MIRAGE: Agentic Framework for Multimodal Misinformation Detection with Web-Grounded Reasoning

Mir Nafis Sharear Shopnil
Fatima Fellowship
sharears4077@gmail

Sharad Duwal
Fatima Fellowship

Abhishek Tyagi
University of Rochester
Rochester, NY, USA

Adiba Mahbub Proma
University of Rochester
Rochester, NY, USA

Abstract

Misinformation spreads across web platforms through billions of daily multimodal posts that combine text and images, overwhelming manual fact-checking capacity. Supervised detection models require domain-specific training data and fail to generalize across diverse manipulation tactics. We present MIRAGE, an inference-time, model-pluggable agentic framework that decomposes multimodal verification into four sequential modules: visual veracity assessment detects AI-generated images, cross-modal consistency analysis identifies out-of-context repurposing, retrieval-augmented factual checking grounds claims in web evidence through iterative question generation, and a calibrated judgment module integrates all signals. MIRAGE orchestrates vision-language model reasoning with targeted web retrieval, outputs structured and citation-linked rationales. On MMFakeBench validation set (1,000 samples), MIRAGE with GPT-4o-mini achieves 81.65% F1 and 75.1% accuracy, outperforming the strongest zero-shot baseline (GPT-4V with MMD-Agent at 74.0% F1) by 7.65 points while maintaining 34.3% false positive rate versus 97.3% for a judge-only baseline. Test set results (5,000 samples) confirm generalization with 81.44% F1 and 75.08% accuracy. Ablation studies show visual verification contributes 5.18 F1 points and retrieval-augmented reasoning contributes 2.97 points. Our results demonstrate that decomposed agentic reasoning with web retrieval can match supervised detector performance without domain-specific training, enabling misinformation detection across modalities where labeled data remains scarce.

Keywords misinformation detection · multimodal learning · vision-language models · agentic systems · fact-checking

1 Introduction

Web platforms process billions of multimodal posts daily, where text and images combine to spread information at unprecedented scale, fact-checking organizations verify thousands of claims annually [1], yet web-based misinformation—encompassing millions of misleading posts—spreads far faster than manual verification can address. Multimodal misinformation, defined as false narratives that exploit coordinated manipulation across textual and visual modalities [2], thrives in this verification gap. Posts on web platforms pair false claims with synthetic images, repurpose authentic images with misleading captions, or use genuine content but change contexts [2, 3]. The stakes are high. Medical misinformation can influence health decisions [4]. Financial fraud costs billions annually [5]. Election interference undermines democratic processes [5]. Therefore, it is necessary to have scalable web-grounded detection approaches that can match the speed and volume of misinformation that propagates across the modern web.

Despite rapid progress in multimodal misinformation detection systems, existing approaches still have some limitations in generalization, robustness, and factual grounding. Supervised deep learning models achieve strong performance on benchmark datasets but require domain-specific training data and fail to generalize across manipulation types [6]. Recent work shows that end-to-end trained detectors are susceptible to semantic shifts and cross-modal inconsistencies [7]. Prompted large vision-language models offer flexibility without training but suffer from critical weaknesses. [8] Comparative evaluations reveal moderate accuracy on fact-checking tasks, with GPT-4 achieving 61-72% on mixed-source misinformation benchmarks [9, 2]. Canonical RAG methods establish that retrieval mitigates parametric
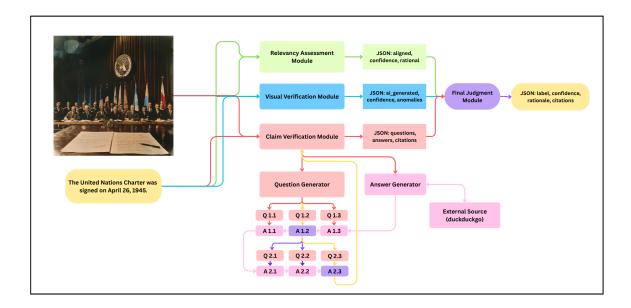
Figure 1: MIRAGE architecture showing the four sequential modules: Visual Verification analyzes images for AI generation and manipulation, Relevancy Assessment evaluates cross-modal alignment, Claim Verification performs retrieval-augmented fact-checking, and Final Judgment integrates all signals to produce calibrated predictions.

knowledge limits and improves provenance [10]. LLMs hallucinate explanations when knowledge is lacking [11]. Unchecked LLM credibility judgments show political bias and diverge from journalist baselines [12]. Rumor detection on social threads exposes limitations of reasoning over conversational structure [13]. To our knowledge, no prior system combines cross-modal verification, retrieval-augmented reasoning, and calibrated confidence scoring in a model-pluggable architecture.

Here, we present MIRAGE (Multi-Modal Agentic Reasoning for Misinformation Detection), an inference-time framework that decomposes multimodal misinformation detection into four explicit verification stages. The visual verification module examines images for AI-generation artifacts and manipulation signatures, addressing the documented rise of synthetic media [5]. The module checks for statistical anomalies, inconsistent lighting, and impossible visual elements. The relevancy assessment module evaluates cross-modal consistency, detecting out-of-context repurposing where real images accompany misleading text. The module produces three-level alignment judgments: true alignment where the image depicts the claim, partial alignment where context matches but specifics differ, and false alignment indicating mismatch or contradiction. The claim verification module generates investigative questions about factual claims through three sequential reasoning chains, queries DuckDuckGo web search to retrieve web evidence, and synthesizes web information with citation-linked answers. Each chain builds on findings from previous questions, enabling deeper investigation than single-pass retrieval. The final judgment module integrates signals from all components using structured decision rules. The rules prevent exploiting class imbalance by requiring factual accuracy, reasonable alignment, and genuine imagery for "Not Misinformation" predictions.

MIRAGE is model-pluggable: Any vision-language model capable of following structured prompts can serve as the reasoning engine [14, 15]. The framework outputs structured JSON containing labels, confidence scores, rationales synthesizing evidence from all modules, and citation links to web sources. Confidence scores from individual modules feed into the judge's aggregation logic. Rationales explain which signals drove decisions. This design enables transparent auditing and supports content moderation workflows requiring human review. The system addresses documented gaps in evidence-based reasoning by explicitly retrieving evidence rather than relying on parametric knowledge [16].

Our contributions are two-fold. First, we demonstrate that decomposed agentic reasoning with web retrieval achieves 81.65% F1 on MMFakeBench validation set (1,000 samples), outperforming the strongest zero-shot baseline (GPT-4V with MMD-Agent prompting at 74.0% F1) by 7.65 points. The system maintains 34.3% false positive rate on authentic

content, compared to 97.3% for a judge-only baseline that uses class distribution. Second, we conduct error analysis of 249 misclassifications, which reveals specific failure modes: high-fidelity AI images passing visual checks, niche topics with limited web coverage producing low-confidence answers, and generic stock photos in legitimate journalism flagged as weakly aligned. These findings provide directions for future improvement in multimodal misinformation detection.

2 Related Work

2.1 Multimodal Misinformation Benchmarks

Early fake news datasets focused on text-only claims or isolated image manipulations. MMFakeBench introduced mixed-source evaluation with 11,000 image-text pairs spanning three distortion categories which are textual veracity distortion, visual veracity distortion, and cross-modal consistency distortion [2]. Textual veracity distortion includes false claims paired with real or AI-generated images. Visual veracity distortion involves authentic text with manipulated or synthetic images. Cross-modal consistency distortion pairs real text with real images from wrong contexts. The benchmark reveals that models achieving high accuracy on single manipulation types struggle when the tactics are combined.Human annotators achieve only 56.8% accuracy on binary classification and 37.9% on multi-class source identification [19], indicating inherent difficulty and establishing an upper bound on task difficulty. LVLM baselines reach 72-74% F1, leaving substantial room for improvement. MIRAGE evaluation on MMFakeBench tests all three distortion types in a mixed-source setting.

SNIFFER targets out-of-context detection with entity-grounded explanations [17], while MDAM3 extends coverage to audio and video alongside text and images [3]. These benchmarks establish that real-world misinformation requires cross-modal reasoning rather than isolated text or image analysis.

2.2 Vision-Language Models for Fact-Checking

Instruction-tuned vision-language models enable multimodal reasoning over image-text pairs. LLaVA trains a large multimodal model by instruction-tuning with GPT-4-generated vision-language data [15]. SoMeLVLM specializes LVLMs for social media artifacts including memes, screenshots, and mixed layouts [18]. The model trains on 654,000 instruction pairs covering social-media-specific capabilities. GPT-4's technical report establishes multimodal capabilities but documents notable limitations including hallucination and reduced reliability without tool use [14]. Direct prompting of these models for fact-checking yields moderate accuracy. Evaluations show that zero-shot prompting produces inconsistent judgments and fabricated explanations when knowledge is lacking.

Systems augmenting LVLMs with external knowledge improve reliability through retrieval-aware reasoning and provenance tracing [19, 20, 21]. These systems show that external knowledge integration improves LVLM reliability, but none combine explicit visual forensics with cross-modal alignment assessment and calibrated confidence aggregation. MIRAGE decomposes verification into specialized modules that examine visual authenticity, alignment, and factual claims separately before integrating signals.

2.3 Agentic Reasoning and Tool Use

Agentic fact-checking frameworks overcome parametric knowledge gaps by grounding language models in external knowledge [22]. Retrieval-augmented generation (RAG) provides the canonical framework for grounding language models in external documents [10]. Retrieval mitigates parametric limits and improves factual accuracy through provenance.

Multi-agent architectures decompose verification into specialized roles. LoCal and LeRuD demonstrate benefits of structured reasoning with external tools for fact-checking and rumor detection [23, 13].

These systems demonstrate benefits of structured reasoning with external tools. Most existing agentic fact-checking frameworks focus primarily on text-only claims [22, 23]. MIRAGE extends agentic decomposition to multimodal settings with explicit modules for visual verification, alignment assessment, and retrieval-augmented question answering. The design enables model-pluggable deployment where any instruction-following LVLM can serve as the reasoning engine.

2.4 Visual Forensics and Cross-Modal Consistency

Detecting AI-generated and manipulated images requires complementary approaches. Surveys identify four detection families: statistical features analyzing noise patterns and compression artifacts, model likelihood methods probing

generator fingerprints, supervised classifiers trained on synthetic datasets, and provenance techniques such as watermarks and C2PA standards [5].

Cross-modal consistency verification detects when authentic images are paired with misleading captions. The system uses reverse-image search and semantic analysis to verify that people, places, and events in images match captions. MDAM3 combines internal visual detectors with external web signals [3]; internal detectors check for manipulation traces using ImageBind embeddings, while external signals from reverse-image search and fact-checking APIs provide context. User studies show that multi-source explanations improve human understanding compared to binary labels [3]. Visual veracity and cross-modal consistency are related but distinct problems. MIRAGE addresses both through separate modules: the visual verification module detects synthetic content, while the relevancy assessment module evaluates image-text alignment using three-level judgments (true, partial, false).

2.5   LLM Reliability, Calibration, and Limitations

Unchecked LLM judgments risk bias and inconsistency. Studies comparing LLM credibility ratings against journalist baselines reveal systematic divergence [12]. GPT-3.5 and GPT-4 exhibit political bias in source assessments, with ratings varying across prompts and models. Comparative evaluations show GPT-4 achieves only 61–72% accuracy on fact-checked claims without retrieval support [9], and performance varies substantially across similar items, indicating unreliability.

Frameworks addressing reliability emphasize uncertainty quantification. Incorporating confidence scores and abstention triggers improves cross-dataset generalization [24]. Models should output calibrated probabilities rather than binary predictions; Brier scores and expected calibration error measure this alignment. Adaptive activation steering further improves truthfulness at inference without fine-tuning [25].

MIRAGE mitigates these concerns through structured decision rules that require factual accuracy, alignment, and genuine imagery for "Not Misinformation" predictions. Module-level confidence scores feed into the judge's aggregation logic instead of a single uncalibrated judgment, and citation-linked rationales expose the evidence driving each decision. These mechanisms enable transparent auditing and iterative refinement targeting known failure modes.

3   Methodology

3.1   Overview

MIRAGE addresses multimodal misinformation detection as a structured reasoning problem. Given an image I and text headline H, the system predicts whether the pair constitutes misinformation, producing a binary classification: Misinformation or Not Misinformation. Unlike supervised models that learn detection patterns from labeled training data, MIRAGE decomposes verification into four sequential modules that each address a distinct aspect of the detection task.

Formally, MIRAGE computes the probability of misinformation through modular composition:

$$P(\text{Misinformation} \mid I, H) = f_{\text{judge}}(v, a, r) \tag{1}$$

where:

$$v = f_{\text{visual}}(I) \quad \text{(visual verification)} \tag{2}$$
$$a = f_{\text{align}}(I, H) \quad \text{(relevancy assessment)} \tag{3}$$
$$r = f_{\text{RAG}}(H, \mathcal{W}) \quad \text{(claim verification)} \tag{4}$$

Here, $f_{\text{visual}} : \mathcal{I} \rightarrow \{0, 1\} \times [0, 1]$ detects AI-generated or manipulated images, outputting a binary indicator and confidence score. $f_{\text{align}} : \mathcal{I} \times \mathcal{H} \rightarrow \{\text{true, partial, false}\} \times [0, 1]$ evaluates cross-modal consistency, producing alignment level and confidence. $f_{\text{RAG}} : \mathcal{H} \times \mathcal{W} \rightarrow \mathcal{Q} \times \mathcal{A}$ generates investigative questions $\mathcal{Q}$ about headline H, retrieves evidence from the web $\mathcal{W}$, and synthesizes citation-linked answers $\mathcal{A}$. Finally, f_{judge} aggregates signals (v, a, r) using structured decision rules to produce the final classification.

The pipeline operates as follows. The Visual Verification Module analyzes image I for signs of AI generation or manipulation. The Relevancy Assessment Module evaluates whether image and text align semantically. The Claim Verification Module generates investigative questions about headline H, retrieves evidence via web search following retrieval-augmented generation principles [10], and synthesizes answers with citations. The Final Judgment Module integrates all signals to produce a calibrated binary decision. Figure 1 illustrates this architecture.

Each module produces structured JSON output. This enables interpretability through transparent reasoning, specialization per detection aspect, and independent module improvements.

Unlike end-to-end approaches that rely solely on model parameters, MIRAGE augments reasoning with external knowledge retrieval. This hybrid approach compensates for the smaller model size while achieving superior performance. We implement all modules using GPT-4o-mini [26], accessed via the OpenAI API with temperature set to 0 for deterministic outputs.

3.2 Visual Verification Module

The visual verification module detects AI-generated or manipulated images through structured prompting of the vision-language model. The module receives a base64-encoded image as input and instructs GPT-4o-mini to analyze the image for characteristic artifacts that indicate synthetic or altered content.

The prompt directs the model to examine both technical artifacts (warped hands, inconsistent lighting, impossible anatomy) and contextual anomalies (surreal object combinations, dreamlike elements, impossible scenarios). The model outputs structured JSON containing a binary AI-generation judgment, confidence score calibrated between 0 and 1, explanation of observed patterns, and a list of specific anomalies detected. See Appendix for complete prompt text.

This approach differs from traditional deepfake detectors that analyze statistical properties or generator fingerprints. Instead, MIRAGE leverages the vision-language model's semantic understanding to identify both technical imperfections and contextual impossibilities that humans would recognize. The confidence calibration guides the model to reserve high scores (0.8-1.0) for clear artifacts or impossible contexts, moderate scores (0.4-0.6) for suspicious but ambiguous cases, and low scores (0.0-0.2) for apparently genuine images.

3.3 Relevancy Assessment Module

The relevancy assessment module evaluates whether the image depicts the specific subject, event, or context described in the headline. This addresses out-of-context misinformation where authentic images are repurposed with misleading captions.

The module produces three-level alignment judgments. True alignment indicates the image clearly depicts the specific subject or event mentioned in the headline. Partial alignment indicates the image shows related content but lacks confirmation of specific details—common in legitimate journalism where stock photos illustrate stories. False alignment indicates the image shows a different subject or contradicts the headline.

Critically, the prompt instructs the model to use confidence scores to distinguish legitimate partial alignment from deceptive mismatches. High-confidence partial alignment ($\geq$0.7) suggests the right subject with incomplete details visible. Low-confidence partial alignment (<0.7) suggests possibly wrong subject with only superficial similarity. This calibration enables the final judgment module to correctly classify legitimate news articles that use illustrative imagery while flagging deceptive out-of-context posts.

The module outputs structured JSON containing the alignment level (true/partial/false), calibrated confidence score, and explanation justifying both the alignment judgment and confidence level. See Appendix for complete prompt.

3.4 Claim Verification Module

The claim verification module performs retrieval-augmented fact-checking through a two-stage process: question generation followed by answer synthesis. This design enables iterative investigation where each reasoning chain builds on findings from previous questions.

3.4.1 Question Generation

The question generator produces investigative queries in three sequential chains. Each chain generates three questions (k=3), with explicit duplicate detection preventing redundant searches. The system maintains a list of previously asked questions and recent answers, enabling adaptive investigation.

The prompt instructs the model to use concrete search strategies. First queries verify core claims exist ("Did [event] happen?"). Follow-up queries check specific details including dates, people, and locations. The generator uses concrete terms rather than abstract phrasing, producing queries suitable for web search engines.

Chain 1 addresses direct fact-checking. Chain 2 generates context questions informed by Chain 1 findings—if Chain 1 confirms an event occurred, Chain 2 investigates surrounding circumstances. Chain 3 asks follow-up questions to

resolve ambiguities from earlier chains. This sequential structure enables deeper investigation than single-pass retrieval while maintaining focus on verifying the specific headline claims.

3.4.2 Web Search and Answer Synthesis

For each generated question, the system queries DuckDuckGo web search and retrieves up to 5 results. The answer synthesis prompt receives the question along with titles, URLs, and text snippets from search results. The model synthesizes concise answers (2-5 sentences) grounded in the provided evidence.

Critically, the prompt requires explicit source citations. The model must cite specific URLs and titles for all factual claims in its answer. When sources present conflicting information, the model summarizes differing perspectives and cites each source separately. This citation-grounded approach ensures answers remain traceable to source material for audit purposes.

The model outputs structured JSON containing the textual answer, array of citation objects with URLs and titles, calibrated confidence score reflecting answer certainty, and a brief rationale explaining how it arrived at the answer. Low confidence scores signal cases where search results were sparse, contradictory, or failed to address the question—information the final judgment module uses to avoid overconfident classification.

Complete prompts for both question generation and answer synthesis appear in the Appendix.

3.5 Experimental Setup

We evaluate MIRAGE on MMFakeBench [2], a multimodal misinformation benchmark containing 11,000 image-text pairs across three distortion categories: textual veracity distortion (false claims with real or AI images), visual veracity distortion (real text with manipulated/synthetic images), and cross-modal consistency distortion (real text with wrong-context real images). The dataset includes authentic news from VisualNews as negative examples.

The benchmark splits into train (6,000 samples), validation (1,000 samples), and test (10,000 samples) sets. Class distribution is approximately 70% misinformation and 30% authentic content, creating a challenging imbalance where naive "classify everything as fake" strategies achieve high accuracy but unacceptable false positive rates.

We conduct our primary evaluation on the 1,000-sample validation set. To validate generalization, we performed stratified evaluation on 5,000 test samples that mirror the full 10,000-sample distribution through deterministic sampling with seed 42.

3.6 Evaluation Metrics

We follow the MMFakeBench binary classification protocol [2] with "Misinformation" as the positive class. We report F1 score (primary metric), accuracy, precision, and per-class recall.

Balanced recall analysis is critical because high accuracy can mislead when achieved by exploiting class imbalance. A naive classifier labeling everything "fake" achieves 70% accuracy on the 70-30 distribution but 0% real recall, misclassifying all authentic news. We report Recall (Misinformation) and Recall (Not Misinformation) separately to expose such failures.

We provide per-class breakdowns by misinformation type (textual distortion, visual distortion, cross-modal mismatch, authentic content) to identify category-specific strengths and weaknesses. For visual verification, we report AI-generated image detection metrics independent of textual veracity.

Calibration is assessed using Brier score [27] and Expected Calibration Error [28] to measure confidence reliability. All metrics use the full validation set (1,000 samples) with uncertain predictions penalized as incorrect.

3.7 Implementation Details

MIRAGE uses GPT-4o-mini (gpt-4o-mini-2024-07-18) via the OpenAI API with temperature 0 for deterministic outputs, enabling exact replication for ablations and baselines. Visual verification and relevancy modules use vision–language prompting with base64-encoded images.

For claim verification, we use DuckDuckGo's free API to support large-scale evaluation. Each sample runs 3 question chains $\times$ 3 questions, retrieving up to 5 results per question; cross-chain deduplication yields 4,406 unique queries for 1,000 validation items (4.4 per sample). Reliability safeguards include exponential-backoff retries (up to 2), 35 s per-query timeouts, a 1.8 s rate limit, and result caching. The final judge consumes structured JSON from all modules.

Table 1: Main Results on MMFakeBench Validation Set

Method	Model	F1	Acc	Prec	Sens	Spec
GPT-4V Standard[†]	GPT-4V	72.3	75.6	72.1	72.8	—
GPT-4V MMD-Agent[†]	GPT-4V	74.0	76.8	73.4	75.5	—
MIRAGE	GPT-4o-mini	**81.65**	75.1	**84.3**	79.14	**65.67**

† Baseline results from Liu et al. [2]; specificity not reported.

Evaluation on the 1,000-sample set used 103M prompt tokens and 1.4M completion tokens. At $0.15/M$ (prompt) and $0.60/M$ (completion), cost is $\approx$$16.29$ per $1,000$ samples; using GPT-4V instead ($5.00/M$ prompt, $15.00/M$ completion) would be $\approx$$536.45$—$33\times$ higher, highlighting MIRAGE's cost efficiency from architecture rather than model scale.

3.8 Baselines

We compare MIRAGE against all baselines reported in Liu et al. [2]. The primary baselines are GPT-4V with standard prompting (72.3% F1, 75.6% accuracy on validation) and GPT-4V with the MMD-Agent framework (74.0% F1, 76.8% accuracy). MMD-Agent decomposes the detection task hierarchically into textual veracity checking, visual veracity checking, and cross-modal consistency reasoning, then integrates reasoning with Wikipedia-based external knowledge retrieval. Additional baselines include open-source LVLMs ranging from 7B to 34B parameters (LLaVA-1.6, InstructBLIP, BLIP2, VILA) with F1 scores between 7.9% and 67.2%, as well as traditional detection methods. Complete baseline results are available in Table 6 of Liu et al. [2].

4 Results and Analysis

We present comprehensive evaluation results on the MMFakeBench validation set (1,000 samples), including main performance comparisons, ablation studies, and per-class breakdowns. All experiments use identical stratified sampling to ensure fair comparisons across configurations.

4.1 Main Results

Table 1 presents our main results compared to baseline methods from Liu et al. [2]. MIRAGE achieves 81.65% F1 on the validation set, outperforming GPT-4V with standard prompting (72.3% F1) by 9.35 percentage points and GPT-4V with MMD-Agent (74.0% F1) by 7.65 points. This represents the strongest reported performance on MMFakeBench validation using vision-language models.

While overall accuracy (75.1%) is comparable to baseline systems (75.6-76.8%), F1 score reveals the critical difference in detection capability. MIRAGE achieves balanced performance across both classes with 79.14% recall on misinformation and 65.67% recall on real news, demonstrating substantially lower false positive rates than naive approaches. The system maintains 84.3% precision, correctly identifying misinformation while minimizing misclassification of authentic content.

The modular agentic design with specialized reasoning stages and external knowledge retrieval outperforms end-to-end approaches for mixed-source misinformation detection. The performance gain stems from decomposing the complex detection task into targeted sub-problems that each module addresses with tailored strategies.

Test set evaluation on 5,000 samples[1] yields 81.44% F1 and 75.08% accuracy, confirming that performance generalizes beyond the validation set with less than 0.3 point variation in both metrics.

4.2 Ablation Studies

Table 2 demonstrates each module's contribution through systematic ablations. The judge-only configuration, using only headline text without any verification modules, illustrates the baseline problem: it achieves 70.8% accuracy by classifying nearly everything as fake (100% fake recall, 2.67% real recall), resulting in a 97.3% false positive rate. This approach exploits the 70-30 class distribution but is unusable for deployment.

---

[1]Stratified random sample (50%) preserving the original 70-30 class distribution of the full 10,000-sample test set, selected once with seed 42 for computational efficiency.

Table 2: Ablation Study Results (Validation Set)

Configuration	Acc	F1	Sens	Spec	FP Rate	Δ F1
Full MIRAGE	**75.1**	**81.65**	79.14	**65.67**	**34.3**	—
No Visual Verification	69.1	76.47	71.71	63.0	37.0	-5.18
No Claim Verification	72.9	78.68	71.43	76.33	23.7	-2.97
Judge Only	70.8	82.74	**100.0**	2.67	97.3	+1.09

Δ F1 = Change in F1 score relative to full MIRAGE.
FP Rate = False Positive Rate on authentic content.
Sens = Recall (Misinformation); Spec = Recall (Not Misinformation).

Table 3: Detection Accuracy by Misinformation Type

Type	Full	No Visual	No RAG	Judge Only
Visual distortion	92.0	56.0	92.0	100.0
Textual distortion	84.33	80.67	68.67	100.0
Cross-modal mismatch	69.67	68.0	67.33	100.0
Authentic content	65.67	63.0	76.33	2.67

Visual module critical for AI images; RAG critical for textual lies.

Removing visual verification drops overall accuracy by 6 percentage points (69.1% vs 75.1%) and F1 by 5.18 points, with fake recall decreasing from 79.14% to 71.71%. More critically, Table 3 reveals this primarily impacts visual distortion detection, dropping from 92% to 56% accuracy—a 36-point degradation. This validates that explicit visual analysis is essential for detecting AI-generated and manipulated imagery.

Removing claim verification (no RAG-based question answering) achieves higher real recall (76.33%) but lower fake recall (71.43%) and reduced F1 (78.68%). The per-class breakdown shows this disproportionately affects textual misinformation detection, dropping from 84.33% to 68.67%—a 15.66-point loss. This demonstrates that external knowledge retrieval is critical for fact-checking claims requiring real-world verification beyond the model's parametric knowledge.

4.3 Per-Class Performance Analysis

Table 3 shows module specialization. Removing visual verification drops AI-image accuracy from 92% to 56%, with little effect on textual distortion (84.33%→80.67%) or cross-modal mismatch (69.67%→68%), confirming its role is specific to synthetic-image detection. Removing claim verification reduces textual distortion from 84.33%→68.67% while leaving visual distortion at 92%, underscoring the need for retrieval beyond parametric memory.

For real news (specificity), the full system recalls 65.67% versus 2.67% for a judge-only baseline; omitting claim verification raises recall to 76.33%, indicating that aggressive fact-checking can over-flag rare but true items. This sensitivity–specificity balance is a deploy-time knob.

Cross-modal mismatches remain hardest (69.67%). Out-of-context posts pair genuine imagery with partly true captions, yielding partial alignment that blurs the boundary between illustrative reporting and miscontextualization; our relevancy module captures this via a "partial" label, but finer reasoning is needed.

4.4 Error Analysis

We analyzed all 249 errors from the 1,000-sample validation run to identify systematic failure patterns. False positives (103 errors, 34.3% of real news) occur primarily due to weak image-text alignment (61.2% of false positives). Legitimate news articles frequently use generic stock photos or illustrative images rather than literal depictions of events. For example, an authentic VisualNews article about a beached whale was flagged as fake because the image showed a whale but could not confirm the specific location mentioned in the headline. The system incorrectly interpreted partial alignment as evidence of manipulation.

Unusual but true events also trigger false alarms when web search returns limited results for niche topics. A genuine sports article about Northwestern's Gator Bowl victory was flagged as fake simply because the specific game had

8

minimal online coverage, demonstrating that absence of search results does not equal evidence of falsehood but that can be tricky to detect.

5 Discussion

5.1 Implications for Web-Scale Misinformation Detection

Our results demonstrate that modular agentic reasoning with web retrieval achieves 81.65% F1 on MMFakeBench, outperforming prior zero-shot approaches (GPT-4V at 74.0% F1) without requiring domain-specific training data. This has direct implications for web-scale detection where manipulation tactics evolve faster than labeled datasets can be curated. However, since our model relies on web retrieval methods, its performance relies heavily on information found in the web. If there isn't sufficient information available regarding the topic in the web, our model might not perform as well.

Our 34.3% false positive rate means approximately one-third of authentic content is incorrectly flagged. However, we consider this to be a conservative approach, where authentic content being labelled "misinformation" is preferred compared to the other way around. In the future, our work can be extended to incorporate human-in-the-loop workflows where moderators make final decisions.

Our findings extend prior work on agentic fact-checking [22, 20, 19] by demonstrating that multimodal verification requires specialized modules for visual authenticity, cross-modal consistency, and textual claims rather than generic vision-language prompting. The framework's model-pluggable design and citation-linked outputs address transparency requirements for automated content moderation [12], enabling deployment across languages and platforms where supervised training data remains scarce.

5.2 Ethical Considerations

Automated misinformation detection systems raise ethical concerns, requiring careful deployment considerations. Vision-language models inherit training data biases that may affect detection accuracy across demographic groups or political viewpoints. Furthermore, the dual-use nature of detection systems means the same techniques can be reverse-engineered to identify weaknesses in the system and bypass them. Researchers should take these considerations into account while designing systems. Therefore, future work should focus on evaluating biases in the system and ensuring safe deployment of the models.

Finally, we position MIRAGE as a tool for fact-checkers, platform moderators, and even end-users. However, it is not a replacement for media literacy education and appropriate policies. To ensure responsible use of such systems, we recommend ongoing dialog among researchers, platforms, policymakers, and civil society.

5.3 Limitations

MMFakeBench's 70:30 class skew means a naïve "all fake" strategy can appear accurate while producing unacceptable false positives; our rules enforce evidence-weighted decisions on ambiguous cases rather than heuristics. Web search mitigates knowledge cutoffs but cannot verify very recent or niche events with sparse indexing; result quality is uneven and potentially susceptible to SEO manipulation. Finally, alignment thresholds require tuning as stricter settings reduce false negatives but raise false positives on legitimate illustrative imagery, a deployment-specific sensitivity–specificity trade-off.

Computational cost remains substantial: processing 1,000 samples triggers 4,400 web searches and 103M tokens and still takes hours even with caching and parallelism, precluding sub-second moderation use. Reliance on commercial APIs (OpenAI, DuckDuckGo) introduces vendor lock-in and potential privacy concerns, and the current system is English-only, requiring additional validation and prompt design for multilingual use.

6 Conclusion

Multimodal misinformation at scale requires verification approaches that do not depend on domain-specific training data. MIRAGE decomposes detection into specialized modules for visual forensics, cross-modal alignment, and retrieval-augmented fact-checking. The model-pluggable architecture enables deployment with any instruction-following vision-language model. The system achieves 81.65% F1 on MMFakeBench validation set, outperforming the strongest zero-shot baseline by 7.65 points while maintaining 34.3% false positive rate on authentic content. Ablation studies confirm that both visual verification and retrieval-augmented reasoning contribute substantially. These results demonstrate that

decomposed agentic reasoning with web retrieval can match supervised detector performance without requiring labeled training data. Future work should address three limitations identified through error analysis: improving detection of high-fidelity AI images that lack obvious artifacts, developing better handling of niche topics with limited web coverage, and extending the framework to multilingual content beyond English. Exploring confidence threshold tuning for deployment-specific false positive and false negative trade-offs would improve practical applicability. Integration with human-in-the-loop workflows could leverage MIRAGE's transparent citation-linked rationales for hybrid human-AI fact-checking at scale. Therefore, our work opens new pathways for multimodal misinformation detection across platforms, and modalities where labeled data remains scarce.

Acknowledgments

References

[1] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[2] Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. *arXiv preprint arXiv:2406.08772*, 2024.

[3] Qingzheng Xu, Heming Du, Szymon Łukasik, Tianqing Zhu, Sen Wang, and Xin Yu. Mdam3: A misinformation detection and analysis framework for multitype multimodal media. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 5285–5296, New York, NY, USA, 2025. Association for Computing Machinery.

[4] Ashley M Hopkins, Bradley D Menz, and Michael J Sorich. Potential of large language models as tools against medical disinformation—reply. *JAMA Internal Medicine*, 184(4):450–451, 2024.

[5] Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1302–1305, New York, NY, USA, 2024. Association for Computing Machinery.

[6] Adrian K. Yee. The limits of machine learning models of misinformation. *AI & SOCIETY*, pages 1–14, 2025.

[7] Yike Wu, Yang Xiao, Mengting Hu, Mengying Liu, Pengcheng Wang, and Mingming Liu. Towards robust evidence-aware fake news detection via improving semantic perception. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16607–16618, Torino, Italia, May 2024. ELRA and ICCL.

[8] Declan Iain Campbell, Sunayana Rane, Tyler Giallanza, C. Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Whittington Webb. Understanding the limits of vision language models through the lens of the binding problem. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[9] Kevin Matthe Caramancion. News verifiers showdown: a comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. In *2023 IEEE Future Networks World Forum (FNWF)*, pages 1–6. IEEE, 2023.

[10] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[11] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali, November 2023. Association for Computational Linguistics.

[12] Kai-Cheng Yang and Filippo Menczer. Accuracy and political bias of news source credibility ratings by large language models. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 127–137, 2025.

[13] Qiang Liu, Xiang Tao, Junfei Wu, Shu Wu, and Liang Wang. Can large language models detect rumors on social media? *arXiv preprint arXiv:2402.03916*, 2024.

[14] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[16] John Dougrez-Lewis, Mahmud Elahi Akhter, Federico Ruggeri, Sebastian Löbbers, Yulan He, and Maria Liakata. Assessing the reasoning capabilities of llms in the context of evidence-based claim verification. *arXiv preprint arXiv:2402.10735*, 2024.

[17] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13052–13062, 2024.

[18] Xinnong Zhang, Haoyu Kuang, Xinyi Mou, Hanjia Lyu, Kun Wu, Siming Chen, Jiebo Luo, Xuanjing Huang, and Zhongyu Wei. Somelvlm: A large vision language model for social media processing. *arXiv preprint arXiv:2402.13022*, 2024.

[19] Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. Lemma: towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*, 2024.

[20] M Abdul Khaliq, Paul Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. *arXiv preprint arXiv:2404.12065*, 2024.

[21] Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*, 2024.

[22] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[23] Jiatong Ma, Linmei Hu, Rang Li, and Wenbo Fu. Local: Logical and causal fact-checking with llm-based multi-agents. In *Proceedings of the ACM on Web Conference 2025*, pages 1614–1625, 2025.

[24] Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*, 2023.

[25] Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2562–2578, New York, NY, USA, 2025. Association for Computing Machinery.

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[27] W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[28] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Appendix

7 System Prompts

This appendix documents the complete prompt templates used in each MIRAGE module. All prompts instruct GPT-4o-mini to output strict JSON for downstream processing.

7.1 Visual Verification Module

The visual verification module uses a two-part prompt structure to detect AI-generated or manipulated images.

7.1.1 System Prompt

```
You are an AI image detection expert. Analyze
images for signs of AI generation or manipulation.
Respond with strict JSON only.
```

7.1.2 User Instruction

```
Task: Examine this image for signs of AI generation
or manipulation.

Detection criteria:

Technical artifacts:
- Warped hands, extra fingers, impossible anatomy
- Nonsensical text, garbled signs
- Inconsistent lighting, impossible shadows
- Unnatural textures, blending errors

Contextual anomalies:
- Surreal object combinations
  (e.g., clown in bathroom mirror)
- Impossible scenarios or physics violations
- Dreamlike or fantastical elements in otherwise
  normal scenes
- Objects that don't belong in the context

Be suspicious when:
- Technically perfect BUT contextually bizarre
- Minor details don't make sense
- Scene feels 'off' even without obvious artifacts

Confidence calibration:
- 0.8-1.0: Clear technical artifacts OR
           impossible context
- 0.6-0.8: Strong evidence, multiple anomalies
- 0.4-0.6: Moderate suspicion, some anomalies
- 0.2-0.4: Minor concerns, could be unusual
           real photo
- 0.0-0.2: Appears genuine

Prioritize contextual impossibility as much as
technical quality.

Output JSON:
{
  "ai_generated": boolean,
  "confidence": 0.0-1.0,
  "explanation": "What you observed",
  "anomalies": ["specific issues"]
}
```

The prompt directs the model to examine both technical artifacts (warped anatomy, inconsistent lighting) and contextual anomalies (surreal combinations, impossible scenarios). Confidence thresholds guide calibrated scoring where higher values indicate stronger evidence of AI generation.

7.2 Relevancy Assessment Module

The relevancy module evaluates image-text alignment through three-level classification to detect out-of-context misinformation.

7.2.1 System Prompt

```
You are an image-headline relevancy assessor.
Evaluate if the image depicts the specific subject
and context described in the headline.
Respond with strict JSON only.
```

7.2.2 User Instruction Template

```
Headline: {headline}

Task: Does the image show the specific subject/event
from the headline?

Classification:
- aligned=true: Image clearly depicts the specific
  subject/event
- aligned=partial: Image shows related content but
  lacks confirmation of specifics
- aligned=false: Image shows different subject/event

Confidence calibration (CRITICAL):
- 0.9-1.0: Can identify specific people/places/events
  mentioned in headline
- 0.7-0.9: Shows correct general context but cannot
  confirm specific details
- 0.5-0.7: Shows related content but connection is
  weak or ambiguous
- 0.3-0.5: Superficial similarity only, likely
  wrong subject
- 0.0-0.3: No meaningful connection

For partial alignment:
- High confidence (0.7+): Right subject, details
  not fully visible
- Low confidence (<0.7): Possibly wrong subject,
  superficial match

Output JSON:
{
  "aligned": "true" | "partial" | "false",
  "confidence": 0.0-1.0,
  "explanation": "What you observed and why this
                  confidence"
}
```

The prompt explicitly defines confidence ranges to distinguish legitimate partial alignment, common in journalistic practices where stock photos illustrate stories, from deceptive cross-modal mismatches where images depict different subjects than captions claim.

7.3 Claim Verification: Question Generation

The question generator produces investigative queries in three sequential chains, enabling iterative investigation where each chain builds on findings from previous questions.

7.3.1 System Prompt

```
You are an investigative assistant. Generate search
queries to verify if an image-headline pairing is
authentic or misleading. Focus on verifying both
```

```
the headline's claims AND whether the image matches.
Respond with strict JSON only.
```

7.3.2 User Instruction Template

```
Task: Generate {k} Google-style search queries to
verify this headline.

Headline: {headline}

Already asked:
{prior_questions}

Recent answers:
{answered_qa_pairs}

Query strategy:
- First query: Verify core claim exists
  ("Did [event] happen?", "Is [fact] true?")
- Follow-up queries: Check specific details, dates,
  people involved
- Use concrete terms: names, places, dates,
  specific events
- Keep queries short (4-8 words)

Generate exactly {k} NEW queries (avoid duplicates).

Output as JSON array only:
["query 1", "query 2", ...]
```

The template includes prior questions and recent answers to enable adaptive investigation. The system generates three questions per chain (k=3), with explicit duplicate detection preventing redundant searches across the three chains. Each chain explores different verification angles: Chain 1 addresses direct fact-checking, Chain 2 generates context questions informed by Chain 1 findings, and Chain 3 asks follow-up questions to resolve ambiguities from earlier chains.

7.4 Claim Verification: Answer Synthesis

The answer generator synthesizes citation-linked responses from web search results, grounding claims in external evidence rather than relying on parametric knowledge.

7.4.1 System Prompt

```
You are a careful fact-checking assistant. Using
the provided web snippets, answer the user's
question concisely and cite sources. If sources
disagree, summarize the differing views and cite
each. Respond with strict JSON only.
```

7.4.2 User Instruction Template

```
Question: {question}

Sources:
[1] {title_1}
URL: {url_1}
Snippet: {description_1}

[2] {title_2}
URL: {url_2}
Snippet: {description_2}
```

```
[... up to 5 sources ...]

Instructions: Produce strict JSON with keys:
  answer: short textual answer (2-5 sentences)
  citations: array of objects {url, title} for
             the sources you used
  confidence: number in [0,1]
  rationale: one or two sentences on how you
             arrived at the answer
```

The prompt instructs the model to synthesize concise answers from provided search result snippets, include calibrated confidence scores between 0 and 1, and cite sources explicitly by URL and title. When sources present conflicting information, the model summarizes differing perspectives and cites each source, enabling transparent assessment of evidence quality. This citation-grounded approach ensures answers remain traceable to source material for audit purposes.

7.5 Final Judgment Module

The judgment module integrates all verification signals using structured decision rules designed to prevent naive classification strategies that exploit class imbalance.

7.5.1 System Prompt

```
You are a misinformation detector. Evaluate
image-headline pairings using multiple signals.

Input signals:
- relevancy: {aligned, confidence, explanation}
- visual_veracity: {ai_generated, confidence,
                    anomalies}
- qa_analysis: Verification of headline claims

Decision logic:

STEP 1 - Definitive misinformation (ANY of these):
- Headline verifiably false (Q/A contradicts)
- Image is AI-generated (ai_generated=true,
  confidence>0.6)
- Image completely wrong (aligned=false)
→ Classify as Misinformation

STEP 2 - Evaluate partial alignment cases:
When aligned=partial, USE confidence to distinguish:

High confidence partial ($\geq$0.7):
- Interpretation: Right subject, incomplete
  details visible
- If headline true AND image genuine
  → Not Misinformation

Low confidence partial (<0.7):
- Interpretation: Possibly wrong subject,
  superficial match
- If headline true AND image genuine
  → Misinformation (likely mismatch)

STEP 3 - Verify genuine content (ALL required):
- Headline accurate (Q/A supports)
```

```
- Image genuinely relates (aligned=true OR
  partial with confidence$\geq$0.7)
- Image authentic (ai_generated=false)
→ Not Misinformation

Return JSON:
{
  "label": "Misinformation" | "Not Misinformation",
  "confidence": 0.0-1.0,
  "rationale": "Which signals were decisive",
  "key_factors": ["signal: value"]
}
```

7.5.2 User Message Template

```
You are given the following analysis JSON for a
headline+image pair. Make a final misinformation
judgment.

Analysis JSON (compact):
{
  "headline": "{headline_text}",
  "image_path": "{path}",
  "relevancy": {aligned, confidence, explanation},
  "visual_veracity": {ai_generated, confidence,
                      anomalies},
  "best_qa_per_chain": [
    {
      "question": "{q1}",
      "answer": "{a1}",
      "confidence": 0.X,
      "citations_count": N
    },
    ...
  ]
}
```

```
Respond ONLY with JSON:
{"label":..., "confidence":..., "rationale":...,
 "key_factors":[...]}
```

The judgment prompt encodes explicit decision rules to prevent the model from defaulting to majority-class predictions. Classification as Not Misinformation requires satisfying all three criteria simultaneously: factual headline accuracy verified through question-answer evidence, genuine image-text alignment (true or high-confidence partial), and authentic non-AI-generated imagery. This asymmetric decision structure forces comprehensive verification before labeling content as legitimate, addressing the class imbalance problem where naive "classify everything as fake" strategies achieve high accuracy but unacceptable false positive rates.

7.6 Implementation Notes

All prompts use temperature 0 for deterministic outputs, enabling exact replication across runs. The visual verification and relevancy modules receive base64-encoded images via GPT-4o-mini's vision capabilities, while the claim verification modules operate on text only. JSON output formatting is enforced through explicit prompt instructions rather than OpenAI function calling to maintain compatibility across model versions and facilitate future integration with alternative vision-language models. The structured output format enables programmatic parsing and integration into content moderation pipelines requiring audit trails.