AI Based Detection of Misleading Content by Comparing Images with Captions

Rahuf M[1], Dr. Bonthu Kotaiah[2]
[1]*Department of Computer Science, Central Tribal University of Andhra Pradesh, Kondkarakam, Vizianagaram[1], India*
[2]*Associate Professor and Head, Department of Computer Science, Central Tribal, University of Andhra Pradesh, Vizianagaram², India*

***Abstract*—With the prevalence of image sharing on social media, a growing concern is the possibility that an image and its caption may present misleading or even contradictory statements about what is actually shown in the image. The caption could misrepresent the people, location, timing, or events shown in the image. The usual approaches to detecting fake news from text are not sufficient for addressing this issue since they focus only on the caption text, not how it relates to the accompanying image. This paper proposes an AI-based approach to identifying potentially misleading image-caption pairs. The current prototype focuses on a vision-language model to analyze an image-caption pair, provide a risk score, identify any evidence mismatch, and generate an explanation. The prototype also includes a heuristic fallback mode, a web interface, and database storage of results. The analytical framework focuses on object, entity, event, temporal, and contextual consistency with the image content, and attempts to classify evidence as supported, contradicted, or unverifiable. The final classification output has three possible labels: consistent, misleading, and uncertain. Unlike previous work on the fake news detection task, this proposed framework is focused on detecting unsupported contextual claims and provides an explicit consideration of evidence mismatch and uncertainty. This prototype is not a fully formed fact-checking system, but rather a research tool that will be evaluated in practice on a future annotated image- caption dataset to formally establish its classification accuracy, explanation quality, robustness, and generalizability.**

***Index Terms*—Multimodal misinformation detection, image-caption consistency, vision-language models, contextual misinformation, artificial intelligence, explainable AI, cross-modal verification.**

I. INTRODUCTION

Digital platforms facilitate widespread information dissemination through news websites, messaging services, and social media. Images are frequently shared with captions describing an event, identifying a location, naming a person, or making a claim. However, the inclusion of images does not guarantee the accuracy of the accompanying information.

The most difficult aspect of detecting "out-of-context" misinformation is that the image itself may be accurate, but the caption may contain false information about the image's content such as the event or location taking place at a specific time or place. This means that an image and caption may seem semantically linked while simultaneously conveying misleading information.

Prior research on fake news detection from text has explored linguistic analysis, machine learning, deep learning, multimodal analysis, and vision-language models [1] – [8]. The recent work on contextual or out-of-context is information further discusses the distinction between superficial similarity and factual or contextual consistency [9] – [12].

This current project focuses on this narrower research problem: determining whether a caption is reasonably consistent with the image or whether the caption contains unsubstantiated claims. This project is not concerned with fully evaluating every caption's truthfulness but rather with providing an initial screening mechanism that identifies whether there is an apparent discrepancy between image and caption and explaining why.

The current prototype implements a website application with an integrated AI-based analysis

engine. A user submits an image-caption pair, the system performs a structured multimodal analysis, and the results are presented and stored for later review. A heuristic fallback is implemented for cases when the main analysis path is unavailable.

The main motivation for this project is to move beyond a simple similarity score to provide richer explanation of image-caption relationship.

II. RELATED WORK

2.1 Text-based fake news detection

The fake news detection literature initially focused on linguistic analysis of text, machine learning models, and deep learning approaches [2] – [3]. Subsequent work has discussed applying multimodal analysis and vision-language models to the task of fake news detection from text [1] [4] – [8].

These approaches are helpful in establishing baseline text analysis for a caption. However, a text-only classifier is not sufficient since it cannot determine whether the caption actually describes what is shown in the image.

2.2 Multimodal fake news detection

Fake news detection research has also explored the use of multimodal analysis to incorporate information from different sources, primarily text and images [5] – [8]. This work has also shown the importance of utilizing cross-modal representations to detect inconsistencies [6]. These approaches are also useful in analyzing the image-caption pair. However, the mere presence of an image and caption is not sufficient to ensure that they describe the same information.

The overall similarity between modalities may be high even when there is an evident contradiction in captions or when the image does not support specific claims.

2.3 Image contextualization and out-of-context misinformation

The recent work on out-of-context misinformation highlights the importance of analyzing images in context [9] – [12]. Tonglet et al. describe image contextualization as a task where an algorithm attempts to retrieve information about an image's original context [9]. Specifically, their work emphasizes the importance of retrieving what an image actually shows rather than focusing only on caption similarity. Similarly, the COVE approach of Tonglet, Thiem, and Gurevych focuses on caption veracity as a specific task [10]. Papadopoulos et al. further emphasize the importance of distinguishing between superficial similarity and factual reasoning in out-of-context misinformation detection approaches [11]. An explicit analysis of the distinction between similarity and factual consistency is also described in the recent work by Papadopoulos et al. on evaluating approaches to out-of-context misinformation detection [12]. This line of research is directly relevant to the current project since it focuses on how to determine whether the image supports the caption's specific claims.

2.4 Vision-language models and robustness

A number of vision-language models can reason about images and text. However, a vision-language model's output is not always reliable evidence about the real world. Several recent works have focused on evaluating vision-language models for disinformation detection tasks [13]. MVI-Bench research is also relevant to the current project since it describes robustness to various types of visually deceptive prompting [14]. This means that a vision-language model may generate confident but incorrect captions for images.

A system designed to evaluate image-caption pairs needs to account for this possibility, particularly when it comes to explaining mismatches.

2.5 Research position

Based on the prior work, this current research project is focused on the following principles:

1. Analyzing image and text together can be necessary to identify misleading claims.
2. Semantic compatibility is not the same as factual or contextual consistency.
3. Evidence and explanations are important for responsible screening.

These principles guide the current research and development work on a web-based image-caption screening application. The current project aims to implement a prototype that structures the evidence for image-caption consistency, identifies possible mismatch, provides a risk score, and explains its findings. Whether this approach is more effective than simpler methods will be determined by the subsequent research.

III. LITERATURE REVIEW

3.1 Evolution of Fake News Detection (2016–2019)

Research on misinformation detection initially concentrated on analyzing textual information, including linguistic patterns, writing styles, and the credibility of online sources. Machine learning and deep learning techniques were increasingly applied to identify misleading news articles and social media posts. Shu et al. [3] reviewed fake news detection from a data-mining perspective and discussed the importance of analyzing information beyond individual textual features.

Although text-based approaches provide useful information about the language used in misleading content, they have an important limitation: they cannot independently determine whether an accompanying image supports the claims made in a caption. For example, a caption may describe a genuine event but use an unrelated photograph to create a misleading impression. This limitation demonstrates the need for methods that analyze both textual and visual information.

3.2 Development of Multimodal Fake News Detection (2020–2021)

Between 2020 and 2021, researchers increasingly explored multimodal approaches that combine textual and visual features. Zhang and Ghorbani

[2] reviewed online fake news characterization and detection methods, highlighting the complexity of identifying misinformation in digital environments. Xue et al. [4] investigated multimodal data consistency as a means of detecting fake news.

Radford et al. [21] introduced Contrastive Language–Image Pre-training (CLIP), which learns relationships between images and natural-language descriptions using large-scale image-text pairs. This research provided an important foundation for comparing visual content with captions. However, image-text similarity alone does not establish whether a caption's specific claims are factually correct.

Luo et al. [22] introduced NewsCLIPpings, a dataset designed to study out-of-context multimodal misinformation. The dataset includes examples in which genuine images are paired with mismatched text. Their work demonstrated that misleading content can be created without manipulating the original image or caption. This is particularly relevant to the present research because it focuses on whether the image and its accompanying caption convey the same information.

3.3 Object-Level and Contextual Mismatch Detection (2022–2023)

Research during 2022 and 2023 increasingly emphasized the importance of identifying inconsistencies between visual objects and textual claims. Segura-Bedmar and Alonso-Bartolome [6] discussed multimodal fake news detection, contributing to the broader understanding of approaches that combine different information modalities.

Aneja et al. [23] proposed COSMOS, a self-supervised approach for detecting out-of-context image misuse. The method examines relationships between image objects and textual claims to identify potentially misleading image-text pairs. Their work is particularly relevant because it addresses situations in which an image is authentic but its surrounding narrative is misleading.

Li et al. [8] also investigated cross-modal consistency learning through a fine-grained fusion network for multimodal fake news detection. Such approaches highlight the importance of analyzing interactions between modalities rather than treating images and text as independent sources of information.

Nevertheless, identifying a semantic relationship between an image and a caption does not necessarily establish that every claim in the caption is supported. A photograph of a public gathering, for example, may be visually compatible with a caption about a political event without proving the stated date, location, or identity of the participants.

3.4 Image Contextualization and Vision-Language Models (2024–2025)

Recent research has placed greater emphasis on contextual verification and explainable misinformation detection. Tonglet et al. [9] investigated the prediction of an image's original context, including information that can help fact-checkers understand where an image originated and what it depicts. Their research demonstrates that reconstructing an image's original context can provide useful evidence for evaluating accompanying claims.

Papadopoulos et al. [11], [12] examined the distinction between similarity and factuality in multimodal out-

of-context misinformation detection. Their work highlights an important limitation of similarity-based methods: an image and caption may appear semantically related even when the caption introduces unsupported contextual information.

Tonglet et al. [10] subsequently introduced COVE, a framework that first predicts the original context of an image and then evaluates the veracity of its caption. The approach illustrates the potential benefits of separating image contextualization from caption verification. It also demonstrates the value of presenting contextual information that can assist human fact-checkers.

These studies directly support the motivation for the present research. Rather than relying exclusively on a general image-caption similarity score, a detection system should examine specific claims and distinguish between supported information, contradictions, and claims that cannot be verified using the image alone.

3.5 Recent Developments and Remaining Challenges (2025–2026)

Recent research on vision-language models has expanded the possibilities for analyzing images and natural-language descriptions within a single framework. These models can describe image content, identify objects, and reason about relationships between visual and textual information. Research on robustness to misleading visual inputs also highlights the need to evaluate how reliably these models interpret deceptive or ambiguous examples.

However, several challenges remain. A model may produce an incorrect explanation, overlook a subtle contradiction, or confidently infer details that are not visible in an image. Furthermore, an image alone generally cannot establish historical dates, precise locations, or the identities of people without additional evidence.

The present research addresses these limitations by proposing a structured image-caption analysis framework. It examines object, entity, event, temporal, location, and contextual claims and classifies the relationship between an image and its caption as consistent, misleading, or uncertain. It also aims to provide a risk score and a human-readable explanation.

Unlike approaches that focus primarily on classification or semantic similarity, the proposed framework emphasizes evidence assessment, explicit uncertainty, and explanation quality. Its effectiveness must be established through experiments using an annotated dataset, baseline comparisons, and suitable evaluation metrics.

3.6 Summary of the Literature Review

The literature from 2016 to 2026 illustrates a progression from text-based misinformation detection to multimodal analysis, object-level consistency checking, image contextualization, and vision-language reasoning.

Existing studies provide important methods and datasets for identifying misleading image-text relationships. However, semantic similarity cannot independently establish factual consistency, and visual evidence may be insufficient to verify contextual claims.

Therefore, the present study proposes an AI-assisted screening framework that identifies possible mismatches, distinguishes contradictions from unverifiable claims, and provides explanations for human review. Its contribution lies in the structured implementation and proposed evaluation of this workflow rather than an assumption that the system already achieves superior detection accuracy.

IV. LIMITATIONS OF EXISTING RESEARCH AND RESEARCH GAP

The taxonomy paper by Al-Kaabi et al. provides a comprehensive overview of approaches to fake news detection from text, image, video, and multimodal content [1]. It also discusses various aspects of robustness, cross-modal verification, and human-AI fact-checking [1]. This paper serves as a helpful for the current research. Based on this taxonomy and the prior work, it is possible to identify several research gaps that this project attempts to address.

First, general fake news classification is different from structured image-caption verification. A fake news classifier may not identify what specific claims in a caption are unsubstantiated.

Second, semantic compatibility is not the same as factual or contextual consistency. An image may have some obvious objects or activities that are similar to the caption but fail to establish more specific contextual details such as the location or time [9] – [12]. A third limitation is the lack of explicit representation of uncertainty. There may be claims

that are not supported by an image but cannot be classified as clearly contradicted either. A screening tool needs to identify unsupported claims while distinguishing between readily apparent and non-obvious contradictions. Finally, explanations are often assumed to be correct even when they are not based on real evidence. If an AI model provides fluent but demonstrably wrong explanations, they must still be identified as incorrect. This means that the value of explanations also needs to be evaluated separately from classification accuracy.

Finally, a working prototype is not sufficient to formally establish scientific validity. A system must be evaluated on a labeled dataset with clearly defined methodologies to identify strengths and weaknesses.

4.1 Research gap

This project focuses on the practical application of an image-caption screening workflow. This workflow separates evidence from unsupported claims, identifies specific mismatch categories, provides an explicit uncertain category, and produces a human-readable explanation. The contribution of this research is the implementation of a working prototype and the structured analysis and evaluation methodology rather than immediate classification performance improvements.

V. PROBLEM STATEMENT

Given the image I and the caption C, determine whether it is consistent or misleading and what specific evidence if any supports that conclusion. The task can be represented with the following formula:

The three possible classification labels are CONSISTENT, MISLEADING, and UNCERTAIN.

Note that simply assessing the similarity of the captions and images is not sufficient to address the problem. It is necessary to evaluate whether the image supports each of the claims in the caption and avoid unsubstantiated assertions

VI. RESEARCH OBJECTIVES

This research project aims to:

- Develop a working image analysis web application;
- Apply a vision-language model to analyze image-caption pairs;
- Identify object, entity, event, time-related, and contextual inconsistencies;
- Separate supported, contradicted, and unsubstantiated claims;
- Provide an interpretable classification and explanation;
- Store analysis records for subsequent analysis;
- Develop a formal evaluation plan to use human annotations;
- Identify limitations and propose subsequent improvements.

VII. RESEARCH QUESTIONS

This research is guided by the following questions:

- RQ1: How effective are vision-language models at identifying inconsistencies between image-caption pairs?
- RQ2: Can a mismatch-aware framework identify contextual misinformation better than generic similarity scores?
- RQ3: How often does the system correctly assign an uncertain label when the image fails to support a caption claim?
- RQ4: Do structured mismatch categories improve the utility of explanations?
- RQ5: How does the proposed approach compare to text-only, image-only, and similarity-based methods?

These research questions define the direction of this study. The questions are not answered immediately since this project focuses on building a working prototype rather than formal evaluation.

VIII. PROPOSED MODEL AND METHODOLOGY

8.1 Overall approach

The current prototype takes an image-caption pair as input and performs a structured analysis with a vision-language model. The system then provides evidence for claims, identifies potential inconsistencies, classifies the caption as misleading, and provides an explanation. The research component of this project focuses on organizing this information into a structured whole rather than attempting to optimize each component.

The analysis approach can be represented with the following sequence: input analysis → multimodal

analysis → evidence assessment → mismatch identification → classification → risk score and explanation.

8.2 Input processing

The user provides an image and a caption as input to the system. The backend processes this request. The image provides visual evidence, and the caption contains claims that need to be assessed for consistency with the image.

The input processing step may need to include validation of user input and the ability to reject unsupported or incorrectly formatted requests. The specific requirements depend on the image formats and model interface that the application uses.

8.3 Vision-language analysis

The current prototype uses a Gemini-based vision-language model for multimodal analysis of an image-caption pair. The backend requests specific information about the input from the model. The model provides a classification, risk score, objects detected, scene description, various consistencies, mismatch categories, detected issues, entities identified, temporal claims, and an explanation.

The model implementation combines its understanding of the image with the caption analysis. Since it is a generative model, these findings should be treated as hypotheses rather than proven facts.

The model's configuration should be explicitly noted with the model identifier used in the current implementation. The model's name and date of experimentation may be relevant if the model becomes unavailable or renamed.

8.4 Caption claim analysis

The caption may contain multiple types of claims. For example, a caption may state what objects are present in the image, identify people shown in the image, describe an event taking place, indicate the time when the event occurred, and explain the reasons for the event.

For the purposes of this research, the following categories can be used to structure the claims contained in a caption:

Object claims: statements about objects shown in the image; Entity claims: identifying statements about people or organizations shown in the image; Event claims: statements describing an on-going or completed event; Temporal claims: statements describing the time when the event took place; Location claims: statements describing the location where the event took place; Contextual claims: additional information not explicitly shown in the image; Narrative claims: more broad claims that go beyond direct evidence provided by the image.

The current prototype conducts caption analysis but does not include an independent claim-extraction component.

8.5 Cross-modal consistency assessment

Assessment of an image-caption pair considers the relationship between claims in the caption and the evidence provided by the image. This analysis focuses on five main areas:

Semantic consistency: does the caption describe the image subject matter accurately? Object consistency: does the image show all the objects described in the caption? Entity consistency: does the image provide sufficient evidence to identify the entities mentioned in the caption? Event consistency: does the image provide sufficient evidence to support the events described in the caption? Temporal and contextual consistency: does the image provide sufficient evidence to support the time-related and other contextual claims made in the caption?

These categories serve as analytical tools rather than immediately providing classification evidence. Each category's contribution to the classification must be established through experimentation.

A crucial distinction must be made between claims that are explicitly supported or contradicted by the image and claims that are merely unsubstantiated. An image typically provides limited evidence about the identity of a person, the historical context of an event, the exact location of an event, or other similar claims.

8.6 Evidence status

Each claim can be assigned one of the following three evidence statuses:

Supported: the available image evidence is compatible with the claim;

Contradicted: the available image evidence contradicts the claim;

Not Verifiable: the image evidence does not provide sufficient information to support or refute the claim.

For example, if a caption states that a certain visible object is red, this claim would be supported if the

image shows this color for the object. A caption that states that the image was taken in a specific city in 2018 would need additional metadata or reliable information to be supported.

The three categories help ensure that unverified claims are not dismissed outright. Note that in the current prototype, each category should be considered a model's hypothesis rather than an established fact.

8.7 Final classification

The classification uses three labels:

Consistent: the caption is compatible with the image evidence;

Misleading: there is sufficient evidence to identify unsubstantiated claims;

Uncertain: there is insufficient evidence to make a classification.

The classification label should be based on a combination of evidence rather than rely on general similarity between captions and images.

As mentioned previously, unsubstantiated claims are not always false, but they should be identified as such.

8.8 Model Risk Score

The prototype uses a Model Risk Score, which is a numeric value from 0 to 100 that represents the model's confidence in its misleading classification. A higher score indicates that the model identified more evidence of unsubstantiated claims.

A higher score is not necessarily an indicator of higher precision since the score reflects only the model's confidence in its own classification rather than formally established likelihood of being correct. The risk score's utility must be explored experimentally.

8.9 Explanation generation

The system generates an explanation in natural language that provides relevant details about the image and caption, identifies potential issues, and explains why the caption may be misleading or uncertain. The explanation is intended to help a human analyst make a final judgment about the caption. For research purposes, such explanations can be evaluated to identify strengths and weaknesses. The most important factors are: agreement with the actual image, relevance to the caption claims, specificity of issues identified, acknowledgment of uncertainty, and the absence of invented image details. A fluent explanation is not necessarily correct. Explanation quality must be evaluated separately.

IX. SYSTEM ARCHITECTURE AND IMPLEMENTATION

9.1 Frontend

The frontend is implemented in React and Vite. This component provides the image analysis workflow for the end-user. The frontend displays the model's classification and other relevant information. It also implements the project's analysis-oriented features.

9.2 Backend

The backend is implemented in Python and FastAPI with additional libraries for request validation and processing. This component is responsible for request processing, model interaction, response formatting, and API management.

The API layer allows the frontend to communicate with the analysis engine. This architecture enables easier model replacement if a different vision-language model is used.

9.3 Database

The database is implemented with SQLite. This component stores analysis records to support subsequent research and experimentation.

The records in the database are useful for research purposes, but a formal evaluation requires a separate labeled dataset that follows specific labeling guidelines and is separated into training, validation, and testing subsets.

9.4 AI analysis engine

The current prototype utilizes a Gemini-based vision-language model to analyze the input image-caption pair. The model provides structured output regarding caption classification, risk score, objects detected, scene description, various consistencies, identified mismatch, detected issues, entities found, temporal claims, and an explanation.

The prototype also contains a heuristic fallback component that implements basic image-caption analysis. This component is useful if the main vision-language model becomes unavailable. The fallback component should not be used as a substitute for a full multimodal analysis.

Some additional model components such as CLIP-style visual-text representation, sentence embeddings, and object detection, might be considered for future development. These should not be characterized as

active contributors to the current final classification until their integration into the tested inference path is established.

9.5 Implementation summary

The current project provides an integrated prototype, with a frontend, backend, model interaction, structured results, fallback analysis, and database-backed records, that demonstrate the feasibility of the application workflow.

It does not, in and of itself, establish the ability of the model to reliably detect all types of misleading contents. This claim requires the experimental evaluation described in the following sections.

X. EXPERIMENTAL METHODOLOGY

10.1 Dataset construction

The controlled dataset of image-caption pairs should be developed or obtained from an appropriately licensed source. Each example should include the image, caption, reference label, mismatch category wherever applicable, evidence status, source information, and annotation note.

The dataset should include the following categories:

- Consistent image-caption pairs. Object mismatches.
- Entity mismatches. Event mismatches. Temporal mismatches.
- Location or contextual mismatches. Unsupported narrative claims.
- Cases that cannot be determined from the image alone.

The dataset should include both clearly contradictory examples and subtle examples wherein the image appears relevant but does not establish the full caption claim.

10.2 Human annotation

At least two annotators should independently label a suitable subset of the dataset using documented annotation guidelines. Disagreements should be reviewed and resolved through a defined adjudication process.

Inter-annotator agreement should be reported where feasible. The annotation guidelines should explain the difference between a contradiction, an unsupported claim, and an unverifiable claim.

10.3 Data separation

The data should be divided into training, validation, and test partitions if model training is performed. If the system is evaluated without training, a fixed, held-out test set should be maintained.

Images or near-duplicate examples should not inadvertently appear across partitions, because this can lead to overly optimistic estimates of performance.

10.4 Baseline comparison

The proposed system should be compared with the following baselines:

- Text-only baseline: Analyses the caption without using the image.
- Image-only baseline: Analyses visual content without assessing caption consistency.
- Similarity baseline: Uses an image-caption similarity method to estimate semantic alignment.
- Current vision-language model: Uses the prototype's main model-based analysis.
- Mismatch-aware evaluation: Examines whether the structured evidence and mismatch categories improve performance over the simpler approaches.

Where feasible, all approaches should be evaluated on the same test examples and under clearly documented settings.

10.5 Evaluation metrics

The following metrics are recommended for the three-class classification task:

Accuracy: Precision and recall for each class. Macro-F1 score.

Confusion matrix: Performance by mismatch category. Rate of uncertain predictions.

Explanation-grounding quality, assessed using a human review protocol.

Macro-F1 is particularly useful when the dataset contains unequal numbers of consistent, misleading, and uncertain examples.

If the risk score is evaluated as a ranking measure, suitable threshold-based and ranking metrics should be selected according to the task and reported with their definitions. Calibration metrics are needed if the score is intended to represent probability.

10.6 Ablation study

An ablation study can examine whether different evidence dimensions contribute useful information.

Possible variants include:
General image-caption assessment. Assessment with object consistency. Assessment with entity consistency. Assessment with event consistency.
Assessment with temporal and contextual consistency. Full mismatch-aware assessment.
The experiment should compare these variants under a controlled protocol. If a component is not independently implemented, it must be reported as a proposed experiment rather than a completed result.

XI. PRELIMINARY PROTOTYPE VALIDATION AND RESULTS REPORTING

The current prototype contains stored development analyses and provides an operational workflow for submitting image-caption pairs and receiving model-generated results.
These records demonstrate that the application can retain and display outputs from its analysis process. They should not be treated as a controlled benchmark, and they do not establish classification accuracy, precision, recall, or generalization.
No quantitative performance claims are made in this paper because a controlled, independently annotated test set and corresponding evaluation results have not yet been established.
Once the evaluation is conducted, this section should report:
Dataset size and category distribution. Annotation procedure and agreement.
Model and configuration used. Evaluation protocol.
Accuracy, macro-F1, class-wise precision and recall.
Confusion matrix.
Error analysis and examples of incorrect explanations.
Comparison with the selected baselines.
Results should be inserted only after the experiments have actually been run. This distinction is necessary to preserve the scientific validity of the paper.

XII. EXPECTED CONTRIBUTIONS

The project is expected to contribute in three main areas.
Practical contribution: A web-based prototype for examining image-caption pairs and presenting structured model assessments.
Methodological contribution: A mismatch-oriented analysis framework that separates broad semantic alignment from object, entity, event, temporal, and contextual claims.
Evaluation contribution: A proposed controlled testing protocol for assessing classification quality, uncertainty handling, explanation grounding, and performance across different mismatch categories.
These are intended contributions. Their effectiveness must be established through implementation verification and empirical evaluation.

XIII. LIMITATIONS

The current system has several limitations.
First, the primary analysis depends on a generative vision-language model. Its output may contain errors, and the model may misidentify objects, misunderstand a caption, or produce explanations that are not fully grounded in the image.
Second, visual evidence alone is insufficient to verify every contextual claim. Historical events, exact dates, geographic locations, and personal identities may require trusted external evidence.
Third, the heuristic fallback is limited by its predefined visual cues. It cannot be assumed to provide the same depth of analysis as the main model.
Fourth, the risk score is not a calibrated probability. Its interpretation and usefulness need to be established empirically.
Fifth, the current development records are not a substitute for a balanced, independently annotated benchmark.
Finally, the present implementation does not establish the independent effectiveness of CLIP, YOLO, sentence embeddings, or other components that may be included in future extensions.
These limitations mean that the system should be treated as an AI-assisted screening tool rather than an autonomous fact-checking authority.

XIV. ETHICAL CONSIDERATIONS

An incorrect misleading-content label could affect the reputation of a person, organization, or news source. For this reason, system outputs should be presented as assessments requiring review, rather than definitive judgments.
The system should distinguish visual contradictions from claims that cannot be verified using the image alone. It should communicate uncertainty, avoid

unsupported allegations, and allow human reviewers to inspect the underlying evidence.

Images and captions may also contain sensitive personal information. Dataset collection, storage, and sharing should follow applicable privacy requirements and the terms of the relevant data sources.

XV. FUTURE WORK

Future development will focus on the following areas:

- Building a larger, balanced, human-annotated image-caption dataset.
- Evaluating the prototype against text-only, image-only, and similarity-based baselines.
- Integrating and independently evaluating CLIP-style representations, object detection, OCR, and sentence-level claim analysis.
- Adding trusted external evidence retrieval for claims involving time, location, and event context.
- Evaluating whether the risk score can be calibrated against human judgments.
- Testing robustness to paraphrased captions, subtle contradictions, and adversarial wording.
- Measuring explanation quality and uncertainty handling through human assessment.
- Improving reproducibility through documented model versions, configuration settings, datasets, and experiment logs.

These extensions could help determine whether a structured mismatch-aware approach offers measurable benefits over general multimodal similarity and model-generated judgments alone.

XVI. CONCLUSION

This research addresses the problem of potentially misleading captions associated with images, including cases where authentic visual content is reused to support an unsupported narrative. The project implements a web-based prototype using a Gemini-based vision-language analysis path, a heuristic fallback, a FastAPI backend, a React frontend, and persistent storage of analysis records.

The proposed research framework emphasizes multiple forms of image-caption inconsistency, explicit evidence statuses, three-way classification, risk scoring, and human-readable explanations. Its main distinction is the focus on whether the image supports the claims made by the caption, rather than whether the two inputs merely appear similar.

The current implementation demonstrates a practical workflow but does not yet establish scientific performance. A controlled dataset, human annotation, baseline comparisons, and quantitative evaluation are necessary to determine its reliability and limitations. With this evaluation, the project can develop into a more rigorous approach to explainable image-caption misinformation screening.

REFERENCES

[1]. H. A. Al-Kaabi, A. N. Kamber, and M. R. Al-Rikab, “Fake News Detection: A Comprehensive Taxonomy of Text, Image, Video, and Multi-Modal Techniques,” AlKadhim Journal for Computer Science, vol. 3, no. 2, pp. 30–45, 2025, doi: 10.61710/kjcs. v3i2.103.

[2]. X. Zhang and A. A. Ghorbani, “An Overview of Online Fake News: Characterization, Detection, and Discussion,” Information Processing & Management, vol. 57, no. 2, Art. no. 102025, 2020.

[3]. K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” ACM Computing Surveys, vol. 53, no. 5, pp. 1–38, 2020.

[4]. J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, “Detecting Fake News by Exploring the Consistency of Multimodal Data,” Information Processing & Management, vol. 58, no. 5, Art. no. 102610, 2021.

[5]. P. Meel and D. K. Vishwakarma, “HAN, Image Captioning, and Forensics Ensemble Multimodal Fake News Detection,” Information Sciences, vol. 567, pp. 23–41, 2021.

[6]. S. Segura-Bedmar and S. Alonso-Bartolome, “Multimodal Fake News Detection,” Information, vol. 13, no. 6, Art. no. 284, 2022.

[7]. “Multi-Modal Fake News Detection via Bridging the Gap between Modals,” Entropy, vol. 25, no. 4, Art. no. 614, 2023.

[8]. J. Li, Y. Bin, J. Zou, G. Wang, and Y. Yang, “Cross-Modal Consistency Learning with Fine-

Grained Fusion Network for Multimodal Fake News Detection," arXiv preprint arXiv:2311.01807, 2023. https://arxiv.org/abs/2311.01807

[9]. J. Tonglet, M.-F. Moens, and I. Gurevych, "'Image, Tell Me Your Story!' Predicting the Original Meta-Context of Visual Misinformation," in Proc. 2024 Conf. Empirical Methods in Natural Language Processing, 2024, pp. 7845–7864, doi: 10.18653/v1/2024.emnlp-main.448.

[10]. J. Tonglet, G. Thiem, and I. Gurevych, "COVE: COntext and VEracity Prediction for Out-of-Context Images," in Proc. 2025 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2025. https://aclanthology.org/2025.naacl-long.102/

[11]. S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, "Similarity over Factuality: Are We Making Progress on Multimodal Out-of-Context Misinformation Detection?" arXiv preprint arXiv:2407.13488, 2024. https://arxiv.org/abs/2407.13488

[12]. S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, "Similarity over Factuality: Are We Making Progress on Multimodal Out-of-Context Misinformation Detection?" in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision, 2025, pp. 5570–5579.

[13]. S. Raza et al., "VLDBench: Vision Language Models Disinformation Detection Benchmark," arXiv preprint, 2025. https://huggingface.co/papers/2502.1136

[14]. H. Chen et al., "MVI-Bench: A Comprehensive Benchmark for Evaluating Robustness to Misleading Visual Inputs in LVLMs," in Proc. 43rd Int. Conf. Machine Learning, Proc. Mach. Learn. Res., vol. 306, pp. 16324–16341, 2026. https://proceedings.mlr.press/v306/chen26dh.html

[15]. J. Wu, F. Li, Z. Fu, M.-Y. Kan, and B. Hooi, "Seeing Through Deception: Uncovering Misleading Creator Intent in Multimodal News with Vision-Language Models," in Proc. Int. Conf. Learning Representations, 2026. Verify final publication metadata before submission.

[16]. "A Fuzzy-Based Multimodal Approach for Interpretable Fake News Detection," Applied Soft Computing, vol. 179, Art. no. 113277, 2025.

[17]. Z. Zhou, H. Guan, M. M. Bhat, and J. Hsu, "Fake News Detection via NLP Is Vulnerable to Adversarial Attacks," arXiv preprint arXiv:1901.09657, 2019. https://arxiv.org/abs/1901.09657

[18]. R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach," Multimedia Tools and Applications, vol. 80, no. 8, pp. 11765–11788, 2021.

[19]. C. Zhang et al., "Fake News Detection: Taxonomy and Comparative Study," Information Fusion, 2024. Verify final bibliographic details and publication year against the publisher's record before submission.

[20]. M. Westerlund, "The Emergence of Deepfake Technology: A Review," Technology Innovation Management Review, vol. 9, no. 11, 2019.

[21]. A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," in Proc. 38th Int. Conf. Machine Learning, vol. 139, pp. 8748–8763, 2021.

[22]. G. Luo, T. Darrell, and A. Rohrbach, "NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media," in Proc. 2021 Conf. Empirical Methods in Natural Language Processing, pp. 6801–6817, 2021, doi: 10.18653/v1/2021.emnlp-main.545.

[23]. S. Aneja, C. Bregler, and M. Nießner, "COSMOS: Catching Out-of-Context Image Misuse Using Self-Supervised Learning," in Proc. AAAI Conf. Artificial Intelligence, vol. 37, no. 12, pp. 14084–14092, 2023, doi: 10.1609/aaai. v37i12.26648.