

Is a Picture Worth a Thousand Words? Agentic Multimodal Fact-Checking for Adaptive Use of Visual Evidence

Anonymous ACL submission

Abstract

Automated fact-checking is a crucial task not only in journalism but also across web platforms, where it supports a responsible information ecosystem and mitigates the harms of misinformation. While recent research has progressed from text-only to multimodal fact-checking, a prevailing assumption is that incorporating visual evidence universally improves performance. In this work, we challenge this assumption and show that indiscriminate use of multimodal evidence can reduce accuracy. To address this challenge, we propose an agentic fact-checking framework, AMuFC (Adaptive Agentic Multimodal Fact-Checking with Visual Evidence Necessity). The proposed method employs an *Analyzer* that determines whether visual evidence is necessary for claim verification and a *Verifier* that predicts claim veracity conditioned on both the retrieved evidence and the Analyzer’s assessment. Experimental results show that incorporating the Analyzer’s assessment of visual evidence necessity into the Verifier’s prediction yields substantial improvements in verification performance. Case studies further support its generalizability across diverse fact-checking scenarios.

1 Introduction

Fact-checking—the process of determining the veracity of claims—is a cornerstone of responsible journalism and an essential mechanism for mitigating the harms of misinformation in broader web environments (Hassan et al., 2015; Adair et al., 2017). Building on early work focused on text-only verification, recent research shifts toward multimodal fact-checking, which addresses scenarios in which evidence is conveyed through multiple modalities, primarily images and text (Hameleers et al., 2020; Alam et al., 2022; Biamby et al., 2022). Various approaches integrate these modalities for fact verification. For example, multimodal fusion-based classifiers combine retrieved image–text evidence

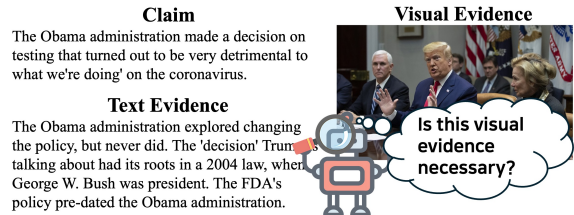


Figure 1: Illustration of the key idea.

for fact verification (Yao et al., 2023), while more recent work leverages reinforcement learning to summarize multimodal inputs (Chen et al., 2024). A common finding across these studies is that leveraging both textual and visual evidence improves fact-checking accuracy (Tahmasebi et al., 2024; Cekinel et al., 2025), echoing the adage that *a picture is worth a thousand words*.

In this study, we challenge a prevailing assumption in prior work—the presumed necessity of visual evidence, as illustrated in Figure 1. We analyze claim types according to their dependence on visual evidence using both manual and automated methods. While some claims are unverifiable without visual evidence (Cekinel et al., 2025), our analysis reveals that a naïve strategy of always incorporating visual evidence can cause models to underperform compared to text-only baselines. We further show that verification performance improves when evidence is selected adaptively, as demonstrated in an oracle setting in which visual evidence is used only when it contributes to verification.

Motivated by these findings, we propose AMuFC, an agentic fact-verification framework that adaptively incorporates visual evidence through two collaborating vision-language model (VLM) agents. An *Analyzer* assesses whether visual evidence is necessary for verification, while a *Verifier* predicts claim veracity conditioned on the Analyzer’s natural-language assessment. Experimental results demonstrate the effectiveness of

Dataset	Primary Inputs	Evidence	Task	Label Source	Claim Source	External Knowledge Source	Access
Fakeddit (Nakamura et al., 2020)	Img/Txt	-	Mixed	Synthetic	Reddit	✗	✓
NeuralNews (Tan et al., 2020)	Img/Txt	-	OOC	Mixed	Grover/GoodNews	✗	✓
NewsCLIPings (Luo et al., 2021)	Img/Txt	-	OOC	Mixed	CLIP/VisualNews	✗	✓
Factify (Suryavardan et al., 2023)	Img/Txt	Img/Txt	OOC	Mixed	Twitter	✗	✓
VERITE (Papadopoulos et al., 2024)	Img/Txt	-	Mixed	Mixed	Snopes/Reuters	✗	✓
Fauxtography (Zlatkova et al., 2019)	Img/Txt	Meta	FV	Expert	Snopes/Reuters	✗	✓
MMM (Gupta et al., 2022)	Img/Txt	Txt	FV	Mixed	FC webs	✗	✗
FactDrill (Singhal et al., 2022)	Txt	Vid/Aud/Img/Txt/Meta	FV	Expert	FC webs	✗	✗
ChartFC (Akhtar et al., 2023a)	Txt	Img	FV	Crowd	TabFact	✗	✓
ChartCheck (Akhtar et al., 2024)	Txt	Img/Txt	FV	Crowd	Wikimedia	✗	✓
MOCHEG (Yao et al., 2023)	Txt	Img/Txt	FV	Expert	FC webs	✓	✓
FIN-FACT (Rangapur et al., 2025)	Txt	Img/Txt	FV	Expert	FC webs	✗	✓
AVerImaTeC (Cao et al., 2025)	Img/Txt	Img/Txt	FV	Expert	FC webs	✓	✓

Table 1: Datasets for multimodal misinformation detection. OOC and FV denote out-of-context misinformation and fact verification, respectively.

adaptive multimodal evidence use: our best configuration, with Llama-3.2-11B-Vision as the Analyzer and Qwen2-VL-7B as the Verifier, achieves an accuracy of 0.612 and outperforms existing methods. We also conduct case studies on test-only datasets, including newly collected claims with web-searched evidence, where performance gains from incorporating the Analyzer’s assessment persist, underscoring the generalizability of AMUFC across diverse fact-checking scenarios.

The key findings and contributions of this study are summarized below.

- **Necessity of visual evidence:** We show that a naïve approach of always incorporating visual evidence can degrade performance.
- **Effectiveness of agentic fact-checking:** We propose AMUFC, an agentic multimodal fact-checking framework that adaptively utilizes visual evidence, and demonstrate substantial improvements over existing methods.
- **Generalizability across diverse fact-checking scenarios:** Case studies on two datasets confirm the effectiveness of AMUFC and support its generalizability.

2 Related Work

2.1 Multimodal Fact-checking

Previous research on multimodal fact-checking, or multimodal misinformation detection, can be broadly categorized into three different tasks according to the verdict prediction settings: *manipulation classification*, *out-of-context detection*, and *fact verification* (Akhtar et al., 2023b). Manipulation classification focuses on detecting manipulated

misinformation within a single modality (Gupta et al., 2013; Boididou et al., 2014; Huh et al., 2018; Heller et al., 2018; Shao et al., 2023). Out-of-context detection aims to determine whether a given image-text pair conveys consistent contextual information (Nakamura et al., 2020; Tan et al., 2020; Luo et al., 2021; Suryavardan et al., 2023; Papadopoulos et al., 2024). Fact verification, which is the focus of this paper, seeks to predict the truthfulness of a claim, typically in textual form (Singhal et al., 2022; Akhtar et al., 2023a, 2024; Yao et al., 2023). Table 1 summarizes widely used multimodal fact-checking datasets.

Following early attempts of zero-shot fact verification (Lee et al., 2020), most subsequent works have adopted the retrieval-augmented generation (RAG) framework (Lewis et al., 2020), which retrieves both image and textual evidence and then leverages the retrieved content to assist a verifier model. Because evidence quality is critical for RAG-based verification, several studies have focused on mitigating the issue of noisy or irrelevant contexts. For example, MetaSumPerceiver was proposed to summarize multimodal document evidence (Chen et al., 2024). Other work introduced re-ranking methods based on predicted token probabilities (Tahmasebi et al., 2024), while additional approaches proposed tool-based RAG pipelines for fact verification. Most recently, Pang et al. (2025) proposed a hypergraph transformer-based fact-checking framework to capture high-order relationships between claim and evidence.

To the best of our knowledge, our study is the first to explicitly examine the necessity of visual evidence and to incorporate it within an agentic fact-verification framework.

2.2 Adaptive Retrieval

We review previous research on adaptive retrieval for fact verification and, more broadly, RAG. Although RAG improves performance on knowledge-intensive tasks (Lewis et al., 2020), inaccurate or irrelevant retrieval can introduce noisy context that degrades performance (Yoran et al., 2024; Shi et al., 2023). Adaptive retrieval strategies mitigate this issue by retrieving on demand or filtering unhelpful context (Tang et al., 2025; Parekh et al., 2025). For example, Self-RAG (Asai et al., 2023) enables dynamic retrieval via reflection tokens, while other approaches select RAG strategies based on query complexity (Jeong et al., 2024) or use trained relevance estimators to assess context utility (Kim and Lee, 2024). More recent work leverages hidden-state representations to decide whether retrieval is necessary (Baek et al., 2025) or dynamically incorporates retrieved passages based on quality (In et al., 2025).

In fact verification, agent-based iterative frameworks integrate retrieval and verification across multiple rounds (Xie et al., 2025), while PASS-FC augments claims with temporal and entity grounding for progressive search (Zhuang, 2025). Related work also explores dynamically adjusting search depth to extract multimodal evidence (Braun et al., 2025). In contrast to prior work that primarily targets textual relevance or pre-filters irrelevant evidence, this study introduces an agentic framework in which a VLM-based agent assesses the necessity of visual evidence relative to the textual evidence and another VLM predicts a verdict based on this natural-language assessment, enabling adaptive and selective use of visual information for fact verification.

3 Task and Dataset

Given a textual claim c and an external knowledge source K , the task of *multimodal fact verification* aims to determine the veracity of c by retrieving and reasoning over both textual and visual evidence. Specifically, a fact-checking system retrieves a set of textual evidence $E_t \subset K$ and visual evidence $E_v \subset K$, and then predicts a verdict label $y \in \{\textit{supported}, \textit{refuted}, \textit{not enough information (NEI)}\}$ based on (c, E_t, E_v) . This formulation extends the standard text-only fact verification setting, as exemplified by FEVER (Thorne et al., 2018) and AVeriTeC (Schlichtkrull et al., 2023).

Among existing multimodal fact-checking

datasets (Table 1), we select MOCHEG (Yao et al., 2023) as the primary testbed for our study based on three criteria aligned with our research objectives. First, it must explicitly target *fact verification*, thereby excluding datasets designed for related but distinct tasks, such as Factify (Suryavardan et al., 2023), which focuses on out-of-context detection. Second, it must take a textual claim c as input, excluding datasets such as AVeriMaTeC (Cao et al., 2025) that assumes text-image input pairs. Third, it must provide the knowledge source K for the retrieval of both textual and visual evidence. To the best of our knowledge, MOCHEG (Yao et al., 2023) is the only publicly available dataset meeting all these requirements, as ChartCheck provides only chart-caption evidence pairs. To assess the generalizability beyond a single dataset, we use two additional datasets only for testing.

MOCHEG comprises claims with corresponding verdicts, supported by evidence curated by professional fact-checkers. The claims are collected from two widely used fact-checking websites: PolitiFact¹ and Snopes². The dataset includes 11,669 claims for training, 1,490 for validation, and 2,442 for testing. The knowledge source K consists of 3,070,563 sentences and 137,621 images. Each claim is paired with gold textual evidence and, when available, gold visual evidence.

FIN-FACT (Rangapur et al., 2025) is a financial fact-checking dataset of claims and verdicts produced by human experts. We applied a sanitization process to the original dataset by removing claims without publicly accessible image evidence and refining the remaining claims, yielding 2,581 claims with associated multimodal evidence. As the dataset does not provide an external knowledge store, we treat the entire set of textual evidence and images as K .

WebFC is a newly constructed corpus developed in this study, consisting of 621 recently published claims and their corresponding verdicts from PolitiFact, spanning January 2024 to September 2025. Treating web documents as the knowledge source K , we retrieved both textual and visual evidence using the Google Custom Search API. To prevent retrieving documents published after the corresponding articles, which would make the setting unrealistic, we restricted retrieval to sources published before the fact-checking articles. As the

¹<https://www.politifact.com/>

²<https://www.snopes.com/>

dataset does not assume a fixed domain, it reflects a real-world, open-domain fact-checking scenario.

Details of dataset preprocessing and construction are provided in Appendix B.

4 Is Visual Evidence Always Needed?

This section examines the role of visual evidence in verifying textual claims. We first analyze the effects of visual evidence on performance, and then conduct qualitative analyses to identify when visual evidence contributes to verification and under what circumstances it is not necessary.

4.1 Effect of Visual Evidence on Performance

We compared the verification accuracy of VLMs across four input configurations: (1) textual evidence only, (2) textual evidence with *gold* image evidence, (3) textual evidence with retrieved image evidence, and (4) an oracle setting in which the input modality that leads to the correct outcome is provided for each instance.

In all configurations, we assumed that the *gold* textual evidence is provided, as our focus is to assess the impact of visual evidence. Following prior work (Cekinel et al., 2025), configuration (2) used the first image as evidence when multiple images are available. In configuration (3), we used the fine-tuned CLIP provided with the MOCHEG baseline to retrieve the top-1 image from K as visual evidence. In configuration (4), the input combination that yields the correct verdict—when such a configuration exists—was always selected, thereby representing the upper bound of achievable performance according under perfect modality selection. This comparison allows us to examine the impact of visual evidence on fact verification.

Table 2 presents the evaluation results. The first column lists the target VLMs, the second specifies the input configuration, and the third and fourth report verification performance measured by accuracy and macro F1 on the MOCHEG test split. We evaluated two open-weight and two proprietary VLMs. The open-weight models are Qwen2-VL-7B (Wang et al., 2024) and Llama-3.2-11B-Vision (Grattafiori et al., 2024), denoted throughout the paper as **Qwen2-VL** and **Llama-3.2-V**, respectively, while the proprietary models are **GPT-4o** (Hurst et al., 2024) and **Gemini-2.5-Pro** (Comanici et al., 2025). Model checkpoints are provided in Appendix A.

We derived three key observations regarding the

Model	Configuration	Accuracy	F1
Qwen2-VL	(1) Text	0.509	0.419
	(2) + Image (G)	0.506	0.410
	(3) + Image (R)	0.490	0.395
	(4) Oracle	0.547	0.451
Llama-3.2-V	(1) Text	0.496	0.495
	(2) + Image (G)	0.462	0.445
	(3) + Image (R)	0.412	0.393
	(4) Oracle	0.590	0.582
GPT-4o	(1) Text	0.642(.001)	0.612(.000)
	(2) + Image (G)	0.635(.003)	0.596(.003)
	(3) + Image (R)	0.630(.001)	0.597(.001)
	(4) Oracle	0.680(.002)	0.649(.001)
Gemini-2.5-Pro	(1) Text	0.627(.001)	0.599(.001)
	(2) + Image (G)	0.620(.001)	0.570(.001)
	(3) + Image (R)	0.596(.001)	0.557(.001)
	(4) Oracle	0.692(.001)	0.660(.001)

Table 2: Fact-verification performance across different evidence configuration strategies, indicating that visual evidence is not always necessary. Values in parentheses denote standard errors, which are omitted for open-weight models due to deterministic results (G: Gold, R: Retrieved).

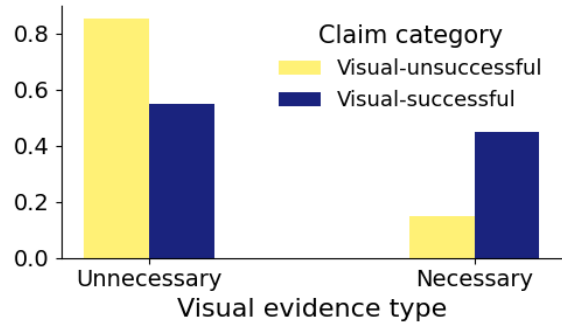


Figure 2: Distribution of visual evidence types across claim categories, illustrating their associations.

role of visual evidence. First, when comparing the text-only setting (configuration 1) with the multimodal setting using gold images (configuration 2), incorporating visual evidence consistently degraded performance across different models. Second, in configuration (3), which uses retrieved image evidence, performance decreased relative to both configurations (1) and (2) across all models. The largest drop was observed for Llama-3.2-V, with decreases of 0.084 in accuracy and 0.102 in F1. This result suggests that imperfect retrieval introduces noisy or misleading context. Third, the oracle setup (configuration 4) consistently outperformed all other settings by optimally incorporating visual evidence. Taken together, these findings support our hypothesis that visual evidence is not always necessary for claim verification. Moreover,

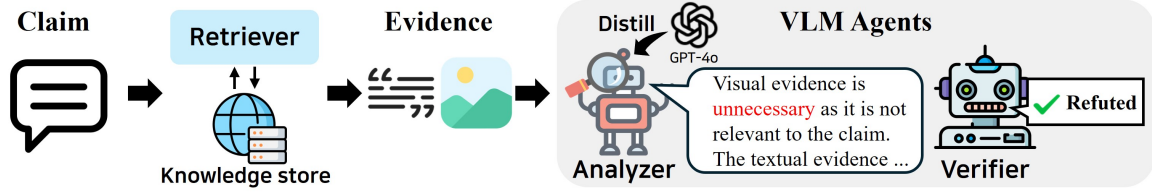


Figure 3: Overall pipeline of AMuFC. Given the retrieved evidence, the two VLM agents—Analyzer and Verifier—are responsible for assessing the necessity of visual evidence and predicting the claim’s veracity, respectively.

the performance upper bound achieved by the oracle setup motivates the need for adaptive use of visual evidence.

4.2 Analyzing the Role of Visual Evidence

Building on the observed impact of visual evidence on verification performance, we conducted an analysis to examine when visual evidence aids verification and when it does not. We first defined two *claim categories* based on whether visual evidence changed GPT-4o’s verification outcome on the validation set, as GPT-4o achieved the best overall performance among the four VLMs evaluated in Table 2. Specifically, we labeled claims as **Visual-successful** (51 claims), where the model predicted the correct verdict only when visual evidence was provided alongside textual evidence, and as **Visual-unsuccessful** (61 claims), where the model verified the claim correctly using textual evidence alone but became incorrect when visual evidence was incorporated. Since our goal is to understand the influence of visual evidence in verification, we excluded claims for which adding images did not affect the text-only model’s prediction.

Next, we manually assessed whether visual evidence is necessary for claim verification. Through an iterative process, we developed an annotation scheme that distinguished two *visual evidence types* by necessity for claim verification. The scheme was reviewed by a journalism Ph.D. expert. We annotated visual evidence as **Unnecessary** (80 pieces) when it was redundant (i.e., it only repeats what the text already conveys) or irrelevant (e.g., it omits the key entity or event mentioned in the claim, or merely shows a portrait of a person without adding contextual value), and as **Necessary** (32 pieces) when it provided unique information required to evaluate the claim’s truthfulness. Two annotators from our institution applied this scheme to the 112 samples, of which the claim categories were labeled, following the annotation guidelines in Table A6. This process achieved high agreement

(Krippendorff’s $\alpha = 0.809$), indicating reliable labeling consistency; Examples are shown in Table A7.

Figure 2 shows how evidence necessity varied across claim categories. We observed a significant association between claim categories and evidence types ($p < 0.001$, chi-square test), indicating that irrelevant or redundant visual evidence is not only unnecessary but can also degrade verification performance. Consistent with the results in Table 2, these findings motivate the adaptive use of visual evidence, as reflected in the proposed method.

5 Methods

5.1 Proposed Method: AMuFC

We propose an agentic fact-checking framework, termed **Adaptive Agentic Multimodal Fact-Checking with Visual Evidence Necessity** (AMuFC). This framework addresses the limitations observed when visual evidence is indiscriminately adopted (Section 4).

As illustrated in Figure 3, AMuFC consists of three components: a Retriever and two collaborating VLM agents. With a focus on open-weight models, our best-performing configuration employs Llama-3.2-V as the Analyzer and Qwen2-VL as the Verifier. The prompts used in this framework are shown in Figure 4.

- *Retriever*: Retrieves textual and visual evidence from the knowledge source K .
- *Analyzer*: Evaluates whether visual evidence is necessary for verifying a claim. It generates natural-language judgments about the necessity of visual evidence given a claim and the retrieved textual and visual evidence. Llama-3.2-V is fine-tuned on claims and gold evidence in the training split using GPT-4o distilled outputs under this prompting scheme. Building on the findings in Section 4, the Analyzer is explicitly instructed to reason about the necessity of visual evidence.

- *Verifier*: Predicts claim veracity as the primary VLM agent. In addition to the claim and retrieved evidence, the Verifier incorporates the Analyzer’s natural-language assessment to adaptively determine whether to use visual evidence during verification.

This agentic design is inspired by the fact-checking practices of human experts (Borel, 2023; Graves and Amazeen, 2019), who assess the relevance and quality of candidate evidence before reaching a verdict. We hypothesize that incorporating the Analyzer’s natural-language assessment is critical for improving overall fact-checking accuracy. This hypothesis is validated through experiments in Section 6.

For the experiments, the Analyzer and Verifier were trained on the training split of MOCHEG, our primary testbed. For model comparison, we adopted the retriever provided with the dataset and used in baseline methods to ensure a fair comparison. Specifically, CLIP (Radford et al., 2021) was used for image retrieval and SBERT (Reimers and Gurevych, 2019) was used for text retrieval. The top-ranked textual and visual evidence was selected via maximum cosine similarity search and passed to subsequent verification stages. For the WebFC case study in Section 6.4, we used the Google Search engine as the Retriever and treated web documents as the knowledge source K .

5.2 Baseline Methods

We compare AMuFC with four existing approaches for multimodal fact verification that incorporate both textual and visual evidence. Our focus is on approaches that assume a fixed knowledge source K , thereby excluding those that do not, such as DEFAME (Braun et al., 2025).

- MOCHEG (Yao et al., 2023): Concatenates the claim, textual evidence, and visual evidence, which are then fed into CLIP to obtain contextual representations. It applies stance detection to generate claim–text and claim–image stance representations, which are averaged and passed to a multimodal fusion classifier for final prediction.
- LVLM4FV (Tahmasebi et al., 2024): Employs a re-ranking strategy that leverages both a LLM and a VLM to retrieve more informative evidence, and uses another VLM to verify the claim.

Your task is to determine whether the provided image evidence is necessary for verifying the given claim or clarifying the accompanying text evidence. Follow these steps:

1. Analyze the claim and the text evidence to understand the context.
2. Assess whether the image provides important information that is not already conveyed by the text.
3. Decide whether the image is necessary for verification and justify your reasoning. Respond only with your analysis.

Claim: {claim}
Image Evidence: {image evidence}
Text Evidence: {text evidence}

(a) Analyzer

Given a claim, your task is to determine the correct verdict based on the provided image evidence and text evidence. Provide a justification for your answer, then choose one of the following verdicts: ‘Supported’, ‘Refuted’, or ‘NEI’ (Not Enough Information).

Claim: {claim}
Image Evidence: {image evidence}
Image Analysis: {analysis generated by Analyzer}
Text Evidence: {text evidence}

(b) Verifier

Figure 4: Prompts used in AMuFC.

- HGTmFC (Pang et al., 2025): Utilizes a Hypergraph Transformer-based model to capture high-order relationships between different modalities of evidence and claims by modeling them as hypergraphs.
- MetaSumPerceiver (Chen et al., 2024): Jointly summarizes the claim, textual evidence, and visual evidence for information extraction, followed by claim verification.

6 Experiments

This section presents the experimental results on the effectiveness of the proposed method. We evaluated fact-verification performance using accuracy (micro F1) and macro F1, which are standard metrics for multi-class classification. Additional experimental details are provided in Appendix A.

6.1 Performance Evaluation

Table 3 shows the fact-verification performance of AMuFC and the baseline methods on the MOCHEG test set. In the *Gold* setting, we provided gold evidence without retrieval, simulating

Method	Gold		Retrieved	
	Accuracy	F1	Accuracy	F1
AMuFC	0.612	0.600	0.546	0.540
MOCHEG	0.520	0.500	0.456	0.438
LVL4M4FV	0.534	0.535	0.451	0.450
HGTMFC	0.541	0.520	0.486	0.468
MetaSumPerceiver	0.556	0.482	0.486	-

Table 3: Verification performance of baseline and proposed methods measured on the test set of MOCHEG, indicating the effectiveness of the proposed framework.

perfect retrieval, whereas the *Retrieved* setting relied on retrieved evidence. We reported baseline scores from their original papers and omitted the macro F1 score for the retrieved setting of MetaSumPerceiver, as its implementation was not publicly available and the reported results were not reproducible.

The results show that AMuFC consistently outperformed existing methods by a substantial margin across different settings and evaluation metrics. Specifically, AMuFC achieved an accuracy of 0.612 and a macro F1 of 0.6 in the gold setting. When retrieved evidence was used, overall performance decreased, but the performance advantage of AMuFC persisted. These results demonstrate the effectiveness of the proposed framework and indicate its ability to adaptively leverage visual evidence across different claims.

6.2 Ablation Analysis

Analyzer-Verifier Integration We evaluated the hypothesis that incorporating the Analyzer’s natural-language assessment into the Verifier’s prompt is critical to the performance gains of AMuFC. To this end, we implemented three alternative strategies. The first, *label-only*, treats the Analyzer as a classifier that generates a visual evidence necessity label without explanations. The second, *pre-filtering*, excludes images predicted as *unnecessary* by the Analyzer from the Verifier’s input, unlike AMuFC, which passes images to the Verifier for all claims. The third, *w/o Analyzer*, discards the Analyzer module entirely. Table 4 reports verification performance for these strategies, all of which employed the fine-tuned Qwen2-VL as the Verifier. The results indicate that the alternative integration strategies yielded lower performance than AMuFC, supporting the hypothesis that the Analyzer’s natural-language assessment is critical to the performance gains of AMuFC and suggest-

Method	Gold		Retrieved	
	Accuracy	F1	Accuracy	F1
AMuFC	0.612	0.600	0.546	0.540
Label-only	0.556	0.530	0.484	0.442
Pre-filtering	0.574	0.551	0.492	0.460
w/o Analyzer	0.563	0.537	0.477	0.435

Table 4: Ablation results on Analyzer-Verifier integration strategies, demonstrating the effectiveness of incorporating the Analyzer’s natural-language assessments, adopted in AMuFC.

Verifier	Analyzer	Gold		Retrieved	
		Accuracy	F1	Accuracy	F1
Qwen2-VL	None	0.563	0.537	0.477	0.435
	Qwen2-VL	0.610	0.598	0.547	0.543
	Llama-3.2-V	0.612	0.600	0.546	0.540
	GPT-4o	0.631	0.620	0.538	0.537
	Gemini-2.5-Pro	0.612	0.606	0.478	0.470
Llama-3.2-V	None	0.491	0.398	0.423	0.314
	Qwen2-VL	0.555	0.455	0.513	0.408
	Llama-3.2-V	0.562	0.461	0.514	0.409
	GPT-4o	0.575	0.473	0.515	0.412
	Gemini-2.5-Pro	0.525	0.445	0.415	0.329

Table 5: Comparison of verification performance with varying VLM choices for the Analyzer and Verifier.

ing that the Verifier does not treat the Analyzer’s assessments as binding but integrates them within its internal reasoning process.

VLM Comparison Table 5 reports verification accuracies under different VLM configurations. For the Verifier, we compared two open-weight VLMs: Qwen2-VL and Llama-3.2-V. For the Analyzer, we additionally evaluated two closed-weight VLMs accessed via API: GPT-4o and Gemini-2.5-Pro. Although the primary focus of this study is on open-weight VLMs, this comparison allows us to assess how effectively the fine-tuned Analyzer in AMuFC performs relative to closed, high-performance models.

For the Verifier, Qwen2-VL consistently performed better than Llama-3.2-V, leading to its selection as the Verifier in our best-performing variant. Relative to the baseline approach without an Analyzer agent, incorporating any Analyzer model consistently improved verification accuracy. For the Analyzer, GPT-4o achieved the best performance in the Gold setting, followed by Gemini-2.5-Pro and Llama-3.2-V. In the Retrieved setting, using open-weight models as the Analyzer yielded better performance than closed models. Based on these results, we adopted Llama-3.2-V as the Analyzer and Qwen2-VL as the Verifier for AMuFC.

True Label	Predicted Label			True Label	Predicted Label		
	Supported	Refuted	NEI		Supported	Refuted	NEI
Supported	457	137	223	Supported	205	377	235
Refuted	70	563	192	Refuted	15	740	70
NEI	220	268	312	NEI	76	504	220

(a) AMuFC (b) Baseline

Figure 5: Distribution of confusion patterns for AMuFC compared with the Verifier-only baseline.

6.3 Error Analysis

To assess the strengths and weaknesses of AMuFC, we compared its error distribution with that of a baseline method that excludes the Analyzer agent. Figure 5 shows the corresponding confusion matrices. Overall, the baseline method overpredicted the *Refuted* label across different classes, whereas AMuFC produced correct predictions more frequently across different classes. The largest gap was observed for the *Supported* class (457 vs. 205). As illustrated by the left example of Table A8, the Analyzer’s natural-language assessment of visual evidence necessity helped the Verifier produce the correct *Supported* verdict. Another notable trend is that AMuFC favored *NEI* predictions relative to the baseline, which improved *NEI* predictions (312 vs. 220). Although this conservative tendency can sometimes lead to incorrect predictions (e.g., the right example in Table A8), it may be preferable in real-world fact-checking settings, where avoiding premature judgment is often important.

6.4 Case Studies

We conducted case studies on two test-only datasets, FIN-FACT and WebFC, to assess the generalizability of the proposed framework across diverse fact-checking scenarios. For AMuFC, we employed Llama-3.2-V as the Analyzer and Qwen2-VL as the Verifier, both of which were fine-tuned on the MOCHEG training split. We compared this with Verifier-only baseline that used all retrieved textual and visual evidence. Table 6 shows that AMuFC outperformed the baseline on both datasets, supporting its generalizability beyond MOCHEG, which served as the primary dataset in earlier experiments. Moreover, the successful integration of a web retriever implies the applicability of AMuFC to real-world fact-checking.

Method	FIN-FACT		WebFC	
	Accuracy	F1	Accuracy	F1
AMuFC	0.491	0.461	0.321	0.236
Verifier-only	0.469	0.384	0.308	0.178

Table 6: Case-study results on test-only datasets, demonstrating AMuFC’s generalizability.

7 Conclusion

This study examines a common assumption in multimodal fact-checking (Tahmasebi et al., 2024; Cekinel et al., 2025): that incorporating visual evidence leads to more accurate fact verification. In-depth analyses with manual annotation reveal that indiscriminate use of multimodal evidence can reduce accuracy, as the necessity of visual evidence varies across claims. This observation aligns with findings from psychology research, indicating that some images are ineffective for misinformation correction (Whitehead et al., 2025). To enable adaptive use of visual evidence, we propose AMuFC, an agentic fact-checking framework with visual evidence necessity, composed of two VLM agents. The Analyzer assesses whether visual evidence is necessary for claim verification, while the Verifier predicts claim veracity by incorporating the Analyzer’s natural-language assessment into its reasoning process. Experimental results show that AMuFC outperforms existing approaches on the MOCHEG dataset, and ablation analyses indicate that integrating the Analyzer’s natural-language assessment of visual evidence necessity is the key factor driving this improvement. Toward real-world fact-checking, future work could extend the agentic framework to multi-agent collaboration, building on recent work that demonstrates its potential for evaluation (Lin et al., 2025), beyond visual evidence necessity by simulating the fact-checking process of human experts (Borel, 2023).

Limitations

This work has several limitations, which opens directions for future research. First, although experiments demonstrated the generalizability of AMuFC, the evaluation is limited to English. Future work could extend this study to broader contexts by developing multilingual benchmarks for multimodal fact-checking. Second, to ensure a fair comparison with existing methods, we used the same retriever, CLIP and SBERT, for the main ex-

periments on MOCHEG (Section 6). To address the limitation, we conducted additional experiments with different retrievers, of which the results are provided in Table A4. Third, the agentic design introduces additional computational costs compared to the single-model baseline (Table A5). Employing smaller models represents a promising direction toward real-world fact-checking (Belcak et al., 2025), ultimately fostering a more trustworthy and responsible information ecosystem.

Ethical Considerations

Two annotators—one graduate and one undergraduate student (both male) from the authors’ institution—were recruited for manual data annotation (Section 4). As fact verification is an objective task, we expect that the annotators’ gender will have a minimal impact on the outcomes. In compliance with local wage regulations, the annotators were compensated at a rate of approximately USD 7 per hour. We constructed WebFC to support the evaluation of multimodal fact-checking methods, using previously fact-checked claims published between January 2024 and September 2025. Since the dataset is derived from publicly available fact-checking verdicts, its use poses minimal privacy concerns. The dataset will be released exclusively for academic purposes, such as benchmarking, through a public software repository (during the review process, available at <https://anonymous.4open.science/r/AMuFC-01C6/> along with code for reproducibility). Its primary purpose is to facilitate the evaluation of fact-checking pipelines without additional training, including zero-shot inference by LLM/VLM models and web search. All prompts and model checkpoints used in this study are provided within the paper. Language editing was conducted using ChatGPT.

References

Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward “the holy grail”: The continued quest to automate fact-checking. In *Computation+ Journalism Symposium*, (September).

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023a. [Reading and reasoning over chart images for evidence-based automated fact-checking](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo,

Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023b. [Multimodal automated fact-checking: A survey](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. [ChartCheck: Explainable fact-checking over real-world chart images](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13921–13937, Bangkok, Thailand. Association for Computational Linguistics.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimitar Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. [A survey on multimodal disinformation detection](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). In *The Twelfth International Conference on Learning Representations*.

Ingeol Baek, Hwan Chang, ByeongJeong Kim, Jimin Lee, and Hwanhee Lee. 2025. [Probing-RAG: Self-probing to guide language models in selective document retrieval](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3287–3304, Albuquerque, New Mexico. Association for Computational Linguistics.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. [Small language models are the future of agentic ai](#). *arXiv preprint arXiv:2506.02153*.

Giscard Biamby, Grace Luo, Trevor Darrell, and Anna Rohrbach. 2022. [Twitter-COMMs: Detecting climate, COVID, and military multimodal misinformation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1530–1549, Seattle, United States. Association for Computational Linguistics.

Christina Boididou, Symeon Papadopoulos, Yiannis Kompatsiaris, Steve Schifferes, and Nic Newman. 2014. [Challenges of computational verification in social multimedia](#). In *Proceedings of the 23rd international conference on world wide web*, pages 743–748.

Brooke Borel. 2023. *The Chicago guide to fact-checking*. University of Chicago Press.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. [DEFAME: Dynamic](#)

708	Evidence-based FAct-checking with Multimodal Experts . In <i>Proceedings of the 42nd International Conference on Machine Learning</i> .	764
709		765
710		
711	Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2025. Averimatec: A dataset for automatic verification of image-text claims with evidence from the web. <i>arXiv preprint arXiv:2505.17978</i> .	766
712		767
713		768
714		769
715		770
716	Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2025. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies . In <i>Proceedings of the 31st International Conference on Computational Linguistics</i> , pages 4622–4633, Abu Dhabi, UAE. Association for Computational Linguistics.	771
717		772
718		773
719		774
720		
721		
722		
723	Ting-Chih Chen, Chia-Wei Tang, and Chris Thomas. 2024. MetaSumPerceiver: Multimodal multi-document evidence summarization for fact-checking . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 8742–8757, Bangkok, Thailand. Association for Computational Linguistics.	775
724		776
725		777
726		778
727		779
728		780
729		781
730	Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. <i>arXiv preprint arXiv:2507.06261</i> .	782
731		783
732		784
733		785
734		786
735		787
736		788
737	Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. <i>Advances in neural information processing systems</i> , 36:10088–10115.	789
738		790
739		791
740		792
741	Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. <i>arXiv preprint arXiv:2407.21783</i> .	793
742		794
743		795
744		796
745		797
746	Lucas Graves and Michelle Amazeen. 2019. Fact-checking as idea and practice in journalism.	798
747		799
748	Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In <i>Proceedings of the 22nd international conference on World Wide Web</i> , pages 729–736.	800
749		801
750		802
751		803
752		804
753		805
754	Vipin Gupta, Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2022. Mmm: an emotion and novelty-aware approach for multilingual multimodal misinformation detection. In <i>Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022</i> , pages 464–477.	806
755		807
756		808
757		809
758		810
759		811
760	Michael Hameleers, Thomas E Powell, Toni GLA Van Der Meer, and Lieke Bos. 2020. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via	812
761		813
762		814
763		815
		816
		817
		818
		819
		820
	social media. <i>Political communication</i> , 37(2):281–301.	
	Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In <i>Proceedings of the 24th acm international on conference on information and knowledge management</i> , pages 1835–1838.	
	Silvan Heller, Luca Rossetto, and Heiko Schuldt. 2018. The ps-battles dataset-an image collection for image manipulation detection. <i>arXiv preprint arXiv:1804.04866</i> .	
	Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In <i>Proceedings of the European conference on computer vision (ECCV)</i> , pages 101–117.	
	Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. <i>arXiv preprint arXiv:2410.21276</i> .	
	Yeonjun In, Sungchul Kim, Ryan A. Rossi, Mehrab Tajim, Tong Yu, Ritwik Sinha, and Chanyoung Park. 2025. Diversify-verify-adapt: Efficient and robust retrieval-augmented ambiguous question answering . In <i>Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 1212–1233, Albuquerque, New Mexico. Association for Computational Linguistics.	
	Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity . In <i>Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.	
	Kiseung Kim and Jay-Yoon Lee. 2024. RE-RAG: Improving open-domain QA performance and interpretability with relevance estimator in retrieval-augmented generation . In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 22149–22161, Miami, Florida, USA. Association for Computational Linguistics.	
	Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In <i>Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)</i> , pages 36–41, Online. Association for Computational Linguistics.	
	Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model .	

821	Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio	In Companion Proceedings of the ACM on Web Con-	878
822	Petroni, Vladimir Karpukhin, Naman Goyal, Hein-	ference 2025, pages 785–788.	879
823	rich Küttler, Mike Lewis, Wen-tau Yih, Tim Rock-		
824	täschel, and 1 others. 2020. Retrieval-augmented gen-	Nils Reimers and Iryna Gurevych. 2019. Sentence-	880
825	eration for knowledge-intensive nlp tasks. <i>Advances</i>	BERT: Sentence embeddings using Siamese BERT-	881
826	<i>in neural information processing systems</i> , 33:9459–	networks . In <i>Proceedings of the 2019 Conference on</i>	882
827	9474.	<i>Empirical Methods in Natural Language Processing</i>	883
		<i>and the 9th International Joint Conference on Natu-</i>	884
828	Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan	<i>ral Language Processing (EMNLP-IJCNLP)</i> , pages	885
829	Zhang, Jing Ma, See-Kiong Ng, and Tat-Seng Chua.	3982–3992, Hong Kong, China. Association for Com-	886
830	2025. FACT-AUDIT: An adaptive multi-agent frame-	putational Linguistics.	887
831	work for dynamic fact-checking evaluation of large		
832	language models . In <i>Proceedings of the 63rd Annual</i>	Michael Schlichtkrull, Zhijiang Guo, and Andreas Vla-	888
833	<i>Meeting of the Association for Computational Lin-</i>	chos. 2023. Averitec: A dataset for real-world claim	889
834	<i>guistics (Volume 1: Long Papers)</i> , pages 360–381,	verification with evidence from the web . In <i>Ad-</i>	890
835	Vienna, Austria. Association for Computational Lin-	<i>vances in Neural Information Processing Systems</i> ,	891
836	guistics.	volume 36, pages 65128–65167. Curran Associates,	892
		Inc.	893
837	Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021.	Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting	894
838	NewsCLIPPings: Automatic Generation of Out-of-	and grounding multi-modal media manipulation. In	895
839	Context Multimodal Media . In <i>Proceedings of the</i>	<i>Proceedings of the IEEE/CVF Conference on Com-</i>	896
840	<i>2021 Conference on Empirical Methods in Natural</i>	<i>puter Vision and Pattern Recognition</i> , pages 6904–	897
841	<i>Language Processing</i> , pages 6801–6817, Online and	6913.	898
842	Punta Cana, Dominican Republic. Association for		
843	Computational Linguistics.	Freda Shi, Xinyun Chen, Kanishka Misra, Nathan	899
		Scales, David Dohan, Ed H Chi, Nathanael Schärli,	900
844	Kai Nakamura, Sharon Levy, and William Yang Wang.	and Denny Zhou. 2023. Large language models can	901
845	2020. Fakeddit: A new multimodal benchmark	be easily distracted by irrelevant context. In <i>Inter-</i>	902
846	dataset for fine-grained fake news detection . In <i>Pro-</i>	<i>national Conference on Machine Learning</i> , pages	903
847	<i>ceedings of the Twelfth Language Resources and</i>	31210–31227. PMLR.	904
848	<i>Evaluation Conference</i> , pages 6149–6157, Marseille,		
849	France. European Language Resources Association.	Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam	905
		Kumaraguru. 2022. Factdrill: A data repository of	906
850	Hui Pang, Chaozhuo Li, Litian Zhang, Senzhang Wang,	fact-checked social media content to study fake news	907
851	and Xi Zhang. 2025. Beyond text: Fine-grained	incidents in india. In <i>Proceedings of the interna-</i>	908
852	multi-modal fact verification with hypergraph trans-	<i>tional AAAI conference on web and social media</i> ,	909
853	formers . <i>Proceedings of the AAAI Conference on</i>	volume 16, pages 1322–1331.	910
854	<i>Artificial Intelligence</i> , 39(6):6389–6397.		
		S Suryavardan, Shreyash Mishra, Parth Patwa, Megha	911
855	Stefanos-Iordanis Papadopoulos, Christos Koutlis,	Chakraborty, Anku Rani, Aishwarya Reganti, Aman	912
856	Symeon Papadopoulos, and Panagiotis C Petranton-	Chadha, Amitava Das, Amit Sheth, Manoj Chin-	913
857	akis. 2024. Verite: a robust benchmark for multi-	nakotla, and 1 others. 2023. Factify 2: A multimodal	914
858	modal misinformation detection accounting for uni-	fake news and satire news dataset. <i>arXiv preprint</i>	915
859	modal bias. <i>International Journal of Multimedia</i>	<i>arXiv:2304.03897</i> .	916
860	<i>Information Retrieval</i> , 13(1):4.		
		Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ew-	917
861	Tanmay Parekh, Pradyot Prakash, Alexander Radovic,	erth. 2024. Multimodal misinformation detection	918
862	Akshay Shekher, and Denis Savenkov. 2025. Dy-	using large vision-language models . In <i>Proceedings</i>	919
863	namic strategy planning for efficient question answer-	<i>of the 33rd ACM International Conference on In-</i>	920
864	ing with large language models . In <i>Findings of the</i>	<i>formation and Knowledge Management, CIKM '24</i> ,	921
865	<i>Association for Computational Linguistics: NAACL</i>	page 2189–2199, New York, NY, USA. Association	922
866	2025, pages 6038–6059, Albuquerque, New Mexico.	for Computing Machinery.	923
867	Association for Computational Linguistics.		
		Reuben Tan, Bryan Plummer, and Kate Saenko.	924
868	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya	2020. Detecting cross-modal inconsistency to de-	925
869	Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-	fend against neural fake news . In <i>Proceedings of the</i>	926
870	try, Amanda Askell, Pamela Mishkin, Jack Clark, and	<i>2020 Conference on Empirical Methods in Natural</i>	927
871	1 others. 2021. Learning transferable visual models	<i>Language Processing (EMNLP)</i> , pages 2081–2106,	928
872	from natural language supervision. In <i>International</i>	Online. Association for Computational Linguistics.	929
873	<i>conference on machine learning</i> , pages 8748–8763.		
874	PmLR.	Xiaqiang Tang, Qiang Gao, Jian Li, Nan Du, Qi Li, and	930
		Sihong Xie. 2025. MBA-RAG: a bandit approach	931
875	Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu.	for adaptive retrieval-augmented generation through	932
876	2025. Fin-fact: A benchmark dataset for multimodal	question complexity . In <i>Proceedings of the 31st Inter-</i>	933
877	financial fact-checking and explanation generation.	<i>national Conference on Computational Linguistics</i> ,	934

935	pages 3248–3254, Abu Dhabi, UAE. Association for Computational Linguistics.		
936			
937	James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In <i>NAACL-HLT</i> .		
938			
939			
940			
941	Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. <i>arXiv preprint arXiv:2502.14786</i> .		
942			
943			
944			
945			
946			
947			
948	Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. <i>arXiv preprint arXiv:2409.12191</i> .		
949			
950			
951			
952			
953	Peter S Whitehead, Charles P Davis, Joyce S Park, and Elizabeth J Marsh. 2025. Pictures are not always worth a thousand words: Nonprobative pictures did not increase the effectiveness of misinformation corrections. <i>Journal of Applied Research in Memory and Cognition</i> .		
954			
955			
956			
957			
958			
959	Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. FIRE: Fact-checking with iterative retrieval and verification. In <i>Findings of the Association for Computational Linguistics: NAACL 2025</i> , pages 2901–2914, Albuquerque, New Mexico. Association for Computational Linguistics.		
960			
961			
962			
963			
964			
965			
966	Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In <i>Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’23</i> , page 2733–2743, New York, NY, USA. Association for Computing Machinery.		
967			
968			
969			
970			
971			
972			
973			
974	Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In <i>The Twelfth International Conference on Learning Representations</i> .		
975			
976			
977			
978			
979	Ziyu Zhuang. 2025. Pass-fc: Progressive and adaptive search scheme for fact checking of comprehensive claims. <i>arXiv preprint arXiv:2504.09866</i> .		
980			
981			
982	Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.		
983			
984			
985			
986			
987			
988			
989			
		Appendix	990
		A Experimental Configurations	991
		For experiments on MOCHEG, we used the fine-tuned CLIP and SBERT retrievers provided with the baseline to ensure a fair comparison by controlling for the effects of the Retriever. We accessed GPT-4o and Gemini-2.5-Pro via API, while Qwen2-VL-7B and Llama-3.2-11B-Vision were used through pretrained checkpoints. Fine-tuning was performed with QLoRA (Detmiers et al., 2023), using a rank of 128, an alpha value of 256, and a dropout rate of 0.05. We employed the AdamW optimizer with a cosine learning rate scheduler, a base learning rate of 2×10^{-5} , a warmup ratio of 0.03, and a batch size of 32. Training was conducted for two epochs. All experiments were run on a single node equipped with eight NVIDIA H100 80GB GPUs, which were used for both training and inference.	992
			993
			994
			995
			996
			997
			998
			999
			1000
			1001
			1002
			1003
			1004
			1005
			1006
			1007
			1008
		We fixed the inference hyperparameters to ensure reproducibility. Specifically, we used greedy decoding for both Qwen2-VL-7B and Llama-3.2-11B-Vision; accordingly, all results from the open-weight models are deterministic. GPT-4o and Gemini-2.5-Pro were evaluated with a temperature of 0.0, with a thinking budget of 128 for Gemini-2.5-Pro. For API-based results, we reported the mean and standard error of model performance over three runs.	1009
			1010
			1011
			1012
			1013
			1014
			1015
			1016
			1017
			1018
		The model IDs of the VLMs used in our experiments are provided below.	1019
			1020
		• GPT-4o (Hurst et al., 2024): gpt-4o-2024-11-20	1021
			1022
		• Gemini-2.5-Pro (Comanici et al., 2025): gemini-2.5-pro	1023
			1024
		• Qwen2-VL-7B (Wang et al., 2024): https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct	1025
			1026
			1027
		• Llama-3.2-11B-Vision (Grattafiori et al., 2024): https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct	1028
			1029
			1030
			1031
		The model IDs of the retrievers used in our experiments are provided below.	1032
			1033
		• CLIP (Radford et al., 2021): sentence-transformers/clip-ViT-B-32	1034
			1035

- **SigLIP (Tschannen et al., 2025):** <https://huggingface.co/google/siglip-so400m-patch14-384>
- **SBERT (Reimers and Gurevych, 2019):** <https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1>
- **mxbai (Lee et al., 2024):** <https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1>

You are a fact-checking assistant. Your task is to determine whether the given claim is complete in intention and clearly stated. If the claim is vague or incomplete (e.g., a keyword like “Google PhoneBook”), refine it into a clear and complete sentence using the provided evidence, justification, and label. If the claim is already clear and complete, return “not needed”.

Claim: {claim}
 Evidence: {evidence}
 Justification: {justification}
 Label: {label}

Return **ONLY** the refined claim or “not needed”.

B Dataset Details

This section describes the dataset sanitization and construction processes.

B.1 FIN-FACT

The original FIN-FACT (Rangapur et al., 2025) dataset contains 3,369 claims, comprising 840 text-only claims and 2,529 claims with multimodal evidence. Among 4,285 referenced image URLs, we successfully downloaded 3,027 images. We then retained 1,741 multimodal claims for which more than one referenced image was successfully retrieved. Combined with the text-only portion, the resulting dataset contains 2,481 claims.

We observed that a substantial portion of FIN-FACT claims are incomplete phrases or keywords rather than self-contained, verifiable statements. For example, “*Google PhoneBook*” or “*Barack Obama’s Net Worth.*” To address this issue, we refined vague and incomplete claims using Gemini-2.5-Pro by providing the original claim along with its evidence, justification, and label. The claim-refinement prompt is shown in Table A1, and examples of refined claims are presented in Table A2.

Because FIN-FACT does not provide an external knowledge source, we constructed a knowledge source K consisting of 39,850 sentences extracted from all textual evidence passages in the original dataset, together with the 3,027 downloaded images. We also transformed labels from *True* and *False* to *Supported* and *Refuted*, respectively, to align with the MOCHEG label space. We used the refined FIN-FACT dataset for retrieval-based evaluation in the experiments.

B.2 WebFC

Using a web-based retriever (Google Custom Search API), we queried the web for each claim to retrieve the top 10 documents and the top 1 image

Table A1: Prompt used in claim refinement

as textual and visual evidence. We excluded claims with more than eight of these URLs that failed due to parsing errors during top 10 URL extraction. As the retrieved documents varied widely in length and were often long, we summarized each of the documents using GPT-4o-mini (gpt-4o-mini-2024-07-18) with the prompt in Table A3. To prevent retrieving documents published after the corresponding articles, which would make the setting unrealistic, we restricted retrieval to sources published before the fact-checking articles.

C Supplementary Results

C.1 Effects of Varying Retrievers

We conducted an ablation study to assess the effectiveness of AMUFC across varying retrievers. For text retrievers, we consider the pretrained SBERT, the fine-tuned SBERT, and mxbai. For image retrievers, we compare the pretrained CLIP, the fine-tuned CLIP, and SigLIP. As shown in Table A4, the effectiveness of AMUFC persists across different retriever configurations.

C.2 Accuracy–Efficiency Trade-offs

While the proposed agentic framework focuses on improving verification accuracy, this design choice introduces additional computational overhead. Specifically, compared to a standard retrieval-based fact-verification pipeline, it incurs extra inference costs due to the Analyzer’s assessment of the necessity of visual evidence. We provide a detailed analysis of these trade-offs in Table A5. We computed inference time on the test dataset with gold evidence using a single-batch request and a basic inference pipeline implemented with Hugging Face

Type	Claim ID 28	Claim ID 35
Claim	Google PhoneBook	Barack Obama’s Net Worth
Refined Claim	Entering a phone number into the Google search engine can produce a home address and a map with directions to that address.	Barack Obama’s net worth increased over \$10 million from 2008 to 2012.

Table A2: Examples of claim refinement in FIN-FACT.

Your task is to read the following document carefully and summarize it into a single, coherent paragraph. Focus on capturing the main ideas and essential details without adding new information or personal opinions.
Document:
{document}

Table A3: Prompt used for document summarization

Method		Verifier-only		AMuFC	
Text	Image	Acc.	F1	Acc.	F1
SBERT	CLIP	0.475	0.434	0.530	0.527
Mxbai	SigLIP	0.502	0.476	0.521	0.523
SBERT(FT)	CLIP(FT)	0.477	0.435	0.546	0.540

Table A4: Ablation results on Analyzer–Verifier integration strategies across different retriever methods. FT denotes fine-tuned.

Transformers, evaluated on a single NVIDIA H100 GPU.

C.3 Qualitative Examples

Table A7 presents two examples with different visual evidence necessity labels, as defined in Section 4. Table A8 shows successful and failed predictions of the proposed method, along with the corresponding Analyzer outputs.

Method	Accuracy	F1	Inference time
AMuFC	0.612	0.600	1.531
Analyzer	-	-	1.420
Verifier	-	-	0.111
Baseline	0.563	0.537	0.112

Table A5: Accuracy-efficiency trade-offs, with inference time averaged per sample (s/it).

Given a claim and its associated textual and visual evidence, determine whether the visual evidence is **necessary** for evaluating the claim.

Inputs:

Claim: {claim}

Textual Evidence: {textual evidence}

Visual Evidence: {visual evidence}

Necessary visual evidence refers to visual content that provides novel, complementary, or clarifying information beyond what is conveyed in the textual evidence, and that meaningfully contributes to interpreting or supporting the claim.

Unnecessary visual evidence refers to visual content that is irrelevant, only depicts entities, is loosely related, or is redundant with the textual evidence.

Select one:

- Necessary
- Unnecessary

Table A6: Annotation guidelines for determining whether visual evidence is necessary.



Type	Unnecessary	Necessary
Claim	Marquette University threatened to rescind student's admission over pro-Trump TikTok video.	In an episode of 'The Simpsons,' Mayor Quimby says he is canceling a trip to the Bahamas while he's in the Bahamas, because of an ongoing epidemic.
Visual Evidence		
Textual Evidence	<ul style="list-style-type: none"> • The pro-Trump post was not at issue. Marquette and the people complaining to the university were examining comments on topics like sexuality and immigration in other social media posts. • The student clearly felt her status was in doubt, but she has stopped short of calling it a threat. The school says no threat was made. • The “threat” element is tricky to pin down, since that claim stems from a private conversation between the student and Marquette admissions staffers. 	After initially saying he didn't do anything wrong, Steve Adler, the mayor of Austin, Texas, says he now realizes he 'set a bad example' by traveling to Cabo San Lucas, Mexico, for vacation last month. An Austin American-Statesman story revealed Wednesday that Adler attended an in-person wedding for his daughter in early November and then flew with others to Cabo for a weeklong vacation. At the same time, Adler was encouraging people to stay home to avoid contracting or spreading COVID-19.

Table A7: Examples for visual evidence types according to their necessity for claim verification.



Type	Successful	Failed
Claim	A Boeing B-17E bomber from World War II was found in the jungle with coffee still in thermoses.	Students no longer say the Pledge of Allegiance in schools.
Visual Evidence		
Textual Evidence	<ul style="list-style-type: none"> • They found the rear door of the airplane, two aluminum receptacles and a third which contained coffee residue. • Both the impactor and the condensate were stored and transported to the laboratory at 2-8 °C. • Print LONG BEACH, Calif. - A B-17 bomber that lay in a New Guinea swamp for decades after being forced down during a World War II combat mission has been returned to the United States after years of salvage efforts. • As food stores were running low at the airplane, Zerbino and Fito went searching for cadavers and although they found one, they were too tired to bring it back to the fuselage. • Ignited by fire, ammunition was exploding inside the cabin. 	<ul style="list-style-type: none"> • Students still say the Pledge of Allegiance in classrooms, though they don't have to. • Virtual learning has presented new challenge for how schools say the Pledge of Allegiance.
Analysis	The image evidence is necessary for verifying the claim. While the text evidence provides detailed information about the discovery of the Boeing B-17E bomber in the jungle, including the presence of coffee residue and thermoses, the image directly shows the plane in its jungle location, confirming the physical context of the claim. The visual evidence complements the text by providing a tangible representation of the scene described, which is essential for verifying the authenticity of the claim.	The image evidence is not necessary for verifying the claim. The text evidence already provides sufficient information to address the claim that students no longer say the Pledge of Allegiance in schools. It clarifies that students still say the Pledge of Allegiance in classrooms, though it is not mandatory, and that it is recited at every school. The image of students raising their hands does not add any new or essential information to verify or refute the claim.
Prediction	Ground truth: <i>Supported</i> Baseline: <i>Refuted</i> ✗ AMuFC: <i>Supported</i> ✓	Ground truth: <i>Refuted</i> Baseline: <i>Refuted</i> ✓ AMuFC: <i>NEI</i> ✗

Table A8: Examples for successful and failed predictions of AMuFC.