International Conference on Intelligent Computing and Explainable AI (ICICEA) – April 2026

RESEARCH ARTICLE OPEN ACCESS

Interpretable Intelligence: A Novel XAI Approach to Multimodal Misinformation Detection

M.Sushmitha Sharan[1], P.R. Tharshini[2], P. Lohitha[3], B. Roshini[4]
Computer Science and Engineering, A.V.C College of Engineering, Mayiladuthurai, TamilNadu
sushmithasharan110706@email.com

Abstract

The proliferation of multimodal misinformation—content combining manipulated imagery with misleading textual narratives—poses unprecedented challenges to information integrity in digital ecosystems. While recent advances in vision-language models have demonstrated remarkable detection capabilities, their black-box nature fundamentally limits deployability in high-stakes scenarios such as electoral integrity and public health communication. Existing explainability approaches predominantly operate within a single modality, failing to capture the cross-modal interactions that often constitute the essence of multimodal misinformation. This paper introduces CMA-X, a novel cross-modal attribution framework specifically designed for explaining multimodal misinformation detection. Unlike conventional approaches that treat visual and textual modalities independently, our method explicitly models interactions between image regions and text tokens through a principled attribution mechanism grounded in integrated gradients and cross-8500-modal Hessian computation. We evaluate our framework across three benchmark datasets comprising real-world misinformation instances spanning political, health, and general news domains. Quantitative results demonstrate that our method achieves 22.4% improvement in faithfulness compared to single-modality baselines, with a 31.7% increase in explanation stability.

Keywords: Explainable AI, Multimodal Misinformation Detection, Cross-Modal Attribution, Integrated Gradients, Trustworthy AI, Vision-Language Models, Factchecking, Interpretability

1. Introduction

1.1 The Rising Challenge of Multimodal Deception

Digital platforms today deliver content in ways that were uncommon just a decade ago. A user scrolling through their feed encounters photographs paired with descriptive captions, videos accompanied by explanatory text, and illustrations integrated within articles. This blend of visual and textual elements has become the standard format for online communication, shaping how billions of people receive information daily. While this multimodal presentation offers advantages in engagement and accessibility, it also opens pathways for those who seek to spread false narratives.

When information arrives through multiple channels simultaneously, people tend to accept it more readily than information presented through a single channel. A photograph creates an immediate sense of authenticity that text alone cannot be established. A well-crafted caption provides context that shapes how viewers interpret what they see. Together, these elements

form a unified impression that feels credible, even when individual components may contain deceptions. Those who create misleading content understand this dynamic well and exploit it deliberately.

The methods for producing deceptive multimodal content have grown more sophisticated over time. Some creators take authentic photographs from past events and attach them to text describing current situations, creating temporal mismatches that casual observers rarely notice. Others use accessible editing tools to modify specific portions of images while leaving surrounding areas untouched, producing visuals that appear genuine except under scrutiny. More recently, advances in generative models have enabled the creation of entirely synthetic images that depict scenarios that never occurred, paired with text crafted to provoke strong reactions. Each new development in content creation technology provides additional tools for those who wish to distribute false information on a large scale.

The impact of such content extends far beyond individual confusion. Election integrity has been threatened in multiple countries by coordinated campaigns distributing manipulated image-text pairs designed to undermine confidence in democratic processes. Public health efforts have faced significant obstacles when fabricated visual evidence contradicts established medical guidance, leading to measurable effects on vaccination rates and treatment adherence. Financial markets have experienced volatility following the release of convincingly fabricated corporate communications or manipulated images of economic indicators. These real-world consequences demonstrate that multimodal deception is not merely an inconvenience but a force capable of influencing outcomes that affect people's lives in tangible ways.

1.2. Introducing CMA-X: A Framework for Cross-Modal Explanation

To address the needs outlined above, we present CMA-X, a framework designed to generate explanations for multimodal deception detection systems. Our approach places cross-modal interactions at the centre of explanation, explicitly modelling how image regions and text tokens combine to influence detection outcomes. We ground our method in established attribution principles while extending these principles to capture interaction effects. The framework produces visual attribution maps, textual importance scores, and cross-modal interaction maps that together provide a comprehensive view of the evidence underlying each detection.

We evaluate CMA-X using three collections of real-world content spanning political, health, and general news domains. Our experimental results demonstrate substantial improvements in faithfulness compared to approaches that treat modalities independently, with corresponding gains in explanation stability. User studies involving participants, including professional fact-checkers, show that cross-modal explanations reduce verification time while improving user trust in system outputs.

2.Related work

2.1 Multimodal Misinformation Detection

Early approaches to detecting deceptive image text pairs relied on handcrafted features extracted separately from each modality, combined through simple fusion mechanisms. The introduction of deep learning enabled automatic feature extraction through convolutional networks for images and recurrent networks for text, leading to improved detection performance.

The emergence of vision-language pretrained models such as CLIP and ViLBERT represented a significant advancement. These models learn joint representations from large-scale image-text corpora and achieve strong performance on misinformation benchmarks after fine-tuning. However, existing detection systems face notable limitations. Models often learn dataset-specific patterns rather than general principles of deception, leading to poor generalization of out of-distribution examples. Additionally, these systems exploit spurious correlations present in training data rather than learning to detect genuine cross-modal inconsistencies. Most

critically for our work, these models operate as black boxes that cannot explain their decisions, limiting their deploy ability in high-stakes contexts such as electoral integrity and public health communication.

2.2. Explainability Artificial Intelligence

The field of explainable AI has developed numerous techniques for understanding model predictions. Attribution-based methods assign importance scores to input features based on their contribution to outputs. LIME generates local explanations by training interpretable surrogate models on perturbed inputs. SHAP provides a unified framework grounded in cooperative game theory, assigning importance values that satisfy properties such as consistency. Integrated Gradients offers a gradient-based approach with axiomatic guarantees including sensitivity and completeness, making it particularly suitable for deep neural networks.

Beyond attribution, example-based explanations identify influential training instances or retrieve similar examples to illustrate model behavior. Concept-based approaches operate at higher levels of abstraction, linking internal representations to human-understandable concepts through techniques such as Concept Activation Vectors.

Evaluation of explanations has traditionally focused on technical metrics including faithfulness, stability, and completeness. However, these measures do not guarantee that explanations help human users make better decisions. Recent work has called for human centered evaluation approaches that assess practical utility with real stakeholders, a direction we adopt in this paper.

2.3. Explainability Multimodal System

Research on explainability for multimodal systems remains comparatively limited. The predominant approach applies single-modality methods independently to visual and textual inputs, presenting explanations side by side. For example, Grad-CAM generates heatmaps for images while LIME identifies important text tokens, with both outputs displayed together. While this provides some transparency, it treats modalities independently and cannot capture interactions where deception arises from relationships between images and text.

Joint visualization approaches attempt to generate presentation through spatial alignment or attention mechanisms but typically do not modify underlying attribution methods to capture interaction effects. Some work has adapted explanation techniques for vision-language models in tasks such as visual question answering and image captioning, demonstrating that cross-modal interactions significantly influence predictions.

2.4 Research Gap and Our Contribution

ASPECTS	EXISTING WORK	CMA-X
Explanation	No or independent modality outputs	Integrated cross-modal attributio n
Interaction Modeling	Not captured	Explicit second order modeling
Cross Modal Coverage	15%	68%
User Validation	None	45 participa nts including fact checkers
Verificati on Time	Baseline	28.3% faster

3. Methodology

3.1. Problem Formulation

Consider a multimodal input comprising an image *I* and a text sequence of *T*. The image is defined as a tensor of pixel values with dimensions $H \times W \times C$. The text is represented as a sequence of tokens $T = [t1, t2 \ldots, tn] T = [t1, t2 \ldots, tn]$, where each token corresponds to a word or sub word unit from a predefined vocabulary. A multimodal classifier $F(I, T)$ processes both inputs through vision and language encoders, fuses the learned representations, and outputs a prediction $y \hat{} \in [0,1]$ representing the probability that the given image-text pair contains misinformation. The goal of this work is to generate three distinct explanation outputs:

- *EI* = Where in the image matters
- *ET* = Which words matter
- *EIT*= How image and word together matter

3.2 Individual Modality Attribution

We begin by computing modality-specific attributions using Integrated Gradients, a method that satisfies the axioms of sensitivity and implementation of invariance. For visual modality, we define a baseline image I' as a black image of identical dimensions to *I*. The attribution for each pixel *p* is defined as the path integral of gradients along the straight line from the baseline to the input:

$$\text{attribution}^{\text{visual}}(p) = \Delta p \int_0^1 \text{gradient}(p' + \alpha \Delta p, p) \text{ d}\alpha$$

For the textual modality, we define a baseline token embedding T' corresponding to a masked or neutral token. The attribution for each token is computed similarly:

$$\text{attribution}^{\text{textual}}(t) = \Delta t \int_0^1 \text{gradient}(t, t' + \alpha \Delta t) \text{ d}\alpha$$

4. Results and Discussion

4.1 Quantitative Results

We compared CMA-X against four baseline methods: LIME-Independent, SHAP Independent, Grad-CAM, and Integrated Gradients-Independent. Across all evaluation metrics, CMA-X consistently outperformed every baseline. For faithfulness, which measures how accurate explanations reflect actual model behavior, CMA-X achieved a score of 0.79 compared to 0.67 from the best baseline, representing a 22.4% improvement. For explanation stability, CMA-X achieved 0.18 while the best baseline scored 0.24, a 31.7% reduction in instability. The most striking difference appeared in cross-modal coverage, where CMA-X scored 0.68 compared to only 0.15 from Integrated Gradients-Independent. This indicates that existing methods capture cross-modal interactions in only 15% of samples, while CMA-X captures them in 68% of samples.

4.2. Ablation study, user study result, case studies result

To understand the contribution of each component in our framework, we conducted an ablation study by removing individual elements and measuring performance. When using only individual attributions without any cross-modal interaction modelling, faithfulness was 0.67 and cross-modal coverage was 0.00. Adding visual interaction contributions improved faithfulness to 0.72 and cross-modal coverage to 0.31. Adding text interaction contributions improved faithfulness to 0.74 and cross-modal coverage to 0.34. When both interactions were included in the full CMA-X framework, faithfulness reached 0.79 and cross-modal coverage reached 0.68, confirming that both components contribute substantially and their combination produces a synergistic effect.

We validated our framework with 45 participants divided into two groups: 30 general users and 15 professional fact-checkers. For general users, CMA-X improved understanding from 3.2 to 4.4, trust from 3.4 to 4.2, and actionability from

2.9 to 4.1. For professional fact-checkers, improvements were even more pronounced with understanding rising from 3.5 to 4.7, trust from 3.8 to 4.5, and actionability from 3.3 to 4.6. Verification time decreased from 28.4 seconds to 18.7 seconds, a reduction of 28.3%.

4.3. Discussion

Our results demonstrate that capturing cross modal interactions is not merely an incremental improvement but a fundamental requirement for meaningful explanations in multimodal misinformation detection. The substantial gap in cross-modal coverage between CMA-X and baseline methods indicates that existing approaches miss most modality interactions that drive detection.

Analysis of the case studies reveals three scenarios where cross-modal explanations provide unique value. The first is temporal mismatches, where image age conflicts with temporal claims in text. The second is semantic contradictions, where visual content contradicts textual claims. The third is contextual manipulations, where authentic content is repurposed with deceptive framing.

5. Conclusion and Future Work

5.1 Conclusion

In this work, we presented CMA-X, a framework designed to explain why multimodal content gets flagged as misinformation. Current methods handle images and text separately, which means they cannot detect deceptions that arise from mismatches between what a picture shows and what a caption says. Our approach bridges this gap by explicitly measuring how visual elements and written words work together to influence detection decisions.

We tested our framework on three real-world datasets containing political, health, and general news content. The results show clear advantages over existing techniques. Our method improved faithfulness scores from 0.67 to 0.79, meaning explanations more accurately reflect actual model behavior. Instability dropped from 0.24 to 0.16, making explanations more consistent. Most notably, cross-modal coverage jumped from 0.15 to 0.68, indicating that our method captures interactions between modalities more than four times better than previous approaches. Taken together, these findings confirm that capturing how images and text interact is essential for explaining multimodal misinformation detection. CMA-X provides a theoretically grounded, empirically validated solution that moves the field beyond treating modalities as isolated information streams.

Metric	Baseline	CMA X (Ours)	Improvement
Faithfulness ↑	0.67	0.79	+22.4%
Explanation Stability ↓	0.24	0.16	-31.7%
Cross Modal Coverage ↑	0.15	0.68	+353%
User Trust (1-5) ↑	3.2	4.3	+1.1 pts
Verification Time (s) ↓	28.4	18.7	-28.3%

5.2 Future Work

Looking ahead, several paths offer opportunities for further advancement. One direction involves adapting our method for models that are not

easily differentiated, such as ensemble systems or commercial application interfaces. Another avenue explores causal explanation techniques that could identify not just correlations but genuine cause-effect relationships between content features and deception labels.

Computational efficiency remains an area for continued improvement. While our current implementation works well for offline fact checking workflows, reducing processing time further would enable deployment in browser extensions or real-time content moderation systems. Testing across multiple languages and cultural contexts would also help determine whether our findings hold universally or require adaptation for different settings. Finally, generating plain language summaries from our attribution outputs could make explanations more accessible to everyday users who lack technical training.

And tables should be properly numbered and cited.

References

[1] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2019.

[2] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *International Conference on Machine Learning*, 2017.

[3] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *IEEE International Conference on Computer Vision*, 2017.

[5] R. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning Transferable Visual Models from Natural Language Supervision," in *International Conference on Machine Learning*, 2021.

[7] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and Language Tasks," in *Advances in Neural Information Processing Systems*, 2019.

[15] A. Vaswani, N. Shazeer, N. Parmar, J.