

Contestable Multi-Agent Debate with Arena-based Argumentative Computation for Multimedia Verification

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

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

Hoang-Loc Cao
University of Science, VNU-HCM
Ho Chi Minh City, Vietnam
chloc22@clc.fitus.edu.vn

Phuc Ho
University of Science, VNU-HCM
Ho Chi Minh City, Vietnam
hdphuc22@clc.fitus.edu.vn

Van Pham
University of Science, VNU-HCM
Ho Chi Minh City, Vietnam
pavan22@clc.fitus.edu.vn

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

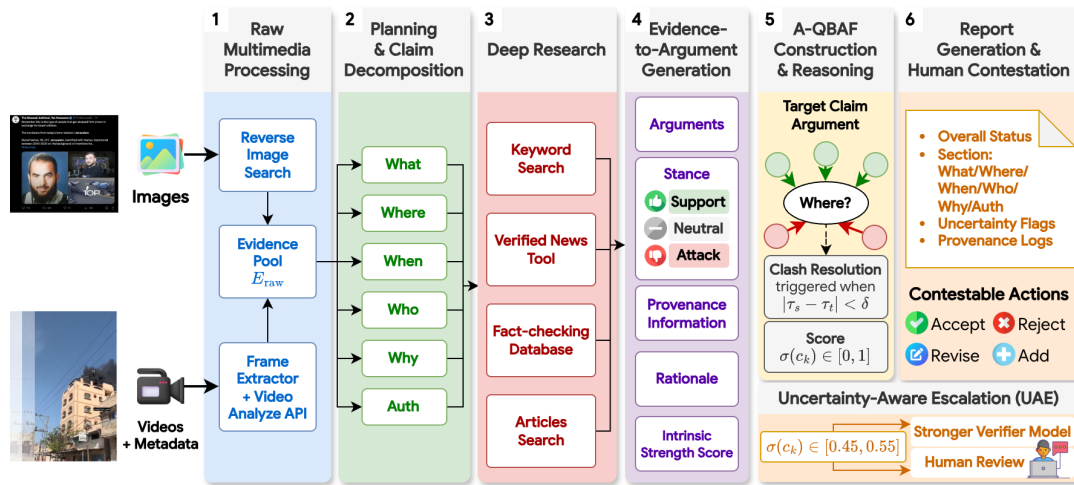


Figure 1: Architecture of Contestable Multi-Agent Debate with A-QBAF for Multimedia Verification.

Abstract

Multimedia verification requires not only accurate conclusions but also transparent and contestable reasoning. We propose a contestable multi-agent framework that integrates multimodal large language models, external verification tools, and arena-based quantitative bipolar argumentation (A-QBAF) as a submission to the *ICMR 2026 Grand Challenge on Multimedia Verification*. Our method decomposes each case into claim-centered sections, retrieves targeted evidence, and converts evidence into structured support and attack arguments with provenance and strength scores. These arguments are resolved through small local argument graphs with selective clash resolution and uncertainty-aware escalation. The resulting system generates section-wise verification reports that are transparent, editable, and computationally practical for real-world

*Both authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

ICMR '26, Amsterdam, Netherlands

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2617-0/2026/06

<https://doi.org/10.1145/3805622.3812606>

multimedia verification. Our implementation is public at: https://github.com/Analytics-Everywhere-Lab/MV2026_the_liems.

CCS Concepts

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

Keywords

Multimedia Verification, Contestable AI, Multimodal LLM

ACM Reference Format:

Truong Thanh Hung Nguyen, Vo Thanh Khang Nguyen, Hoang-Loc Cao, Phuc Ho, Van Pham, and Hung Cao. 2026. Contestable Multi-Agent Debate with Arena-based Argumentative Computation for Multimedia Verification. In *International Conference on Multimedia Retrieval (ICMR '26)*, June 16–19, 2026, Amsterdam, Netherlands. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3805622.3812606>

1 Introduction and Related Works

Multimedia verification is increasingly important because online misinformation is often conveyed through images and videos rather than solely through text [4–6]. Prior work has focused on two main

problems: detecting manipulated or synthetic media such as deep-fakes [26], and identifying authentic but miscontextualized content, often called cheapfakes or out-of-context misuse [2, 10, 24]. As a result, research has expanded from classical media forensics to evidence-based verification that combines visual analysis, metadata inspection, reverse image search, fact-checking resources, and source credibility assessment [14, 22, 25, 28].

Recent studies increasingly use multimodal large language models (MLLMs) and hybrid vision-language systems to jointly reason over images, videos, and text [11, 20, 21, 27]. These systems can also incorporate external tools such as reverse image search, metadata analysis, knowledge graphs, and fact-checking databases [7, 9, 13, 18, 19]. More recently, agent-based approaches have decomposed verification into multi-step workflows in which LLMs coordinate retrieval, tool use, and evidence synthesis [12, 15].

However, most existing approaches still emphasize prediction accuracy or final evidence synthesis rather than *contestability* [1, 16, 17, 23]. In practice, they often compress heterogeneous evidence into a single conclusion, making it difficult for users to inspect supporting and attacking evidence, challenge intermediate reasoning, or revise conclusions when evidence is incomplete or disputed [8]. This limitation is especially important in high-stakes multimedia verification, where conflicting signals, uncertain provenance, and interpretive claims such as *why* a post was shared are common.

To address this gap, we propose a *contestable multimedia verification* framework that combines MLLMs, external verification tools, and an arena-based quantitative bipolar argumentation framework (A-QBAF) [3]. Instead of directly summarizing evidence into a final answer, our method decomposes verification into claim-centered questions and converts evidence into structured support and attack arguments [29]. These arguments are organized into small local graphs, enabling transparent reasoning, efficient conflict handling, and human-editable contestation. The framework also includes uncertainty-aware escalation so that ambiguous cases can be deferred to a stronger verifier or human review. Our goal is to produce verification reports that are evidence-grounded, transparent, editable, and practical for real-world multimedia analysis. Our contributions are as follows:

- (1) We propose a **contestable multi-agent multimedia verification** framework that integrates MLLMs, external verification tools, and claim-centered A-QBAF reasoning to assess both authenticity and contextual integrity.
- (2) We introduce an **evidence-to-argument formulation** that converts retrieved multimodal evidence into structured support and attack arguments with provenance, rationale, and strength scores, enabling users to inspect, accept, reject, edit, or add arguments during verification.
- (3) We design an **efficient reasoning pipeline** based on small claim graphs, top- k evidence selection, sparse relation construction, selective clash resolution, and uncertainty-aware escalation, producing structured reports that remain transparent and contestable without excessive computation.

2 Proposed Method

We propose a contestable multimedia verification framework that combines MLLMs, external verification tools, and the A-QBAF [3].

The main goal of our method is to verify both the authenticity and the context of multimedia content while keeping the reasoning process transparent, editable, and computationally efficient.

Our system contains six stages: (1) raw multimedia processing, (2) planning and claim decomposition, (3) section-wise deep research, (4) evidence-to-argument conversion, (5) A-QBAF reasoning with selective conflict resolution, (6) report generation and human contestation with uncertainty-aware escalation (UAE).

Our system replaces direct evidence synthesis with a lightweight argumentation layer. Retrieved evidence is first converted into structured arguments, then organized into support-attack relations around six target claims: *what*, *where*, *when*, *who*, *why*, and *authenticity*. The framework is designed for practical multimedia verification: it handles both images and videos, integrates evidence from multiple tools and sources, identifies contradictions, and produces a structured report rather than only a binary label. To remain efficient, we use a *section-wise micro-graph* design, where each claim is verified with a small local argument graph instead of a single global graph. This improves traceability and makes the final decision easier to inspect and contest.

2.1 Raw Multimedia Processing

The first stage prepares the multimedia input for later verification. Given a case package \mathcal{I} , the system receives:

$$\mathcal{I} = \{m, d, u\}, \quad (1)$$

where m is the primary media content (image or video), d is the available metadata and textual context, and u represents optional external clues such as source links, captions, or fact-checking notes.

For video input, the system uses an MLLM analyzer to generate frame-level descriptions, extract visible text, identify key objects and events, and collect technical metadata. A keyframe extractor selects the most informative frames to reduce redundancy and later search cost. For image input, the system performs reverse image search and collects visually related web pages, reposts, and possible original sources. This stage outputs a normalized evidence pool:

$$E_{raw} = \{e_1, e_2, \dots, e_n\}, \quad (2)$$

where each evidence may come from a frame, metadata field, OCR result, ASR transcript, reverse search result, or retrieved article.

2.2 Planning and Claim Decomposition

The planner agent coordinates the verification workflow. Instead of treating the case as one single prediction problem, the planner decomposes it into a set of claim-centered verification questions:

$$C = \{c^{what}, c^{where}, c^{when}, c^{who}, c^{why}, c^{auth}\}. \quad (3)$$

These claims are defined as follows:

- c^{what} : what source event or situation is shown,
- c^{where} : where the content was recorded,
- c^{when} : when the event happened,
- c^{who} : which people, groups, or sources are involved,
- c^{why} : why content was posted or what narrative it supports,
- c^{auth} : whether the media is authentic, edited, synthetic, or recaptured.

This decomposition is important for two reasons. First, some claims are more objective than others. For example, *where* and *when*

can often be grounded by landmarks, timestamps, or metadata, while *why* is usually more interpretive and requires stronger explanation. Second, not every case needs the same amount of reasoning. By separating claims, the system can apply more computation only to the difficult parts.

2.3 Deep Research

After claim decomposition, the deep researcher agent gathers targeted evidence for each claim. For a claim $c_k \in C$, the agent retrieves a small relevant evidence subset:

$$E_k = \text{Top}^k(E_{\text{raw}}, c_k, k), \quad (4)$$

where Top^k selects the most relevant evidence items for the target claim. In practice, we keep k small to control cost.

The deep researcher combines three types of operations:

- (1) **keyword-guided search**: generate search queries from the media content and textual clues;
- (2) **tool-assisted verification**: use reverse image search, metadata inspection, OCR/ASR, and fact-checking databases;
- (3) **source analysis**: inspect article credibility, publication time, repost chains, and cross-source consistency.

This stage is especially important for claims such as *where*, *when*, *who*, and *why*. For example, a verified news source may provide explicit location and date information, while a video frame may contain overlays or landmarks that either support or contradict that source. Our method keeps all such signals instead of collapsing them too early into a free-form summary.

2.4 Evidence-to-Argument Generation

The main contribution of our method begins at this stage. We convert retrieved evidence into structured arguments. For each evidence item $e_i \in E_k$, the system creates an argument card:

$$a_i = \{t_i, s_i, p_i, r_i, \tau_i\}, \quad (5)$$

- t_i is the textual form of the argument,
- $s_i \in \{\text{support, attack, neutral}\}$ is its stance toward the claim,
- p_i stores provenance information,
- r_i is a short rationale explaining why the evidence matters,
- $\tau_i \in [0, 1]$ is the intrinsic strength score.

Each argument is written in a simple, evidence-grounded form. For instance, in a *when* claim, one argument may state that “The post was published on *May 7, 2025*, during the India–Pakistan escalation.” While another may attack the same claim by noting that the event timestamps were mismatched, i.e., “Pakistani press coverage of the crash appeared on *April 16, 2025*, well before Operation Sindoor.”

Intrinsic Strength Attribution. We assign a base score to each argument using four lightweight criteria:

$$\tau_i = \lambda_1 q_i^{\text{src}} + \lambda_2 q_i^{\text{corr}} + \lambda_3 q_i^{\text{mod}} + \lambda_4 q_i^{\text{rel}}, \quad (6)$$

where q_i^{src} is source reliability, q_i^{corr} is cross-source corroboration, q_i^{mod} is cross-modal consistency, q_i^{rel} is claim relevance. The weights satisfy $\sum_j \lambda_j = 1$. In our implementation, this score is intentionally simplified. We avoid a fully debate-based scoring process for every argument because it would be overly expensive for multimedia cases with many noisy observations.

A direct multi-agent debate over all evidence would increase computational cost and often lead to redundant information. Our formulation is cheaper because we first select a small evidence subset for each claim, then convert only those items into short argument cards. This retains structure without creating unnecessarily long reasoning traces.

2.5 A-QBAF Construction and Reasoning

For each claim c_k , we build one small A-QBAF [3]:

$$\mathcal{G}_k = \langle X_k, R_k^-, R_k^+, \tau_k \rangle, \quad (7)$$

where $X_k = \{c_k\} \cup A_k$ is the set of nodes, including the target claim and its arguments; R_k^+ is the set of support edges; R_k^- is the set of attack edges; τ_k is the base strength function. The target claim node is initialized with neutral strength $\tau(c_k) = 0.5$.

If argument a_i supports the claim, then $(a_i, c_k) \in R_k^+$. If it attacks the claim, then $(a_i, c_k) \in R_k^-$. We also allow limited argument-to-argument edges when two arguments clearly reinforce or contradict each other. However, to keep the graph sparse, we only add these edges in obvious cases, such as:

- two arguments based on the same visual cue but opposite interpretations,
- metadata directly contradicting a reported date or location,
- two independent trusted sources confirming the same fact.

This sparse graph design is important. Full pairwise relation prediction across all arguments is expensive and often unnecessary in multimedia verification. In many cases, the main structure is already captured by argument-to-claim edges.

2.6 Report Generation and Human Contestation

2.6.1 Quantitative Reasoning and Conflict Calibration. After graph construction, we propagate argument influence using quantitative bipolar reasoning. For a node $x \in X_k$, its total incoming energy is:

$$E(x) = \sum_{y \in \text{Sup}(x)} \sigma(y) - \sum_{z \in \text{Att}(x)} \sigma(z), \quad (8)$$

where $\sigma(\cdot)$ is the propagated strength.

We use the quadratic energy function: $h(x) = \frac{\max(x, 0)^2}{1 + \max(x, 0)^2}$. The final equilibrium score of node x is:

$$\sigma(x) = \tau(x) + (1 - \tau(x))h(E(x)) - \tau(x)h(-E(x)). \quad (9)$$

The final claim score $\sigma(c_k)$ reflects the balance between supporting and attacking evidence. A high score means the claim is strongly supported by the available evidence. A low score means the claim is weakened by contradictions. A score near 0.5 means the evidence remains inconclusive.

Selective Clash Resolution (CR). A common problem in LLM-based reasoning is score saturation, in which both support and attack arguments appear equally strong. To address this, we added a lightweight CR step. For a support–attack pair (a_s, a_t) , if their base scores are close $|\tau(a_s) - \tau(a_t)| < \delta$, we trigger one additional comparison by a judge model. This judge decides which argument is stronger, given the claim and the case evidence. The score update is: $\Delta\tau(a) = \beta(2w - 1)$, where $w \in [0, 1]$ is the win rate of the argument across clashes and β is a small adjustment magnitude. We

only apply CR to ambiguous sections that benefit most from conflict calibration. For easier sections, such as locations with strong landmark evidence, this extra step is usually skipped.

2.6.2 Human Contestation and Editable Verification. One important goal of our method is contestability. The user should not only read the final report, but also inspect and challenge the reasoning process. Therefore, for each claim, the system presents: (1) the claim statement; (2) the main support and attack arguments; (3) provenance links to frames, metadata, or articles; (4) both intrinsic and propagated scores; and (5) a short explanation of why each argument affects the conclusion.

Users may perform four contestable actions: (1) accept an argument, (2) reject an argument, (3) edit an argument, or (4) add a missing argument with evidence. After any accepted edit, the graph is updated and all claim scores are recomputed. In this way, contestation changes the final result directly rather than appearing as post-hoc feedback. This design is especially useful for difficult claims such as *why*. In many real cases, the system cannot fully prove the motivation behind a post, but it can show which arguments support a likely interpretation and which arguments weaken it. This makes the result more honest and easier to audit.

2.6.3 Uncertainty-Aware Escalation (UAE). However, not all cases warrant a definitive system-level decision. If a claim score remains close to the neutral region, the system should admit uncertainty and request a stronger review. For claim c_k , if: $\sigma(c_k) \in [0.45, 0.55]$, we mark the section as uncertain. For uncertain claims, we use one of two escalation options: (1) invoke a stronger verifier model for one final judgment, or (2) defer the claim to human review. This mechanism prevents the system from giving overconfident conclusions when the evidence is balanced or incomplete. It is particularly important for authenticity judgments and narrative interpretation, where both false positives/negatives are costly.

2.7 Final Report Generation

The final report combines section scores into a structured verification document. Instead of a single binary label, the report includes: (1) overall verification status, (2) section-wise findings for *what*, *where*, *when*, *who*, *why*, and *authenticity*, (3) key supporting and attacking evidence, (4) uncertainty notes for inconclusive claims, (5) provenance records and contestation logs. This report format matches the practical needs of multimedia fact-checking. It explains not only what the system concludes, but also why the conclusion was reached, which evidence supports it, which evidence challenges it, and which parts still remain uncertain.

Our method is designed to improve contestability without incurring high computational cost. It achieves this through several efficiency choices, i.e., claim decomposition into small local graphs, top- k evidence selection for each claim, sparse relation construction, selective CR only for near-tied arguments, and escalation only for uncertain cases, where the input evidence is multimodal, noisy, and often large-scale. Our method adds argument structure only where it is most useful, rather than turning the whole pipeline into a heavy debate system.

3 Demonstrative Results and Conclusion

To illustrate the performance of our framework, we present an argumentative reasoning example using sample ID01 from the *ICMR 2026 Grand Challenge on Multimedia Verification* [6] validation set. In this experiment, Gemini 3.1 Pro, GPT 5.4, and Claude Opus 4.6 were used as debaters to generate evidence-grounded arguments, which were then organized and resolved through the proposed reasoning A-QBAF.

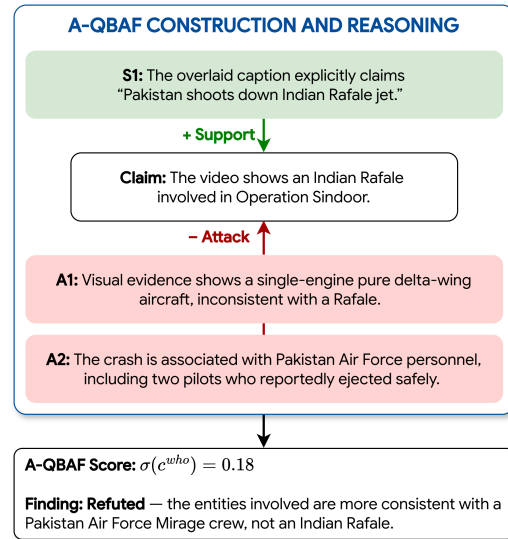


Figure 2: The final section-wise conclusion for *who* (ID01 Validation) via argumentative A-QBAF computation.

Fig. 2 shows the final reasoning outcome for the *who* claim. The target claim states that the video shows an Indian Rafale involved in Operation Sindoor. One supporting argument is derived from the overlaid caption. However, two attacking arguments receive stronger overall influence in the graph. After A-QBAF reasoning, the final claim score is $\sigma(c^{who}) = 0.18$, indicating that the claim is strongly weakened by the available evidence. The resulting section-level finding is therefore *refuted*.

This example shows the advantages of our framework, i.e., transparent support-attack reasoning, human contestability through editable arguments and score recomputation, and practical efficiency through a small local graph. It also highlights uncertainty calibration. Since the final score lies well outside the uncertainty band defined in Sec 2.6.1, this claim does not require escalation to a stronger verifier or human review. More broadly, the same claim-centered process can be applied to *what*, *where*, *when*, *who*, *why*, and *authenticity*, producing a section-wise report that is more informative than a single binary label.

In conclusion, we propose a contestable multimedia verification framework by combining evidence retrieval, evidence-to-argument conversion, A-QBAF reasoning, selective CR, and UAE. The framework produces verification reports that are transparent, editable, and practical for high-stakes multimedia analysis.

Acknowledgments

This work was supported by NSERC Discovery Grant No RGPIN-2025-04478 and NSERC Discovery Supplement Award No DGECR-2025-00129.

References

- [1] Kars Alfrink et al. 2023. Contestable AI by design: Towards a framework. *Minds and Machines* 33, 4 (2023), 613–639.
- [2] 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).
- [3] Hoang-Loc Cao, Phuc Ho, Truong Thanh Hung Nguyen, Phuc Truong Loc Nguyen, Dinh Thien Loc Nguyen, and Hung Cao. 2026. Adaptive Collaboration of Arena-Based Argumentative LLMs for Explainable and Contestable Legal Reasoning. *arXiv preprint arXiv:2602.18916* (2026).
- [4] Duc-Tien Dang-Nguyen, Sohail Ahmed Khan, Michael Riegler, Pål Halvorsen, Anh-Duy Tran, Minh-Son Dao, and Minh-Triet Tran. 2024. Overview of the grand challenge on detecting cheapfakes at acm icmr 2024. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*. 1275–1281.
- [5] 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*.
- [6] Duc-Tien Dang-Nguyen, Kha-Luan Pham, Minh-Anh Pham, Silje Førsund, Henrik B. Vold, Minh-Triet Tran, and Anh-Duy Tran. 2026. The 2026 Grand Challenge on Multimedia Verification: Overview and Key Directions. In *Proceedings of the 2026 International Conference on Multimedia Retrieval (Amsterdam, Netherlands) (ICMR '26)*. Association for Computing Machinery, New York, NY, USA.
- [7] 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.
- [8] Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2025. Argumentative Large Language Models for Explainable and Contestable Claim Verification. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence, AAAI 2025*, Vol. 39. Philadelphia, PA, USA, 14930–14939.
- [9] 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).
- [10] Haiying Guan. 2025. *NIST Open Media Forensics Challenge (OpenMFC Briefing for IIRD)*.
- [11] Tuan-Vinh La, Quang-Tien Tran, Thanh-Phuc Tran, Anh-Duy Tran, Duc-Tien Dang-Nguyen, and Minh-Son Dao. 2022. Multimodal cheapfakes detection by utilizing image captioning for global context. In *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*. 9–16.
- [12] Huy Hoan Le, Van Sy Thinh Nguyen, Thi Le Chi Dang, Vo Thanh Khang Nguyen, Truong Thanh Hung Nguyen, and Hung Cao. 2025. Multimedia Verification Through Multi-Agent Deep Research Multimodal Large Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 14034–14040.
- [13] Dale Meredith. 2024. *The OSINT Handbook: A practical guide to gathering and analyzing online information*. Packt Publishing Ltd.
- [14] Eivind Moholdt, Sohail Ahmed Khan, and Duc-Tien Dang-Nguyen. 2023. Detecting out-of-context image-caption pair in news: A counter-intuitive method. In *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*. 203–209.
- [15] 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.
- [16] Hung Nguyen, Alireza Rahimi, Veronica Whitford, H el ene Fournier, Irina Kondratova, Ren e Richard, and Hung Cao. 2026. Heart2Mind: Human-Centered Contestable Psychiatric Disorder Prediction System Using Wearable ECG Monitors. *ACM Transactions on Computing for Healthcare* (2026).
- [17] Loc Phuc Truong Nguyen, Hung Thanh Do, Hung Truong Thanh Nguyen, and Hung Cao. 2025. Motion2Meaning: A Clinician-Centered Framework for Contestable LLM in Parkinson’s Disease Gait Interpretation. In *Proceedings of 9th International Symposium on Chatbots and Human-centred AI (CONVERSATIONS) 2025*.
- [18] Minh-Nhat Nguyen, Trong-Nghia Tran, Minh-Triet Tran, Duc-Tien Dang-Nguyen, and Trong-Le Do. 2025. Robust Multimedia Verification of Cheapfakes and Deepfakes via External Context Leveraging. In *2025 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–8.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] Kha-Luan Pham, Manh-Thien Nguyen, Anh-Duy Tran, Minh-Son Dao, and Duc-Tien Dang-Nguyen. 2023. Detecting cheapfakes using self-query adaptive-context learning. In *Proceedings of the 4th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*. 60–63.
- [23] Thomas Ploug and S oren Holm. 2020. The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial intelligence in medicine* 107 (2020), 101901.
- [24] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nie bner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.
- [25] Jangwon Seo, Hyo-Seok Hwang, Jiyoung Lee, Minhyeok Lee, Wonsuk Kim, and Junhee Seok. 2024. A Multi-Stage Deep Learning Approach Incorporating Text-Image and Image-Image Comparisons for Cheapfake Detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*. 1312–1316.
- [26] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information fusion* 64 (2020), 131–148.
- [27] 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.
- [28] Simon S Woo. 2022. Advanced Machine Learning Techniques to Detect Various Types of Deepfakes. In *Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes*. 25–25.
- [29] Yuqicheng Zhu, Nico Potyka, Daniel Hern andez, Yuan He, Zifeng Ding, Bo Xiong, Dongzhuoran Zhou, Evgeny Kharlamov, and Steffen Staab. 2025. ArgRAG: Explainable Retrieval Augmented Generation using Quantitative Bipolar Argumentation. In *19th International Conference on Neurosymbolic Learning and Reasoning*.