Multimedia Verification Through Multi-Agent Deep Research Multimodal Large Language Models

Huy Hoan Le*
hoanlh4@fpt.com
Quy Nhon AI, FPT Software
Quy Nhon, Vietnam

Van Sy Thinh Nguyen*
thinhnvs@fpt.com
Quy Nhon AI, FPT Software
Quy Nhon, Vietnam

Thi Le Chi Dang
chidtl@fpt.com
Quy Nhon AI, FPT Software
Quy Nhon, Vietnam

Vo Thanh Khang Nguyen
khangnvt1@fpt.com
Quy Nhon AI, FPT Software
Quy Nhon, Vietnam

Truong Thanh Hung Nguyen†
hung.ntt@unb.ca
University of New Brunswick
Fredericton, New Brunswick, Canada

Hung Cao
hcao3@unb.ca
University of New Brunswick
Fredericton, New Brunswick, Canada

Abstract

This paper presents our submission to the *ACMMM25 - Grand Challenge on Multimedia Verification*. We developed a multi-agent verification system that combines Multimodal Large Language Models (MLLMs) with specialized verification tools to detect multimedia misinformation. Our system operates through six stages: raw data processing, planning, information extraction, deep research, evidence collection, and report generation. The core Deep Researcher Agent employs four tools: reverse image search, metadata analysis, fact-checking databases, and verified news processing that extracts spatial, temporal, attribution, and motivational context. We demonstrate our approach on a challenge dataset sample involving complex multimedia content. Our system successfully verified content authenticity, extracted precise geolocation and timing information, and traced source attribution across multiple platforms, effectively addressing real-world multimedia verification scenarios.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → **Privacy policies**; **Digital rights management**.

Keywords

Multimedia Verification, Multimodal Large Language Models

1 Introduction

Multimedia misinformation has become a critical challenge in the digital information landscape, with visual content serving as a primary vector for spreading false narratives. Recent studies show that around 80% of fact-checked misinformation cases involve images or videos, making multimedia verification essential for maintaining information integrity [11]. Modern misinformation employs two main strategies: deepfakes involving sophisticated content manipulation, and cheapfakes where genuine media is repurposed in misleading contexts. Current multimedia verification approaches typically focus on either technical forensics or contextual analysis in isolation. Traditional media forensics tools excel at detecting pixel-level manipulations but struggle with context verification, while text-based fact-checking systems cannot adequately process visual information. Although Multimodal Large Language Models (MLLMs) offer new possibilities for multimedia analysis, they face challenges with hallucination and lack grounding in verifiable evidence sources.

The *ACMMM25 - Grand Challenge on Multimedia Verification* [8] provides an opportunity to address these limitations through comprehensive verification systems that can handle real-world misinformation scenarios. The challenge requires systems to process complex multimedia content and provide detailed verification reports with evidence-based assessments of authenticity and contextual accuracy. In this paper, we present a multi-agent multimedia verification system designed specifically for this challenge. Our approach integrates MLLMs with specialized verification tools through a systematic six-stage pipeline that processes multimedia content from initial data extraction through comprehensive report generation. Our contributions are as follows:

- **Multi-agent verification architecture:** We propose a six-stage pipeline that systematically processes multimedia content from data extraction through comprehensive report generation, ensuring thorough coverage of both technical and contextual verification requirements.
- **Deep Researcher Agent with specialized tools:** We introduce an agent that employs four verification tools, including reverse image search, metadata analysis, fact-checking databases, and a novel verified news processor that extracts spatial, temporal, attribution, and motivational context.
- We demonstrate the effectiveness of our proposed system through evaluation on the *ACMMM25 - Grand Challenge on Multimedia Verification* dataset [8], successfully handling complex multimedia content involving geopolitical events.

2 Related Work

2.1 Multimedia Verification

The field of multimedia verification has experienced substantial growth in recent years, driven by the increasing prevalence of misinformation and the sophisticated nature of modern content manipulation techniques. As visual content becomes a major vector for misinformation, research on verifying the authenticity and context of images/videos has grown significantly. Recent studies found that around 80% of fact-checked misinformation cases online include an image or video [11].

*Both authors contributed equally to this research.
†Corresponding author.

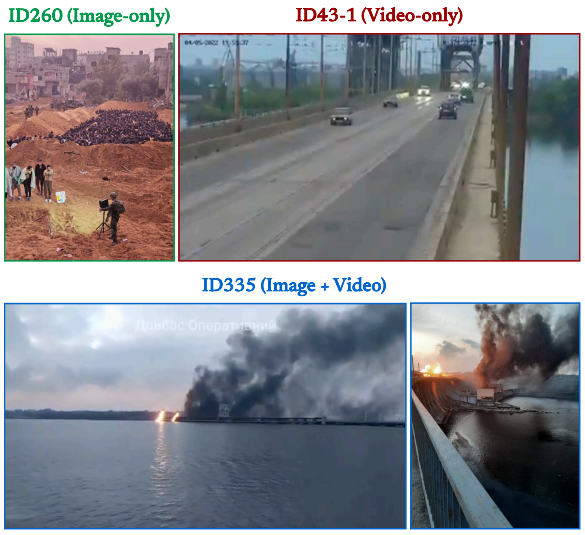

Figure 1: *ACMMM25 - Grand Challenge on Multimedia Verification* dataset for verifying the authenticity and context of multimedia content.

Prior studies highlight two major challenges: (1) "deepfakes" (i.e., content tampering and synthesis, where images/videos are digitally altered or generated) and (2) "cheapfakes" (i.e., content miscontextualization, where genuine media is reused in a false context to mislead). Early efforts in media forensics tackled image manipulations like splicing, copy-move, or GAN-generated fakes. For example, the MediaEval 2016 Verifying Multimedia Use task aimed to automatically detect manipulated and misleading use of web images and videos [4]. They defined a post as fake if the visual content does not actually depict the event described by the accompanying text, e.g., an old or unrelated photo falsely presented as current news. Large benchmark datasets such as the NIST Media Forensics Challenge [16] and FaceForensics++ [31] have driven progress in detecting tampered images and AI-generated videos. In parallel, the recent "Detecting Cheapfakes" Grand Challenges at ACM Multimedia and related venues specifically target out-of-context (OOC) image misuse, reflecting a growing research emphasis on context verification alongside traditional media forensics [3].

2.2 Multimodal LLMs (MLLMs) for Multimedia Verification

Given the inherently multimodal nature of online misinformation, recent research has trended toward joint image–text verification. Prior studies have shown that leveraging both the visual and textual modalities can improve accuracy over text-only or image-only methods [5, 13, 17, 20, 22, 24, 28, 29, 33]. Previous attempts at multimedia verification have focused on developing comprehensive Open Source Intelligence (OSINT) methodologies combining reverse image search platforms [2, 12, 15, 23], metadata analysis tools [12], structured fact-checking frameworks [24], and knowledge graphs [9, 11, 27].

The advent of MLLMs has opened new avenues for multimedia verification. These models bring vast parametric world knowledge from their training data, which can help recognize when a caption makes implausible claims about an image. While MLLMs excel at flexible reasoning, they have drawbacks. A known issue is hallucination, where a MLLM may fabricate plausible-sounding details if it lacks actual evidence [14, 19, 34]. This is problematic in fact-checking, as the system might incorrectly justify a decision with false information.

To counter this, recent research has explored hybrid approaches that combine LLM reasoning with external knowledge retrieval or structured representations, such as treating the verification task as a multi-step reasoning process orchestrated by AI agents. Lakara et al. [21] involved an agent or multiple agents that can dynamically query tools and data sources, e.g., performing image analysis, web searches, and knowledge base lookups, guided by an LLM's logic. Duwal et al. [11] integrated information from knowledge graphs or used graph neural networks to ensure decisions are grounded in real data. Such methods aim to retain the interpretability of LLM-based reasoning while improving accuracy and trustworthiness by anchoring the model's output to verifiable evidence. Also, Liu et al. [22] developed FKA-Owl, a framework that leverages forgery-specific knowledge to augment MLLMs, enabling them to reason about manipulations effectively by incorporating semantic correlations between text and images and artifact traces in image manipulation. Similarly, Gao et al. [13] proposed a knowledge-enhanced vision and language model that integrates information from large-scale open knowledge graphs to augment the ability to discern the veracity of news content. Such multimodal LLM agents can dynamically choose operations (e.g., reverse image search, object recognition) and reason about the results in a conversational manner. Braun et al. [5] introduced DEFAME, a zero-shot multimodal fact-checking pipeline that uses an LLM backbone to retrieve and analyze text and image evidence, producing structured verification reports. Nguyen et al. [24] introduced a two-stage cheapfake-detection pipeline that first performs fast reputation checking by reverse-image-searching a photo and verifying whether it appears on trusted news domains, then passes the image–caption triplet to a visual-language network trained primarily on LLM-generated synthetic data to judge contextual integrity.

Our work follows this trend, focusing on the main task of multimedia verification by leveraging an MLLM agent that can analyze content and use online information. Our agent-based approach builds upon prior research by marrying the strengths of MLLM reasoning and tool-assisted verification, aiming for high accuracy and reliability in tackling real-world misinformation challenges.

3 Dataset

In this paper, we tackle the main task of the *ACMMM25 - Grand Challenge on Multimedia Verification* [8]. The dataset comprises a collection of 50 high-quality samples designed to reflect real-world multimedia verification challenges. Each sample contains multimedia content files including videos in `.mp4` format and images in `.jpg` format, with audio and video data constituting 72% of the total dataset, while static images account for the remaining. While in the validation set, as shown in Figure 1, most data samples are video-only, case `ID260` is image-only, and case `ID335` contains both video and images.

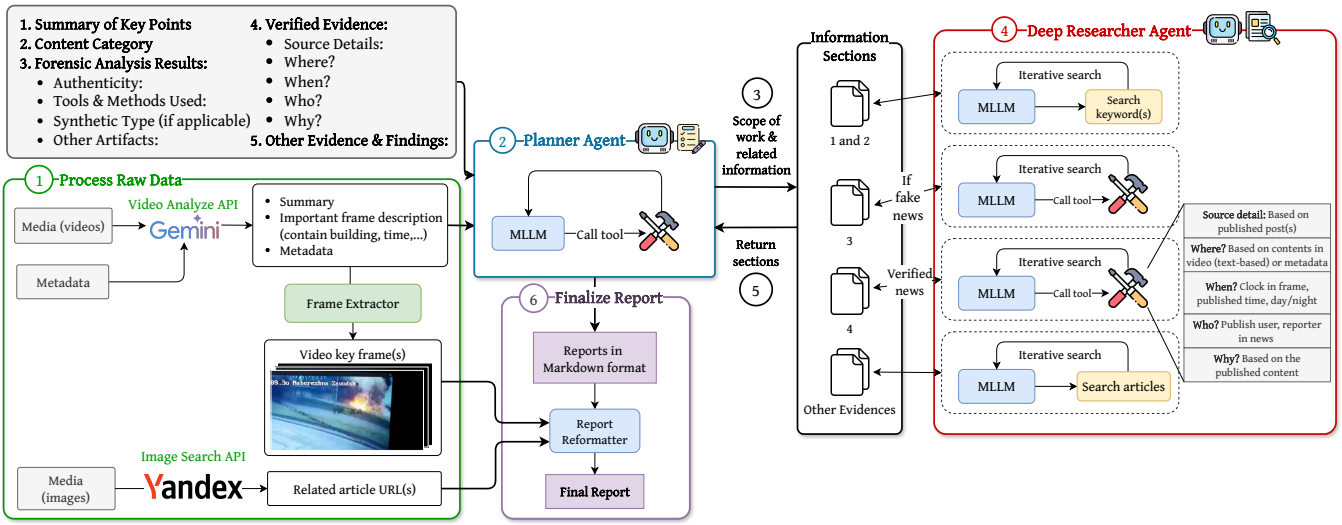

Figure 2: Our proposed Multi-Agent Deep Research MLLMs architecture for the multimedia verification system.

Each verification case is provided as a comprehensive package including: (1) Primary multimedia content consisting of image(s) or video(s) that may originate from various sources and potentially feature content in languages other than English; (2) Contextual information such as captions, descriptions, social media posts, news articles, and available metadata that provide background for the multimedia content; and (3) Additional investigative clues including possible source attributions, claims made about the content, or preliminary fact-checker notes when relevant to the case.

The dataset requires participants to produce comprehensive verification that systematically addresses multiple dimensions of content authenticity and context. The verification framework encompasses three primary analytical components: temporal and spatial verification requiring precise determination of location (geological coordinates) and timing of events (specific dates); forensic authenticity analysis involving detection of synthetic content, modifications, or recapturing using specialized tools and techniques; and evidential cross-verification demanding corroboration through both the provided multimedia data and external internet sources.

Importantly, the challenge's objective extends beyond binary authenticity determination. Instead, it emphasizes comprehensive evidence assessment that identifies which specific elements of the multimedia content can be independently verified through external sources and which aspects remain unverifiable or inconclusive. This approach recognizes the nuanced nature of multimedia verification in real-world scenarios where complete verification may not always be feasible, requiring participants to clearly distinguish between confirmed facts, uncertain elements, and information that cannot be substantiated through available evidence.

4 Methodology

Our multimedia verification system employs a multi-agent architecture that combines MLLMs with external verification tools to systematically analyze and fact-check multimedia content. The system operates through six distinct stages: (1) raw data processing, (2) planning and coordination, (3) information extraction and sectioning, (4) deep research and verification, (5) evidence collection, and (6) comprehensive report generation, as outlined in Figure 2.

4.1 Stage 1 – Raw Data Processing

The initial stage handles diverse multimedia inputs through specialized processing pipelines designed for different media types.

4.1.1 Video Processing Pipeline. For video content, we utilize the Gemini 2.0 Flash as the core MLLM to analyze videos and extract comprehensive metadata and contextual information. The system generates frame-by-frame descriptions that capture temporal dynamics, identify key objects and scenes, and extract technical metadata including timestamps, resolution, and encoding information. A Frame Extractor component automatically identifies and extracts the most informative keyframes that represent critical visual moments in the video sequence.

4.1.2 Image Processing Pipeline. Static images are processed through the Yandex Image Search API [7], which performs image searches to identify potential source materials, related articles, and previous uses of the image across the web. This component generates a comprehensive list of related article URLs and metadata that forms the foundation for subsequent verification steps.

4.2 Stage 2 – Planner Agent

The Planner Agent serves as the central coordination hub, implemented using an LLM with tool-calling capabilities. This agent analyzes the processed multimedia data and associated metadata to develop a systematic verification strategy. The planner determines which verification tools and methods are most appropriate for the specific content type and potential misinformation vectors identified during initial processing. The Planner Agent organizes the verification workflow by creating structured information packets

that guide subsequent analysis stages. It identifies key claims, potential inconsistencies, and areas requiring detailed investigation, then delegates specific verification tasks to specialized components.

4.3 Stage 3 – Information Extraction and Sectioning

Based on the planner's strategic assessment, the system extracts relevant information and organizes it into discrete, analyzable sections. Each section contains specific aspects of the multimedia content that require independent verification, such as:

- **Temporal Claims:** Date and time assertions made by or about the content.
- **Geographical Claims:** Location-based information and spatial context.
- **Entity Recognition:** People, organizations, or objects featured in the content.
- **Contextual Metadata:** Technical and social context surrounding content creation and distribution.

This sectioning approach enables parallel processing of different verification aspects while maintaining systematic coverage of all potentially misleading elements.

4.4 Stage 4 – Deep Researcher Agent

The Deep Researcher Agent represents the core verification engine, implementing an iterative search and analysis framework. This agent employs multiple search strategies:

- **Keyword-Based Search:** The agent generates contextually relevant search terms based on content analysis and performs systematic web searches to gather corroborating or contradicting evidence.
- **Tool-Assisted Verification:** The system integrates multiple external verification tools, including reverse image search engines, metadata analysis utilities, and fact-checking databases. When processing verified news content, this tool systematically extracts four critical source details:
  - Source detail (details of the published posts),
  - Where? (spatial context based on contents in video text, or metadata),
  - When? (temporal context using clocks in frames, published time, day/night indicators,
  - Who? (attribution context identifying published users and reporters in news),
  - Why? (motivational context based on the published content analysis).
- **Source Analysis:** For each piece of evidence discovered, the agent performs detailed source verification, examining publication details, publisher credibility, temporal consistency, and cross-referencing with multiple independent articles.

The Deep Researcher Agent maintains detailed provenance tracking, documenting the complete chain of evidence discovery and analysis decisions to ensure transparency and reproducibility.

4.5 Stage 5 – Evidence Collection and Synthesis

The system aggregates findings from all verification components, organizing evidence according to reliability, relevance, and consistency. This stage performs aggregation of different evidence sources and identifies potential conflicts or contradictions that require additional investigation. Evidence is categorized into verified facts, related information, and disputed claims, with confidence scores assigned based on source reliability and evidence consistency. The system also identifies gaps in available evidence and areas where verification remains inconclusive.

4.6 Stage 6 – Report Generation and Formatting

The final stage synthesizes all verification findings into a comprehensive, structured report. The Report Generator creates detailed documentation following a standardized format:

- **Executive Summary:** Key findings and overall assessment of content authenticity and contextual accuracy.
- **Content Classification:** Categorization of the multimedia content type and potential misinformation vectors identified.
- **Forensic Analysis Results:** Technical analysis of content authenticity, including manipulation detection results, metadata analysis, and synthetic content identification where applicable.
- **Verified Evidence Documentation:** Comprehensive source attribution including temporal, geographical, and contextual details with complete provenance tracking.
- **Additional Findings:** Supplementary evidence, related content analysis, and potential implications for broader misinformation patterns.

The Report Reformatter ensures consistent formatting and presentation, producing both human-readable summaries and machine-readable structured outputs suitable for integration with broader fact-checking workflows.

Altogether, our system leverages MLLMs for reasoning and coordination, while maintaining grounding through external tool integration and evidence-based verification. The modular architecture enables scalable processing of diverse multimedia content while maintaining systematic verification standards. The iterative nature of the Deep Researcher Agent allows for adaptive investigation strategies that can handle novel misinformation patterns and evolving content manipulation techniques.

5 Demonstrative Results

To demonstrate the effectiveness of our multimedia verification system, we present a comprehensive case study using sample `ID43-3` from the *ACMMM25 - Grand Challenge on Multimedia Verification*.

Figure 3 shows the key frames extracted by our system's raw data processing stage from the submitted video content. The Frame Extractor component successfully identified the most informative visual elements, including the moment of impact, the resulting explosion, and the aftermath with visible smoke and structural damage. These extracted frames provided the visual foundation for subsequent verification stages. Our system processed this case

ID43-3: Verificiation Report

 # Case Summary: Missile Strike on Dnipro Bridge Verification Report

- Overview: Multiple video sources and associated metadata depict a missile strike on a bridge in Dnipro, Ukraine. The footage, reportedly recorded from a hotel, shows the explosion at or near Naberezhna Zavodska, with visible aftermath including a significant fireball, rising smoke, and infrastructural damage to a bridge frequently used by vehicles (and possibly trams). Data overlays in the video provide location details and timestamps.
- Verification Status: Verified
- Geolocation Coordinates: Approx. 48.4647° N, 35.0462° E (Dnipro, Ukraine – approximate central coordinates)
- Exact Date: 04/05/2022, with one video timestamp noted at 19:58:37 (local time)

 # Content Classification

- **Tags:** Content Classification:
- Platforms: Twitter (social media video post), Multiple Video Sharing Platforms
- Involved Entities: Hotel occupants filming, local eyewitnesses, potential military actors (inferred from combat context)
- Event Type: Missile strike in a combat zone
- Content Category: Combat actions and high violence explosive event

 # Verified Evidence

- **Source Details:**
  - Original Publisher: The footage was initially shared via a Twitter account (`https://twitter.com/cesarnews4/status/1521949838916521984`) and subsequently reshared on various video platforms.
  - Platform Specifics: The videos contain embedded overlays showing details such as time (e.g., 9:30 and a timestamp reading '04/05/2022 19:58:37') and location (Naberezhna Zavodska, Dnipro) that corroborate the metadata.
- **Where? (Location):**
  - Verified Location: Naberezhna Zavodska in Dnipro, Ukraine. Visual confirmation through recognizable landmarks (bridge structure and urban skyline) aligns with known images of Dnipro.
  - Approximate Coordinates: 48.4647° N, 35.0462° E.
- **When? (Time):**
  - Verified Timestamps: Overlays and metadata in the videos indicate the event occurred on 04/05/2022 with one video showing an exact timestamp of 19:58:37 local time. Additional visual cues (shadow positions, morning/evening light differences) support temporal claims.
- **Who? (Entities Involved):**
  - Individuals: The videos appear to be captured by hotel occupants or local eyewitnesses.
  - Groups: Indirect evidence suggests involvement of military forces executing a missile strike.
- **Why? (Motivation or Intent):**
  - Evidence-Based Assessment: The content was created to document a military/missile strike and highlight the ensuing damage in a combat zone, likely aimed at informing the public and possibly influencing opinion regarding the conflict in Ukraine.

`[Evidence Image](report/ID43-2.mp4_00_02_2.50s.jpg)`
`[Evidence Image](report/ID43-2.mp4_00_02_to_00_11_3.50s.jpg)`
`[Evidence Image](report/ID43-1.mp4_00_00_to_00_02_2.00s.jpg)`

 # Forensic Analysis

- **Authenticity Assessment:** Detailed frame analysis confirms the presence of embedded overlays (time, location, MAC-address information) that are consistent across multiple video segments, indicating the footage is not synthetically manipulated. The consistency in lighting, shadow patterns, and environmental cues further supports the video's authenticity.
- **Verification Tools & Methods:**
  - Frame-by-frame analysis and timestamp cross-referencing using digital forensic software.
  - Metadata inspection of video files confirming the overlay data (timestamps, location detail).
  - Comparison with satellite imagery and known landmarks of Dnipro to verify visual landmarks.
- **Synthetic Type (if applicable):** No anomalies typically associated with synthetic or deepfake content were identified using state-of-the-art detection tools.
- **Other Artifacts:** Minor compression artifacts were observed but are consistent with the distribution across social media platforms and do not indicate tampering.

 # Other Evidence & Findings

- **Supporting Sources:**
  - Multiple video analysis YAML files (`ID43-3_ID43-2`, `ID43-3_ID43-3`, `ID43-3_ID43-1`) consistently report the explosion at a bridge in Dnipro, Ukraine, with details such as missile impact, smoke plume, and associated timestamps.
  - The Twitter post (`https://twitter.com/cesarnews4/status/1521949838916521984`) serves as a primary reference, with additional reshares corroborating the event across platforms.
- **Cross-Checking Information:**
  - The location details (Naberezhna Zavodska, Dnipro) are cross-referenced with urban landmark images and satellite data of Dnipro, confirming the recognizable bridge structure.
  - Timestamps and temporal clues (morning/evening light, timestamp overlay '04/05/2022 19:58:37') align consistently across the different video sources.
- **Other Info:**
  - The metadata info regarding a missile strike on a bridge is supported by visual evidence in terms of large explosion flash, resultant fireball, and rising smoke observed in the videos.
  - The violent nature and high level of damage are further emphasized by the associated violence classification tags and detailed description across multiple evidentiary reports.

through the complete verification pipeline, generating a comprehensive verification report that demonstrates the practical application of our methodology. The Deep Researcher Agent successfully employed all four specialized verification tools, with the Verified News Tool extracting critical source details across the four key dimensions: spatial context (Where?), temporal context (When?), attribution context (Who?), and motivational context (Why?). The verification process successfully confirmed the authenticity of the submitted content, classifying it as "Verified" based on converging evidence from multiple sources. The system's ability to extract and cross-reference precise geolocation coordinates (48.4647° N, 35.0462° E), exact timestamps (04/05/2022, 19:58:37 local time), and

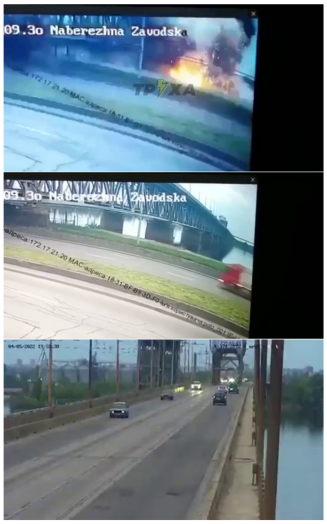

Figure 3: Detected key frames from raw data ID43-3.

detailed source attribution demonstrates the effectiveness of our multi-modal approach. The system's comprehensive source analysis is particularly noteworthy, tracing the content's origin to a specific Twitter account and documenting its subsequent distribution across multiple platforms. The forensic analysis component detected no signs of synthetic manipulation or deepfake content, while identifying minor compression artifacts consistent with legitimate social media distribution.

This result validates several key aspects of our proposed multi-agent deep research MLLMs multimedia verification system. The iterative nature of the Deep Researcher Agent enabled systematic evidence gathering from diverse sources, while the system's ability to handle complex, real-world content involving geopolitical events demonstrates its robustness and practical applicability to contemporary misinformation challenges.

6 Conclusion and Future Works

This paper presents our multi-agent multimedia verification system for the *ACMMM25 - Grand Challenge on Multimedia Verification*. Our approach combines MLLMs with specialized verification tools through a systematic six-stage pipeline that addresses both technical manipulation detection and contextual verification challenges. The Deep Researcher Agent's four specialized tools, particularly the verified news processor that extracts spatial, temporal, attribution, and motivational context, demonstrate effective integration of diverse verification approaches. Our evaluation on the challenge dataset cases shows the system can successfully handle complex scenarios involving geopolitical content while maintaining detailed source tracking and generating comprehensive verification reports.

Future work will focus on developing contestable AI multimedia verification systems that bridge automated efficiency with human agency across multiple dimensions. Our research future works encompass three primary directions: (1) **Scalability and Integration**, including real-time processing capabilities for cross verification responses [30], expanded multilingual verification across diverse linguistic contexts, and seamless integration with existing fact-checking workflows through standardized APIs and collaborative interfaces; (2) **Explainable AI (XAI)** methods that enable dynamic multimodal explanation generation [25], uncertainty quantification with confidence assessments [10], and source tracking for evidence-based decision making [6]; or going beyond explaining through (3) **Human-Centered Contestable AI** with tiered appeal systems, culturally-adaptive explanation interfaces that account for diverse stakeholder needs, and interactive verification workflows that preserve human oversight while maintaining system efficiency [1, 18, 26, 32].

References

[1] Kars Alfrink et al. 2023. Contestable AI by design: Towards a framework. *Minds and Machines* 33, 4 (2023), 613–639.

[2] Rajaa Alqudah, Mohammed Al-Qaisi, Rakan Ammari, and Yazan Abu Ta'a. 2023. OSINT-Based Tool for Social Media User Impersonation Detection Through Machine Learning. In *2023 International Conference on Information Technology (ICIT)*. IEEE, 752–757.

[3] Shivangi Aneja, Cise Midoglu, Duc-Tien Dang-Nguyen, Sohail Ahmed Khan, Michael Riegler, Pål Halvorsen, Chris Bregler, and Balu Adsumilli. 2022. Acm multimedia grand challenge on detecting cheapfakes. *arXiv preprint arXiv:2207.14534* (2022).

[4] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc Tien Dang Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, et al. 2015. Verifying multimedia use at mediaeval 2015. In *MediaEval 2015*. Vol. 1436. CEUR-WS.

[5] Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2024. DEFAME: Dynamic Evidence-based FAct-checking with Multimodal Experts. *arXiv preprint arXiv:2412.10510* (2024).

[6] Swathi Chundru. 2021. Leveraging AI for Data Provenance: Enhancing Tracking and Verification of Data Lineage in FATE Assessment. *International Journal of Inventions in Engineering & Science Technology* 7, 1 (2021), 87–104.

[7] Yandex Cloud. 2025. *Yandex Cloud Documentation | Yandex Search API | Web Search API, gRPC: ImageSearchService*. https://yandex.cloud/en/docs/search-api/api-ref/grpc/ImageSearch/

[8] Duc-Tien Dang-Nguyen, Morten Langfeldt Dahlback, Henrik Vold, Silje Førsund, Minh-Son Dao, Kha-Luan Pham, Sohail Ahmed Khan, Marc Gallofré Ocaña, Minh-Triet Tran, and Anh-Duy Tran. 2025. The 2025 Grand Challenge on Multimedia Verification: Foundations and Overview. In *Proceedings of the 33rd ACM International Conference on Multimedia*. xxxx–yyyy.

[9] Minh-Son Dao and Koji Zettsu. 2023. Leveraging knowledge graphs for cheapfakes detection: Beyond dataset evaluation. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 99–104.

[10] Jessica Deuschel, Andreas Foltyn, Karsten Roscher, and Stephan Scheele. 2024. The role of uncertainty quantification for trustworthy AI. In *Unlocking Artificial Intelligence: From Theory to Applications*. Springer, 95–115.

[11] Sharad Duwal, Mir Nafis Sharear Shopnil, Abhishek Tyagi, and Adiba Mahbub Proma. 2025. Evidence-Grounded Multimodal Misinformation Detection with Attention-Based GNNs. *arXiv preprint arXiv:2505.18221* (2025).

[12] Dhanvi Ganti. 2022. A novel method for detecting misinformation in videos, utilizing reverse image search, semantic analysis, and sentiment comparison of metadata. *Utilizing Reverse Image Search, Semantic Analysis, and Sentiment Comparison of Metadata (June 5, 2022)* (2022).

[13] Xingyu Gao, Xi Wang, Zhenyu Chen, Wei Zhou, and Steven CH Hoi. 2024. Knowledge enhanced vision and language model for multi-modal fake news detection. *IEEE Transactions on Multimedia* (2024).

[14] Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. 2024. Logical Consistency of Large Language Models in Fact-checking. *arXiv preprint arXiv:2412.16100* (2024). arXiv:2412.16100

[15] Sonal Goel, Niharika Sachdeva, Ponnurangam Kumaraguru, A. V. Subramanyam, and Divam Gupta. 2016. PicHunt: Social Media Image Retrieval for Improved Law Enforcement. In *Social Informatics*, Emma Spiro and Yong-Yeol Ahn (Eds.). Springer International Publishing, Cham, 206–223.

[16] Haiying Guan. 2025. *NIST Open Media Forensics Challenge (OpenMFC Briefing for IIRD)*.

[17] Arash Heidari, Nima Jafari Navimipour, Hasan Dag, and Mehmet Unal. 2024. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 14, 2 (2024), e1520.

[18] Naveena Karusala, Sohini Upadhyay, Rajesh Veeraraghavan, and Krzysztof Z Gajos. 2024. Understanding Contestability on the Margins: Implications for the Design of Algorithmic Decision-making in Public Services. In *Proceedings of the*

2024 CHI Conference on Human Factors in Computing Systems. 1–16.

[19] Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401* (2024).

[20] Ashish Kumar, Divya Singh, Rachna Jain, Deepak Kumar Jain, Chenquan Gan, and Xudong Zhao. 2025. Advances in DeepFake detection algorithms: Exploring fusion techniques in single and multi-modal approach. *Information Fusion* (2025), 102993.

[21] Kumud Lakara, Juil Sock, Christian Rupprecht, Philip Torr, John Collomosse, and Christian Schroeder de Witt. 2024. MAD-Sherlock: Multi-Agent Debates for Out-of-Context Misinformation Detection. *arXiv preprint arXiv:2410.20140* (2024).

[22] Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 10154–10163.

[23] Dale Meredith. 2024. *The OSINT Handbook: A practical guide to gathering and analyzing online information*. Packt Publishing Ltd.

[24] Bao-Tin Nguyen, Van-Loc Nguyen, Thanh-Son Nguyen, Duc-Tien Dang-Nguyen, Trong-Le Do, and Minh-Triet Tran. 2024. A Hybrid Approach for Cheapfake Detection Using Reputation Checking and End-To-End Network. In *Proceedings of the 1st Workshop on Security-Centric Strategies for Combating Information Disorder*. 1–12.

[25] Hung Nguyen et al. 2024. LangXAI: Integrating Large Vision Models for Generating Textual Explanations to Enhance Explainability in Visual Perception Tasks. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 8754–8758. doi:10.24963/ijcai.2024/1025 Demo Track.

[26] Hung Nguyen, Alireza Rahimi, Veronica Whitford, Hélène Fournier, Irina Kondratova, René Richard, and Hung Cao. 2025. Heart2Mind: Human-Centered Contestable Psychiatric Disorder Diagnosis System using Wearable ECG Monitors. *arXiv preprint arXiv:2505.11612* (2025).

[27] Minh-Tam Nguyen, Quynh T Nguyen, Minh Son Dao, and Binh T Nguyen. 2025. Multimodal scene-graph matching for cheapfakes detection. *International Journal of Multimedia Information Retrieval* 14, 2 (2025), 17.

[28] Thanh-Son Nguyen, Vinh Dang, Minh-Triet Tran, and Duc-Tien Dang-Nguyen. 2023. Leveraging cross-modals for cheapfakes detection. In *Proceedings of the 4th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*. 51–59.

[29] Thanh-Son Nguyen and Minh-Triet Tran. 2023. Multi-Models from Computer Vision to Natural Language Processing for Cheapfakes Detection. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 93–98.

[30] Van-Hoang Phan, Long-Khanh Pham, Dang Vu, Anh-Duy Tran, and Minh-Son Dao. 2025. E-FreeM2: Efficient Training-Free Multi-Scale and Cross-Modal News Verification via MLLMs. *arXiv preprint arXiv:2506.20944* (2025).

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

[32] Timothée Schmude. 2025. Explainability and Contestability for the Responsible Use of Public Sector AI. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[33] Quang-Tien Tran, Thanh-Phuc Tran, Minh-Son Dao, Tuan-Vinh La, Anh-Duy Tran, and Duc Tien Dang Nguyen. 2022. A textual-visual-entailment-based unsupervised algorithm for cheapfake detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7145–7149.

[34] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided Direct Preference Optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 13258–13273. doi:10.18653/v1/2024.findings-emnlp.775