and natural flow of the video's storyline and thematic elements. AI-generated content may exhibit unusually simplistic and disjointed narratives. The prompt focus on one core aspect *"Evaluate if the storyline appears unusually bland of lacks expected complexity, which can be indicative of AI generation."*

Category 2: Visual Details & Realism This prompts the VLM to scrutinize the fine-grained visual elements within the video, checking for consistency with real-world physics, materials and object morphology. AI models may falter in rendering these details accurately. The prompt insights focus on three aspects: *"Assess if object materials or textures appear overly uniform, lack natural gradients, or are inconsistent with their real-world counterparts."*, *"Examine perspective relationships within the video for unscientific or illogical inconsistencies."*, and *"Inspect bends or curves in objects, particularly looking for hard-edged, computer-generated characteristics rather than natural flexions."*

Category 3: Image Integration & Composition This category directs attention to how various visual elements and figures are combined within the frame and across scenes. AI generation can result in unnatural seams, layering or repetitive modularity. The prompt insights focus on two aspects: *"Analyze the transitions and boundaries between foreground and background elements for overly clean, sharp or artificial-looking separations.*, and *"Identify if figures or elements within the video exhibit significant modular characteristics, resembling a 'college' or 'copy-paste' effect."*

4.2 Prompt Flow Design

The prompting flow is designed to systematically query the VLM regarding each category of insights for a given video. This involves a hierarchical prompting structure where VLM queries each insight and finally returns the ensembled result. The goal is to elicit a reasoned "thought process" from the VLM, mirroring human-like multi-faceted critical analysis, before it arrives at a final identification decision. This structured approach contrasts with naive prompting methods by providing explicit, explainable dimensions for the VLM to consider. We employ a multi-tiered prompt architecture that follows a coarse-to-fine progression and is organized into three principal analytical frameworks corresponding to the above categories.

Prompt Engineering Methodology

Our prompt generation follows a systematic process: **(1) Insight Extraction:** Human insights from interviews are codified into evaluative criteria following established qualitative analysis protocols [Wei *et al.*, 2022]; **(2) Linguistic Formulation:** Each insight is translated into clear VLM directives using consistent grammatical structures; **(3) Validation:** Generated prompts are tested on a pilot set of videos to ensure reliable VLM responses and alignment with human insights through expert review.

5 Experiment Settings

We conducted an experiment including various models and datasets to validate the results. This section details the model, dataset and parameter settings.

5.1 Model Settings

We leverage state-of-the-art VLMs, including qwen-vl-max, glm-4v-plus, claude-3-7-sonnet, gemini, gpt-4o, gpt-4o-mini to assess the generalizability of the prompting methods. These models span across different architectures, brand, sizes and open/close source status.

5.2 Datasets

We utilize several benchmark datasets to evaluate our proposed methods. These datasets encompass various types of AI-generated videos, from T2V generations to deepfakes.

GenVidBench [Ni *et al.*, 2025] is a challenging contemporary benchmark for detecting AI-generated videos, featuring synthetic content from advanced text-to-video and video-to-video models such as Pika, Mora and SVD. It includes videos with diverse resolutions and frame rates

T2VQA-DB [Kou *et al.*, 2024] specifically targets text-to-video generations, incorporating content from ten distinct generation processes (nine unique models like Text2Video-zero and AnimateDiff, with Tune-a-video using two pre-trained weights).

Celeb-DF-v2 [Li *et al.*, 2020] is a large-scale dataset extensively used for deepfake detection, focusing on face-swapped videos of celebrities created using common deepfake techniques. It presents a data distribution with 590 real videos and 5,639 deepfake videos, all averaging 13 seconds at 30 FPS, primarily sourced from YouTube interviews.

FaceForensics++ (C23) [Rossler *et al.*, 2019] is a dataset of manipulated videos using four different techniques (Deepfakes, Face2Face, FaceSwap and NeuralTextures). It applied these four state-of-the-art automated face manipulation methods to 1,000 pristine videos originally sourced from YouTube.

UADFV [Li *et al.*, 2018] is an earlier deepfake detection dataset focusing on physical signals such as eye blinking, facial expressions, and lip movements. It generates 49 deepfake videos from web-sourced interviews and presentations.

5.3 Parameter Settings

According to the input requirements of the VLM we use, we utilize the DashScope SDK and OpenAI API to pass the local video path or segmented frames. As part of this process, we set the sampling parameter to **fps=10**, which means the VLM samples at 10 frames per second. If the total frames exceeded the maximum allowing context of the models, we took the maximum number of frames with equal-interval sampling. In actual processing, if a video causes errors because its frame rate does not meet this requirement, we skip the video. Temperature and other hyper-parameters were set as default to facilitate reproduction. We evaluated three times and reported the average accuracy to avoid fluctuation. Input videos were processed at their original resolution, with VLMs automatically resizing according to their internal requirements (e.g., maximum 10MB for some models). The baseline prompt simply asked: *"Please determine whether this video is AI-generated or real. Directly provide your answer as either 'AI-generated' or 'Real'."*

6 Results

We present the empirical outcomes of our human-insight-driven prompting methodology. We first discuss the overall accuracy improvements achieved by our approach in comparison to baseline methods. This is followed by an ablation study designed to elucidate the distinct and combined contributions of the human insight categories.

6.1 Accuracy

Our primary finding, as evidenced in Table 2, is that the proposed human-insight-driven prompting strategy ('Ours') generally achieves superior accuracy in identifying AIGC misinformation videos compared to conventional baseline prompting techniques ('baseline'). This advantage was observed across a diverse range of state-of-the-art VLMs and challenging benchmark datasets utilized in our evaluation.

Several prominent VLMs demonstrated consistent and often substantial performance enhancements when guided by our human-centric prompts. For example, the Claude-3-7-sonnet model, when employing our method, outperformed its baseline counterpart on all five datasets, with accuracy increasing from 0.9203 to 0.9849 on GenVidBench and from 0.7725 to 0.9087 on FaceForensics++. Similarly, the gpt-4o-mini model exhibited marked improvements, notably achieving perfect accuracy (1.0000) on the UADFV dataset with our method, compared to the baseline's 0.9796, and a significant gain on T2VQA-DB (0.9700 'Ours' vs. 0,8900 'baseline'). The gemini model also consistently benefited across all datasets, registering a particularly wide margin on UADFV (0.6021 'Ours' vs. 0.2857 'baseline'). Overall, the results confirm the general efficacy of our approach and highlight the significant potential of systematically integrating human-derived analytical cues to enhance VLM accuracy in misinformation video detection.

Table 2: Accuracy of different models across different datasets.

Model	Strategy	GenVidBench	T2VQA-DB	Celeb-DF-v2	FaceForensics++	UADFV
Qwen-vl-max	Ours	0.9850	0.8280	0.9660	0.9036	0.6415
	baseline	0.9450	0.6774	0.9260	0.8798	0.7062
GLM-4v-plus	Ours	0.9350	0.8265	0.5583	0.9348	0.8834
	baseline	0.9509	0.6700	0.4500	0.8199	0.7000
Claude-3-7-sonnet	Ours	0.9849	0.9950	0.5540	0.9087	0.8544
	baseline	0.9203	0.9800	0.5283	0.7725	0.8029
Gemini	Ours	0.9565	0.8200	0.6272	0.7025	0.6021
	baseline	0.9189	0.6947	0.4799	0.6760	0.2857
GPT-4o	Ours	0.9600	0.9000	0.3600	0.2027	0.5249
	baseline	0.9700	0.9500	0.3250	0.2460	0.5200
GPT-4o-mini	Ours	0.9900	0.9700	0.9869	0.9909	1.0000
	baseline	0.9800	0.8900	0.9637	0.9913	0.9796

6.2 Ablation

To dissect the impact of the constituent components within our human-insight-driven prompting strategy, we conducted an ablation study. The results, presented in Table 3, evaluate configurations where one or more of the three core human insight categories – A (Content & Narrative Coherence), B (Visual Details & Realism), and C (Image Integration & Composition) – were systematically removed or isolated.

The results primarily revealed that the 'All' configuration, integrating all three human insight categories, consistently and often substantially outperformed the 'Vanilla' (naive prompting) baseline across the evaluated datasets. This outcome highlights the collective benefit derived from employing structured human insights. For instance, on the FaceForensics++ dataset, the 'All' configuration achieved perfect accuracy (1.0000), a marked improvement over the 'Vanilla' baseline's 0.7314. Similar significant performance gains for the 'All' configuration were also obsrved on other datasets, including Celeb-DF-v2 (0.6600 vs. 0.5399) and GenVidBench (0.8160 vs. 0.7741).

Furthermore, the results elucidated the nuanced, dataset-contingent roles of individual insight categories. A clear synergistic effect among the three categories was evident on the FaceForensics++ dataset. The removal of any single insight category (configurations 'All-A', 'All-B', 'All-C') led to a significant degradation in accuracy from the 1.0000 achieved by the 'All' configuration (e.g., 'All-A' accuracy dropped to 0.7903). This demonstrates that for this specific dataset, all three components were critical and contributes positively to achieving optimal performance. Despite these dataset-specific complexities in component interactions, the consistent margin by which the 'All' configuration surpassed the 'Vanilla' baseline reinforces the overall value of the comprehensive, multi-faceted human-insight prompting strategy.

6.3 Case Analyses

To illustrate our prompting method's analytical accuracy and enhanced explainability, we analyze two distinct cases using the VLM's articulated reasoning. Case 1 (Figure 2, corridor scene) was assessed by the VLM correctly as AI-generated, a conclusion reached through multiple suspicious indicators identified via our structured prompts (reasoning detailed in Appendix C). For **Content & Narrative Coherence,** prompted by human insights on AIGC's narrative simplicity, the VLM identified a "rudimentary plot devoid of complex developments." This assessment, aligning with AIGC characteristics, demonstrated the prompt's utility in pinpointing a key human-perceived marker of artificiality. Under **Visual Details & Realism**, prompts concerning artificial cleanliness, texture uniformity, perspective, and flexion (common AIGC flaws) led to several key findings. The VLM noted the corridor's "high degree of uniformity and lacks discernible imperfections," the attire's "uniform coloration, lacking naturalistic gradients," "notable evenness' in lighting, "excessive idealization" and "unnaturally precise" geometry, and necessitated scrutiny for "mechanistic or robotic movement patterns." These observations, highlighting a lack of real-world irregularities and natural variations, collectively supported suspicious as critical human-perceived indicators of AI. For **Image Integration & Composition,** prompts about modularity guided the consideration of signs like abrupt changes or repetition. The VLM's synthesis, stating "multiple factors ... converge to suggest an AI-generated origin," shows how this structured interrogation built a compelling case from accumulated human-perceptible evidence.

(a) The sixth extracted frame. (b) The thirtieth extracted frame.

Figure 2: (a) the sixth and (b) thirtieth extracted frame.

Case 2 (Figure 3, interview scene) led to a contrasting

Table 3: Accuracy of different ablation configurations across different datasets. Baseline denoted prompting with no designed strategy.

Model	GenVidBench	T2VQA-DB	Celeb-DF-v2	FaceForensics++	UADFV
All	0.8160	0.9500	0.6600	1.0000	0.5431
All-A	0.6979	0.9600	0.7025	0.7903	0.5769
All-B	0.8348	0.9250	0.6316	0.8083	0.6194
All-C	0.8732	0.9600	0.5634	0.7135	0.5778
All-A/B	0.7864	0.9150	0.5731	0.7787	0.6075
All-A/C	0.8595	0.9200	0.5615	0.7183	0.5046
All-B/C	0.6474	0.9146	0.5187	0.7036	0.5572
Baseline	0.7741	0.9141	0.5399	0.7314	0.4741

conclusion: the video could not be definitively determined to be AI-generated, thus suggesting authenticity (reasoning from Appendix C). For **Content & Narrative Coherence,** the VLM found the storyline "conventional ... and does not appear unusually bland." This response, unlike Case 1, indicated authenticity by not aligning with AIGC's typical narrative shallowness. In **Visual Details & Realism**, prompts guided the VLM to confirm "normal textual appearance, devoid of anomalous smoothness," "positioning and scale ... consistent with realistic proportions," and "gestures and bodily movements appear natural and fluid." The absence of AI-associated flaws in texture, perspective, and motion consistently supported the video's authenticity. Similarly, for **Image Integration & Composition,** prompts led to findings of a "natural, without conspicuous segmentation" foreground-background transition and a "cohesive whole, with no discernible traces of sectional assembly." The lack of artificial compositing or modularity further indicated authenticity. The VLM's summary for Case 2, noting "... no two or more ... factors present conditions that warrant significant suspicious," sharply contrasts with Case 1, underscoring the framework's ability to guide nuanced, evidence-based assessments. These analyses highlights the effectiveness in guiding VLMs to distinct, explainable identification.

(a) The fifth frame.

(b) The twenty-ninth frame

Figure 3: The fifth and twenty-ninth frame.

7 Discussions and Future Directions

Building on these findings, key future directions emerge. First, further investigation into nuanced insight contributions is crucial, including exploring dynamic integration, an expanded perceptual heuristic taxonomy potentially leveraging cognitive science, and sophisticated prompt engineering beyond current hierarchies. Second, enhancing VLM explainability and intrinsic capabilities is vital; while our approach grounds detection in human-understandable terms, VLMs should provide detailed, localized (e.g., spatio-temporal) justifications. Third, broadening analytical scope and enhancing robustness remain paramount. While our English-based prompts show strong performance, the universality of visual artifacts suggests potential cross-linguistic applicability, as the identified human insights represent cognitive heuristics that transcend linguistic boundaries [Wang *et al.*, 2025]. For extended video content, our modular framework allows hierarchical analysis where videos are segmented temporally and insights aggregated across segments [Azad *et al.*, 2025]. The analytical scope could broaden from visual cues to incorporate multimodal elements (audio, metadata, platform-specific context [Micallef *et al.*, 2022]) for holistic assessment and improved resilience against sophisticated multimodal deepfakes. Finally, addressing the rapid evolution of AIGC techniques [Papadopoulos *et al.*, 2023; Xu *et al.*, 2023] is paramount. Detection models require inherent explainability, robustness, and adaptability to novel generation methods. Our framework's modularity enables incremental updates as new artifact types emerge [Faber *et al.*, 2024]. However, there still remains need for research into few-shot prompting adaptation for new artifact types and domain generalization [Ni *et al.*, 2025]. While our user study with 14 participants achieved thematic saturation for the core insight categories, future work could validate these findings with diverse participant pools to enhance generalizability across different cultural contexts and user demographics.

Acknowledgements

This work was supported by the Natural Science Foundation of China under Grant No. 62472243, 62132010 and 2022YFB3105201. This work was also supported by Deng Feng Fund.

Ethical Statement

We acknowledged that our paper has ethical concerns and tried our best to address these concerns. We followed Menlo report [Bailey *et al.*, 2012] and Belmont report [Beauchamp and others, 2008] in organizing, designing and carrying out studies. Participants in our study has freedom in quitting the study without any reason, and was properly compensated according to the local wage standard. At the designing stage, our study encouraged participants to discern misinformation videos and critically think about the misinformation. Participants also agreed unanimously they reflected more critically after using the systems. After the study, we debriefed the study's content, the fake videos to the participants and encouraged them to critically examine the misinformation videos.

A Prompt Template for AI Video Detection

The below is the complete prompt template used to guide the evaluation of whether a given video is AI-generated. In this template, the evaluator is instructed to adopt a critical stance—presuming a high likelihood of AI generation—and to consider any two or more of the listed factors as sufficient grounds to deem a video AI-generated.

> Please carefully assess whether the video is AI-generated, adopting a critical stance regarding its authenticity and presuming a high probability of AI generation. If two or more of the following factors raise suspicion, the video should be considered AI-generated. The factors to consider are as follows:
>
> 1. Bland or flat storylines increase the probability of AI generation.
> 2. Too clean transitions between background and foreground raise doubts.
> 3. Excessively uniform material or texture without realistic gradients appears unnatural.
> 4. Unscientific or illogical perspective relationships suggest AI origin.
> 5. Computer-generated hard folds or creases raise suspicion.
> 6. Modular or collage-like visual structures suggest AI compositing.
>
> In addition to the above factors, you should also incorporate the video's content for a more personalized evaluation; if any blatant inconsistencies appear, you may determine that it is AI-generated. Finally, please provide a definitive judgment: either output "This video is AI-generated" or "This video is not AI-generated." No rationale is needed—only the result.

B Baseline Prompt Template

For comparison purposes, we also provide the baseline prompt template used in our experiments. This baseline prompt asks for video authenticity assessment without the structured human-insight-driven guidance that characterizes our main approach.

> Please carefully examine the provided video and determine whether it is AI-generated or authentic. Consider all aspects of the video including visual quality, content consistency, and any potential artifacts that might indicate artificial generation.
>
> Based on your analysis, please provide a definitive judgment: either output ”This video is AI-generated” or ”This video is not AI-generated.” No detailed rationale is needed—only the final result.

C The reasoning processes for the Case Analyses

The below is the reasoning process of the VLM for the first case.

> A systematic analysis of the provided video is undertaken to determine its potential AI-generated origin. The evaluative framework stipulates that the presence of two or more suspicious indicators, as per user-defined considerations, warrants a classification of the video as AI-generated.
>
> Firstly, an assessment of narrative simplicity was conducted. The video portrays an individual ambulating within a corridor, a scenario characterized by a rudimentary plot devoid of complex developments. This observation aligns with a common characteristic of AI-generated content, as AI systems may exhibit limitations in constructing intricate narratives.
>
> Subsequently, the interface between background and foreground elements was examined for excessive demarcation or cleanliness. The visual representation of the corridor, encompassing walls, doors, and flooring, exhibits a high degree of uniformity and lacks discernible imperfections. Such an overly pristine presentation may suggest artificial generation, given that AI synthesis might not fully replicate the subtle irregularities inherent in real-world environments.
>
> Attention was then directed towards the homogeneity of material textures. The attire of the individual (a sweater and trousers) displays uniform coloration, lacking naturalistic gradients or signs of wear. Furthermore, the illumination and shadowing within the corridor are rendered with notable evenness, absent of the nuanced interplay of light typically observed in authentic settings. These aspects could also be indicative of AI synthesis.
>
> Regarding perspectival relationships, the depth of the corridor and the positioning of doors appear plausible. However, the overall spatial representation may exhibit an excessive idealization, lacking the inherent irregularities of physical environments. For instance, the linearity of wall corners and door frames appears unnaturally precise, without the subtle curvatures or wear characteristic of real-world structures.
>
> The presence of hard-edged characteristics in flexed regions was also considered. Although no specific objects exhibiting overt, unnatural flexure were prominent, the fluidity of the depicted individual's motion was scrutinized. Specifically, the naturalness of limb articulation during ambulation was evaluated for any mechanistic or robotic movement patterns, which are frequently observed artifacts in AI-generated videos.
>
> Finally, an examination for modular characteristics or collage-like effects was performed. Abrupt positional

changes of the individual across frames or the repetition of background elements could indicate the use of pre-existing modular assets. For instance, repetitive patterns or structures in the corridor's doors and walls might lead to a disjointed overall visual composition.

Synthesizing these observations, multiple factors within the video converge to suggest an AI-generated origin: the simplistic narrative, the overly clean transitions between background and foreground, the excessive uniformity of material textures, and the idealized perspective. Therefore, it is concluded that the video is AI-generated.
Conclusion: The video is assessed to be AI-generated.

The below is the reasoning process of the VLM for the second case.

A meticulous analysis is now required to determine if the video in question is AI-generated. The assessment will be based on six user-provided factors, each of which will be examined systematically.

First, the mundanity of the storyline is considered. The video depicts an interview setting: a guest, attired in a suit and tie, is seated and speaking, with a backdrop of a nocturnal cityscape and mountains. Such a scene is conventional for authentic interview programs and does not appear unusually bland. Therefore, this aspect is unlikely to support a conclusion of AI generation.

Second, the cohesion between the background and foreground is addressed, specifically whether the transition is unnaturally seamless or exhibits overly clean separation. Observation of the video frames indicates that the transition between the background elements (city lights and mountains) and the foreground subjects (host and guest) appears natural, without conspicuous segmentation or incongruity. Although AI-generated imagery can sometimes exhibit a detachment between background and subject, the elements in this video appear well-integrated. Thus, this factor does not presently arouse suspicion.

Third, the potential for excessive uniformity in the materials or textures of objects is evaluated. The guest's suit, tie, and watch display a normal textural appearance, devoid of anomalous smoothness or unnatural reflections. Similarly, the background lights and mountains do not present evident textural irregularities. Consequently, this aspect is also not considered suspicious.

Fourth, the scientific plausibility and rationality of the perspective are examined. Visual inspection of the frames confirms that the positioning and scale of the individuals are consistent with realistic proportions, showing no distortions or illogical angles. The depiction of the background mountains and city lights also adheres to principles of depth and perspective. Therefore, no issues are identified concerning the perspective.

Fifth, the presence of hard, artificial-looking lines or unnatural creases at points of flexion (e.g., in clothing or body posture) is assessed. The guest's gestures and bodily movements appear natural and fluid, lacking the rigid, computer-generated appearance of sharp folds or unnatural articulations. For instance, his hand gestures transition smoothly across different frames, without abrupt breaks or disjointed movements.

Sixth, the existence of modular characteristics or a collage-like effect is investigated. The entire scene presents as a cohesive whole, with no discernible traces of sectional assembly or patching. The guest's expressions and actions are coherent, and the background and foreground elements are harmoniously integrated, obviating any sense of a 'blocked' or piecemeal construction.

In summary, based on the preceding analysis, no two or more of these six factors present conditions that warrant significant suspicion. Therefore, according to the user's specified criteria, it cannot be definitively determined that the video is AI-generated.

D Failure Case Analysis

To illustrate our method's limitations, we present a failure case where our framework incorrectly classified an authentic presentation video as AI-generated.

D.1 Case Description and VLM Analysis

The video depicts a speaker delivering a professional presentation. Despite being authentic, our VLM framework misclassified it as AI-generated based on three suspicious factors out of six evaluated dimensions:

Facial and Body Movement Fluidity: The speaker's facial expressions appear somewhat rigid, particularly during gestural movements, lacking smooth, spontaneous quality and raising suspicions about authenticity.

Eye and Mouth Synchronization: Analysis reveals slight temporal delay between lip movements and audio, with blinking frequency notably below normal ranges, creating an uncanny visual effect.

Background and Foreground Consistency: The transition between background and foreground elements appears smooth without obvious edge artifacts or blurring zones that would indicate digital compositing.

Lighting and Shadow Effects: The illumination appears realistic, with shadow directions and intensities consistent with apparent light sources in the scene.

Detail and Texture Quality: Skin and hair textures appear somewhat coarse, with pixelation phenomena becoming apparent under close inspection, suggesting potential compression artifacts.

Content Coherence: The speech content maintains logical consistency throughout, with no apparent narrative inconsistencies or absurd elements.

Based on three suspicious factors (movement fluidity, synchronization, and texture quality) meeting the "two or more" threshold, the video is determined to be AI-generated.

D.2 Error Analysis

This misclassification reveals key limitations: **(1) Technical Artifact Misinterpretation:** The VLM incorrectly attributed compression artifacts and natural speaking variations to AI generation. **(2) Context Insensitivity:** The framework failed to account for professional presentation settings where controlled, formal demeanor is expected.

This case underscores the need for more sophisticated contextual understanding and refined evaluation criteria in future iterations.

E Frame Sampling Rate Analysis

To determine the optimal frame sampling rate, we systematically evaluated different fps values (2, 5, 10, 20) across representative video subsets from each dataset using three VLM architectures.

E.1 Key Findings

Performance Consistency: Accuracy variations between different fps settings remained below 2% across all tested combinations. This consistency suggests that human-derived insights are robust to moderate temporal resolution variations.
Frame Adequacy Issues: A critical challenge emerged with videos containing insufficient frames relative to sampling requirements. When videos had fewer frames than required by the fps setting, all sampling rates encountered similar processing difficulties.

E.2 Selection Rationale

fps=10 was selected for: (1) consistent accuracy across video types and VLM architectures, (2) sufficient temporal resolution for artifact detection, (3) reasonable computational efficiency, and (4) better handling of edge cases involving variable frame rates. The negligible performance differences validate that our human-insight-driven approach effectively guides VLMs to focus on semantically meaningful artifacts rather than high-frequency temporal patterns.

References

[Abdali *et al.*, 2024] Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Computing Surveys*, 57(3):1–29, 2024.

[Aldiabat and Le Navenec, 2018] Khaldoun M Aldiabat and Carole-Lynne Le Navenec. Data saturation: The mysterious step in grounded theory methodology. *The qualitative report*, 23(1):245–261, 2018.

[Allison *et al.*, 1994] Truett Allison, Heidi Ginter, Gregory McCarthy, Anna C Nobre, AINA Puce, MARIE Luby, and DENNIS D Spencer. Face recognition in human extrastriate cortex. *Journal of neurophysiology*, 71(2):821–825, 1994.

[Azad *et al.*, 2025] Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. Hierarq: Task-aware hierarchical q-former for enhanced video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8545–8556, 2025.

[Bailey *et al.*, 2012] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The menlo report. *IEEE Security & Privacy*, 10(2):71–75, 2012.

[Bakhshi *et al.*, 2014] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 965–974, 2014.

[Banerjee *et al.*, 2022] Snehasish Banerjee, Monica Lens, and Anjan Pal. Put on your sunglasses and smile: The secret of airbnb hosts' profile photos? *International Journal of Hospitality Management*, 103:103219, 2022.

[Beauchamp and others, 2008] Tom L Beauchamp et al. The belmont report. *The Oxford textbook of clinical research ethics*, pages 149–155, 2008.

[Clarke and Braun, 2017] Victoria Clarke and Virginia Braun. Thematic analysis. *The journal of positive psychology*, 12(3):297–298, 2017.

[Dai *et al.*, 2018] Yue Nancy Dai, Gregory Viken, Eunsin Joo, and Gary Bente. Risk assessment in e-commerce: How sellers' photos, reputation scores, and the stake of a transaction influence buyers' purchase behavior and information processing. *Computers in Human Behavior*, 84:342–351, 2018.

[de Leeuw den Bouter *et al.*, 2023] Merel de Leeuw den Bouter, Javier Lloret Pardo, Zeno Geradts, and Marcel Worring. Protoexplorer: Interpretable forensic analysis of deepfake videos using prototype exploration and refinement, 2023.

[Faber *et al.*, 2024] Kamil Faber, Dominik Zurek, Marcin Pietron, Nathalie Japkowicz, Antonio Vergari, and Roberto Corizzo. From mnist to imagenet and back: benchmarking continual curriculum learning. *Machine Learning*, 113(10):8137–8164, 2024.

[Joslin *et al.*, 2024] Matthew Joslin, Xian Wang, and Shuang Hao. Double face: Leveraging user intelligence to characterize and recognize {AI-synthesized} faces. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1009–1026, 2024.

[Kong *et al.*, 2024] Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C Kot. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *arXiv preprint arXiv:2404.08452*, 2024.

[Kou *et al.*, 2024] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7793–7802, 2024.

[Kundu *et al.*, 2025] Rohit Kundu, Athula Balachandran, and Amit K Roy-Chowdhury. Truthlens: Explainable deepfake detection for face manipulated and fully synthetic data. *arXiv preprint arXiv:2503.15867*, 2025.

[Li *et al.*, 2018] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. Ieee, 2018.

[Li *et al.*, 2020] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[Li *et al.*, 2025] Yongze Li, Ning Li, Ajian Liu, Hui Ma, Liying Yang, Xihong Chen, Zhiyao Liang, Yanyan Liang, Jun Wan, and Zhen Lei. Fa^{3}-clip: Frequency-aware cues fusion and attack-agnostic prompt learning for unified face attack detection. *arXiv preprint arXiv:2504.00454*, 2025.

[Mansoor and Iliev, 2025] Nazneen Mansoor and Alexander I Iliev. Explainable ai for deepfake detection. *Applied Sciences*, 15(2):725, 2025.

[Micallef *et al.*, 2022] Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. Cross-platform multimodal misinformation: taxonomy, characteristics and detection for textual posts and videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 651–662, 2022.

[Morrisey *et al.*, 2019] Marcus N Morrisey, Ruth Hofrichter, and MD Rutherford. Human faces capture attention and attract first saccades without longer fixation. *Visual Cognition*, 27(2):158–170, 2019.

[Ni *et al.*, 2025] Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A challenging benchmark for detecting ai-generated video. *arXiv preprint arXiv:2501.11340*, 2025.

[Papadopoulos *et al.*, 2023] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, pages 36–44, 2023.

[Pellicer *et al.*, 2024] Alvaro Lopez Pellicer, Yi Li, and Plamen Angelov. Pudd: towards robust multi-modal prototype-based deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3809–3817, 2024.

[Rashtian *et al.*, 2014] Hootan Rashtian, Yazan Boshmaf, Pooya Jaferian, and Konstantin Beznosov. To befriend or not? a model of friend request acceptance on facebook. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 285–300, 2014.

[Rossler *et al.*, 2019] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[Ruffin *et al.*, 2024] Margie Ruffin, Haeseung Seo, Aiping Xiong, and Gang Wang. Does it matter who said it? exploring the impact of deepfake-enabled profiles on user perception towards disinformation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1328–1341, 2024.

[Seng *et al.*, 2020] Sovantharith Seng, Akash Chintha, and Bao Thai. Defaking deepfakes: Understanding journalists' needs for deepfake detection. 2020.

[Shahid *et al.*, 2022] Farhana Shahid, Srujana Kamath, Annie Sidotam, Vivian Jiang, Alexa Batino, and Aditya Vashistha. " it matches my worldview": Examining perceptions and attitudes around fake videos. In *proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–15, 2022.

[Soltandoost *et al.*, 2025] Elahe Soltandoost, Richard Plesh, Stephanie Schuckers, Peter Peer, and Vitomir Štruc. Extracting local information from global representations for interpretable deepfake detection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1629–1639, 2025.

[Tahmasebi *et al.*, 2024] Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2189–2199, 2024.

[Tsigos *et al.*, 2024] Konstantinos Tsigos, Evlampios Apostolidis, Spyridon Baxevanakis, Symeon Papadopoulos, and Vasileios Mezaris. Towards quantitative evaluation of explainable ai methods for deepfake detection. In *Proceedings of the 3rd ACM international workshop on multimedia AI against disinformation*, pages 37–45, 2024.

[Walther *et al.*, 2001] Joseph B Walther, Celeste L Slovacek, and Lisa C Tidwell. Is a picture worth a thousand words? photographic images in long-term and short-term computer-mediated communication. *Communication research*, 28(1):105–134, 2001.

[Wang *et al.*, 2024] Jingwei Wang, Ziyue Zhu, Chunxiao Liu, Rong Li, and Xin Wu. Llm-enhanced multimodal detection of fake news. *PloS one*, 19(10):e0312240, 2024.

[Wang *et al.*, 2025] Hao Wang, Pinzhi Huang, Jihan Yang, Saining Xie, and Daisuke Kawahara. Traveling across languages: Benchmarking cross-lingual consistency in multimodal llms. *arXiv preprint arXiv:2505.15075*, 2025.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Xu *et al.*, 2023] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. Combating misinformation in the era of generative ai models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9291–9298, 2023.

[Xu *et al.*, 2025] Qingzheng Xu, Heming Du, Szymon ? ukasik, Tianqing Zhu, Sen Wang, and Xin Yu. Mdam3: A misinformation detection and analysis framework for multitype multimodal media. In *Proceedings of the ACM on Web Conference 2025*, pages 5285–5296, 2025.

[Yu *et al.*, 2025] Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of vision-language models for generalizable and explainable deepfake detection, 2025.

[Zhang *et al.*, 2025] Guosheng Zhang, Keyao Wang, Haixiao Yue, Ajian Liu, Gang Zhang, Kun Yao, Errui Ding, and Jingdong Wang. Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models. *arXiv preprint arXiv:2501.01720*, 2025.