

MultiReflect: Multimodal Self-Reflective RAG-based Automated Fact-Checking

Uku Kangur¹ Krish Agrawal² Yashashvi Singh³ Ahmed Sabir¹ Rajesh Sharma^{1,4}

¹University of Tartu, Institute of Computer Science, Estonia, ²Indian Institute of Technology Indore ³Indian Institute of Information Technology Dharwad, ⁴Plaksha University, India

Abstract

In this work, we introduce MultiReflect, a novel multimodal self-reflective Retrieval Augmented Generation (RAG)-based automated fact-checking pipeline. MultiReflect is designed to address the challenges of rapidly outdated information, limitations in human query capabilities, and expert knowledge barriers in fact-checking. Our proposed pipeline leverages the latest advancements in Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) to enhance fact verification across text and images. Specifically, by integrating multimodal data processing with RAG-based evidence reflection, our system improves the accuracy of fact-checking by utilizing internet-sourced verification. We evaluate our results on the VERITE benchmarks and using several multimodal LLMs, outperforming baselines in binary classification.¹

1 Introduction

Information plurality, particularly on the internet, presents both opportunities and challenges in identifying accurate and up-to-date information. Given the increasing reliance on online platforms for news consumption, learning, and interaction (Eurostat, 2022), developing effective mechanisms to distinguish between truthful and false information has become more critical. However, the increase of coordinated misinformation movements by spam bots, and other forms of informational chaos have significantly complicated this process. Therefore, more advanced and systematic approaches are required to evaluate and verify (*fact-check*) the credibility of information sources.

With the emergence of Large Language Models (LLMs), which can understand and learn from billions of texts, automated fact-checking has grown in popularity as an alternative to traditional manual

methods (Guo et al., 2022). While LLMs are state-of-the-art tools for various language understanding and reasoning tasks, they still face several limitations, such as hallucinations, overconfidence, and bias (Xu et al., 2024b; Li et al., 2024). To address these issues, several studies have employed Retrieval Augmented Generation (RAG) techniques, which allow the model to check based on externally verified information (Lewis et al., 2021; Gao et al., 2024).

Language, however, is only part of the challenge when it comes to fact-checking information on the internet. Online information is presented in various forms, including text, images, video, and sound. As a result, fact-checking also requires the retrieval and reasoning of information across multiple modalities (Akhtar et al., 2023b; Martin et al., 2025). More recently, several state-of-the-art models, such as GPT-4V (OpenAI, 2023), GPT-4o (OpenAI, 2024), DeepSeek-VL2 (Wu et al., 2024) and Claude 3 (Anthropic, 2024), have made reason across multimodal data possible. These rapid advancements highlight the need for multimodal fact-checking, which has grown with the increased prevalence of complex information that spans various data types: text, image, video, and audio. Systems like COSMOS (Aneja et al., 2021), TwitterCOMMs (Biamby et al., 2022), EXMULF (Amri et al., 2022), ChartBERT (Akhtar et al., 2023a), RED-DOT and (Papadopoulos et al., 2024a) have made significant progress in tackling the challenges posed by multimodal data.

However, despite their successes, these systems have not fully taken advantage of RAG, a crucial component for dynamic and context-aware evidence retrieval in the multimodal setting. To address this gap, we introduce **MultiReflect**, illustrated in Figure 1, which integrates multimodal fact-checking (image + text) with a self-reflective RAG framework. Our system is designed to dynamically retrieve, evaluate, and rank supporting

¹<https://github.com/ukangur/MultiReflect>

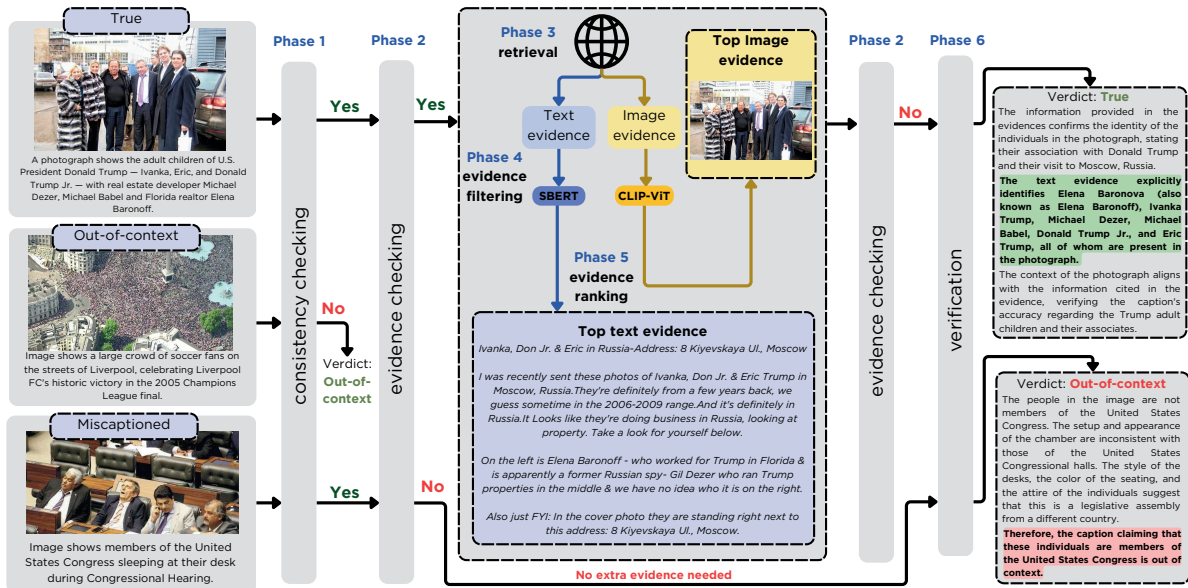


Figure 1: **MultiReflect** system overview. The proposed pipeline contains six phases: (1) consistency checking, (2) evidence checking, (3) retrieval, (4) evidence filtering, (5) evidence ranking and (6) verification. The colors indicate using both modalities in **gray/black**, or only image data in **yellow**, or only text data in **blue**.

evidence, improving reasoning capabilities and accuracy of multimodal fact verification.

We summarize our contributions as follows:

- We propose a novel pipeline **MultiReflect**, a multimodal self-reflective RAG-based automated fact-checking pipeline.
- The novelty of the approach is in combining RAG-based evidence reflection with multimodal fact-checking.
- Our MultiReflect system achieves state-of-the-art results in binary classification in the Multimodal fact-checking VERITE benchmark.

2 Data

For our experiments, we utilize the VERITE dataset, a multimodal *fack-checking* benchmark dataset (Papadopoulos et al., 2024b). The dataset contains 892 different image-text pairs with the labels "True" (302), "Miscaptioned" (302), and "Out-of-context" (288). The dataset incorporates a wide range of real-world data while specifically excluding "asymmetric multimodal misinformation" (Asymmetric-MM), which refers to scenarios where one form of modality significantly amplifies misinformation while others have minimal impact. Also, the data implements "modality balancing," ensuring that each image and caption are represented twice in the dataset: once within truthful contexts and once within misleading pairs.

3 Proposed Method: MultiReflect

In this section, we introduce our proposed six-phase pipeline: (1) consistency checking, (2) evidence checking, (3) retrieval, (4) evidence filtering, (5) evidence ranking, and (6) verification.

3.1 Phase 1: Consistency checking

In this phase, we filter inputs by checking the alignment between the image and the caption. If inconsistent, the post is marked as OUT-OF-CONTEXT (as shown in Figure 1 with the second example). Three strategies are evaluated to determine the best method for consistency checking. The best strategy is used in the pipeline for the consistency checking phase.

Image-to-Text consistency: Using CLIP Large-336 (Radford et al., 2021), cosine similarity between image-caption embeddings determines consistency based on the best threshold of 0.28 estimated via grid search within the range [0.10 - 0.39].

Text-to-Text consistency: BLIP-2 (Li et al., 2023) generates descriptions for images, compared to captions using cosine similarity via SBERT (Reimers and Gurevych, 2019), with best threshold 0.10 for all model (BLIP-2_{2.7B}, BLIP-2_{6.7B} and BLIP-2 FLAN) estimated similar to Image-to-Text method.

Multimodal consistency: Since multimodal LLMs can comprehend and perform reasoning on both text and images, we use the image-caption pairs to evaluate their consistency. We adopt

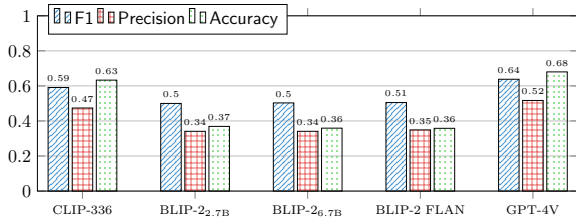


Figure 2: Performance Metrics of Different Consistency Checking Strategies (Phase 1). We rely on three validation methods: Image-to-Text consistency using CLIP, Text-to-Text consistency using BLIP (via SBERT), and Multimodal consistency using GPT-4V model.

a prompt-based approach wherein each image-caption pair is evaluated to ascertain the alignment between the provided text and the associated image. Specifically, the prompt instructs the model to assign a binary score $[0,1]$ whether the caption accurately describes the depicted image. We evaluate multimodal consistency using GPT-4V (OpenAI, 2023). Figure 2 shows that the LLM multimodal consistency strategy has the highest F1-score, therefore, we adopt this strategy in our pipeline.

3.2 Phase 2: Evidence checking

The aim of this phase is for the model to evaluate if extra evidence is needed, inspired by the work of (Asai et al., 2023). We do this to differentiate between information that contains changing or unchanging events. For example, political events may require external evidence for up-to-date information, while physics or nature-related statements, like the world being round, generally remain consistent over time. Additionally, this phase allows the model to be more dynamic when evaluating how much information is needed to fact-check something. We can see this phase in Figure 1, where the first example requires evidence and the third example does not.

We employ the use of multimodal LLM (e.g. GPT-4o-mini) for this task as it allows the model to evaluate the post and the evidence text and image data at the same time. This phase can occur several times during fact-checking one post, as the model can ask for additional evidence several times. To keep this process more efficient, in the evidence retrieval phase, we collect more evidence the first time around and then only provide additional evidence if it is asked in the next iterations of phase 2.

3.3 Phase 3: Evidence retrieval

In this phase, we retrieve both textual and visual evidence required to fact-check the original input post. We collect the evidence for both modalities using at least 3 sources to lower the chances of bias brought by single-source dependency. In addition, we collect all evidence straight from the internet without using any static databases. We do this to ensure the most up-to-date information. We explain the full procedure of evidence collection in the following subsections, as shown in Figure 1 with the top (True) example.

Textual evidence: We retrieve textual evidence from three sources - Wikipedia, Google search, and Bing search. For Wikipedia (Wikimedia, 2024), we search for the top 10 articles. For Google, we use the Google Custom Search API (Google, 2024) to get the top 10 Google search results and collect their textual data. We also use the Google Cloud Vision API (Cloud, 2024) to collect textual information from pages that include a fully or partially matching reverse image search result with our original multimodal post. For Bing search, we use the Bing Web Search API (Microsoft, 2024b) to get the top 10 Bing search results and collect the textual data from each of them. We additionally use the Bing Visual Search API v7 (Microsoft, 2024a) to collect textual information from pages that include a matching image search result with our multimodal post.

Visual evidence: We retrieve visual evidence from three sources - Wikimedia Commons, Google Image Search, and Bing Image Search. For Wikimedia Commons, we use the Wikimedia Commons API (Wikimedia, 2024) to retrieve the top 10 images by querying for each entity from the textual caption of the original post. For Google Image Search, we use Google Custom Search (Google, 2024) to get the top 10 regular image search results by querying all the entities from the textual caption. With Bing Image Search, we use the Bing Image Search API v7 (Microsoft, 2024a) to get the top 10 regular image search results by querying all the entities from the textual caption.

3.4 Phase 4: Evidence filtering

In this phase, we filter the retrieved evidence based on consistency with the original post data to ensure we do not rank unrelated evidence (as shown in Figure 1). The differences in filtering for textual and visual evidence are introduced as follows:

Textual evidence: With textual evidence, we first split each piece of evidence into paragraphs or if paragraphs are not given, then into sentence chunks of 250 words maximum. We use SBERT (Reimers and Gurevych, 2019) to find the top 3 most semantically similar paragraphs to the original post caption from each online source (*i.e.* Wikipedia, Google Search, Google Inverse Search, Bing Search, Bing Visual Search). Then, we extract the top matching paragraph from each textual evidence and dismiss all other paragraphs. By filtering irrelevant details, we retain only text relevant to domain-specific fact-checking. For example, focusing on Biden’s political decisions while excluding information about his private life events.

Visual evidence: With visual evidence, we embed the images with CLIP Large-336 (Radford et al., 2021) and then use cosine similarity to filter out irrelevant images to the original post and find the top 3 images from each source (*i.e.* Wikimedia Commons, Google Image Search and Bing Image Search).

3.5 Phase 5: Evidence ranking

We use this phase to evaluate the quality of the given evidence based on five attributes: (1) **Authority**, (2) **Timeliness**, (3) **Relevancy**, (4) **Support** and (5) **Usefulness**. We compute a unified score to rank the evidence based on these attributes. After this we extract the top ranking text evidence and top ranking image evidence (as shown in Figure 1 with the first example). We keep the other ranking scores in case the pipeline requires additional evidence.

Authority This attribute captures how authoritative is the source of the evidence. We check the authority on how factual, biased, and reliable the sources are. For example, if a source contains factual content, which is neutral and is also reliable, then it is considered authoritative. To label the sources with these attributes in mind, we use the source bias dataset as introduced by Kangur et al. (2024). This dataset provides aggregated factuality, bias and reliability annotations of the top 500 sources used in X Community Notes (Community Notes, 2024) using pre-defined labels from three trusted media monitoring institutions: Media bias fact-check², Allsides³, and Adfontes⁴. As these labels are ordinal (*i.e.* they can be ordered), we

²mediabiasfactcheck.com

³allsides.com

⁴adfontes.com

transform the labels into predefined scores ranging from 0 to 1, except for the factuality score, which is calculated on a scale from -1 to 1, to additionally penalize unfactual sources. The authority score for evidence is calculated as the sum of the factuality, bias and reliability scores as shown:

$$S_{\text{Authority}} = A_{\text{Factuality}} + A_{\text{Bias}} + A_{\text{Reliability}}$$

$$A_{\text{Factuality}} = \begin{cases} 1.0 & \text{if rated } \textit{Very High Factuality} \\ 0.66 & \text{if rated } \textit{High Factuality} \\ 0.33 & \text{if rated } \textit{Mostly Factual} \\ 0.0 & \text{if rated } \textit{Mixed Factuality} \\ -0.33 & \text{if rated } \textit{Low Factuality} \\ -0.66 & \text{if rated } \textit{Very Low Factuality} \\ -1.0 & \text{if rated } \textit{Satire} \\ 0.0 & \text{otherwise.} \end{cases}$$

$$A_{\text{Bias}} = \begin{cases} 0.0 & \text{if rated as } \textit{Left} \text{ or } \textit{Right} \\ 0.5 & \text{if rated as } \textit{Left-Center} \text{ or } \textit{Right-Center} \\ 1.0 & \text{if rated as } \textit{Center} \\ 0.0 & \text{otherwise} \end{cases}$$

$$A_{\text{Reliability}} = \begin{cases} 1.0 & \text{if rated as } \textit{Reliable} \\ 0.5 & \text{if rated as } \textit{Generally Reliable} \\ 0.0 & \text{if rated as } \textit{Mixed Reliability} \\ 0.0 & \text{otherwise} \end{cases}$$

Relevancy evaluates how well the evidence pertains to the multimodal post. The goal is to assess if the evidence is relevant to the factual accuracy of the image or caption. We use a multimodal LLM (*e.g.* GPT-4o-mini) to label evidence as relevant ($S_{\text{Relevancy}} = 1$) or irrelevant ($S_{\text{Relevancy}} = 0$).

Support evaluates how well the evidence backs the claims in the post. We use a multimodal LLM (*e.g.* GPT-4o-mini) to assess the factual accuracy of the input text and image, by examining their alignment with the evidence based solely on the provided information. An entailment scale is used to assign scores based on the degree of support:

$$S_{\text{Support}} = \begin{cases} 1 & \text{if Fully Supported} \\ 0.5 & \text{if Partially Supported} \\ 0 & \text{if No Support/Contradictory} \end{cases}$$

Usefulness measures how informative and relevant the evidence is for accepting or rejecting the claim in the post. We use a multimodal LLM (*e.g.* GPT-4o-mini) to assess how well the evidence helps determine the factuality of the input image and caption. A 5-point scale is used to score the evidence, with utility scores mapped to numeric values as follows:

$$S_{\text{Usefulness}} = \begin{cases} +1 & \text{if score = 5 (Highly informative)} \\ +0.5 & \text{if score = 4 (Mostly sufficient)} \\ 0 & \text{if score = 3 (Adequate)} \\ -0.5 & \text{if score = 2 (Limited)} \\ -1 & \text{if score = 1 (Irrelevant)} \end{cases}$$

Timeliness evaluates how recently the information in the evidence is provided. Evidence E is considered timely if its date $t(E) < 2$ years, and it has a positive score in at least one of **Relevancy**, **Support**, or **Usefulness**. This ensures that only relevant and meaningful recent evidence is prioritized, avoiding the ranking of irrelevant but recent content. The score is assigned as follows:

$$S_{\text{Timeliness}} = \begin{cases} 1 & \text{if } t(E) < 2 \text{ years and} \\ & S_{\text{Relevancy}} + S_{\text{Support}} + \\ & S_{\text{Usefulness}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Combined Evidence Score: The overall evidence score is calculated as the sum of all of the five attributes. Based on this score, we extract the top ranking (highest scored) image and textual evidence. These are passed into the evidence checking (phase 2) and verification (phase 6) phases.

However, our human evaluation showed that **Timeliness** and **Authority** are hard to discern from visual evidence alone due to potential reuploads that may not reflect the original context. Therefore, we use all five attributes to rank textual evidence, but only **Relevancy**, **Support**, and **Usefulness** for visual evidence.

3.6 Phase 6: Verification

This phase verifies whether the original post is FALSE, OUT-OF-CONTEXT, or TRUE. For verification, we prompt a multimodal LLM model (*e.g.* GPT-4o-mini) to assess the factual accuracy of the input image and caption using the provided evidence, labeling the output as OUT-OF-CONTEXT, MISCAPTIONED, or TRUE. If the pipeline fails at any stage (*e.g.* due to LLM policy filters), we mark the original post as TRUE during verification, adhering to the principle of innocent until proven guilty. As baselines, we used the available benchmarks of the VERITE dataset. In Figure 1, we can also see the verdicts and explanations for the first and third examples. These explanations also allow the user to understand the reasoning process of the model.

4 Experimental Results

In the following section, we introduce our (1) baselines and experimental results (2).

4.1 Baselines

VERITE (Papadopoulos et al., 2024b). The VERITE dataset paper introduces a transformer-

based model for detecting misinformation by combining image and text information. It uses CLIP ViT-L/14 to extract visual and textual features, which are merged into a single vector representing the image-caption pair. The vector is then processed by a transformer encoder that omits positional encodings and applies average pooling with multi-head self-attention. A final classification layer predicts the label of the image-caption pair. The model is trained on datasets like CLIP-NESr and CHASMA-D, which include synthetic multimodal misinformation. To handle class imbalance, random down-sampling was used, and the model was trained using categorical cross-entropy loss for multiclass classification.

RED-DOT (Papadopoulos et al., 2024a). The Relevant Evidence Detection Directed Transformer (RED-DOT) is a model for multimodal fact-checking that focuses on identifying and leveraging relevant evidence. It uses CLIP-ViT-L/14 to extract visual features from images and textual features from captions. An evidence re-ranking module emphasizes relevant content via intra-modal similarity, while irrelevant items are filtered using hard negative sampling. Features from both modalities are fused using element-wise operations and concatenation, then processed by a transformer to predict evidence relevance and the overall class. RED-DOT is trained on the NewsCLIPings+ dataset with multi-task learning and evaluated using Out-of-Distribution Cross-Validation (OOD-CV).

MultiReflect models. We compare the efficiency of our pipeline using five different vision LLM models: GPT-4V (OpenAI, 2023), GPT-4o-mini (OpenAI, 2024), Gemma 3 (Team et al., 2025), LLaVA-CoT (Xu et al., 2024a) and DeepSeek-VL2 (Wu et al., 2024). GPT-4V is a large vision-language model that integrates advanced visual and textual reasoning across different domains. GPT-4o-mini builds on this by offering a lighter, faster variant optimized for real-time, low-latency interaction. Gemma 3 (12B) is a general-purpose multimodal foundation model using a modified SigLIP vision encoder. LLaVA-CoT (11B) brings visual inputs together with step-by-step reasoning, improving performance on tasks that require both understanding and explanation. We select LLaVA-CoT and Gemma 3 as they perform on par with GPT-4o-mini on reasoning benchmarks. DeepSeek-VL2 (4.2B) similarly focuses on multimodal reasoning, using techniques like mixture-of-experts, dynamic image tiling and multi-head latent attention to ex-

Type	Class	GPT-4V				GPT-4o-mini				LLaVA-CoT (11B)				DeepSeek-VL2 (4.2B)				Gemma 3 (12B)			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
M	ALL	0.49	0.53	0.49	0.46	0.50	0.53	0.50	0.50	0.38	0.59	0.38	0.30	0.36	0.38	0.36	0.31	0.44	0.52	0.44	0.37
	TRUE	0.69	0.62	0.65	-	0.60	0.48	0.53	-	0.41	0.69	0.51	-	0.38	0.53	0.44	-	0.47	0.71	0.56	
	MC	0.54	0.16	0.25	-	0.56	0.34	0.42	-	1.00	0.00	0.01	-	0.34	0.49	0.40	-	0.41	0.59	0.49	
	OOC	0.37	0.69	0.48	-	0.43	0.70	0.53	-	0.34	0.45	0.38	-	0.42	0.04	0.07	-	0.71	0.02	0.03	
B	ALL	0.78	0.75	0.74	0.74	0.72	0.70	0.72	0.71	0.56	0.64	0.56	0.57	0.54	0.59	0.54	0.56	0.63	0.68	0.63	0.64
	TRUE	0.69	0.62	0.65	-	0.60	0.48	0.53	-	0.41	0.69	0.51	-	0.38	0.53	0.44	-	0.47	0.71	0.56	
	FALSE	0.81	0.86	0.83	-	0.76	0.84	0.80	-	0.75	0.49	0.59	-	0.70	0.55	0.62	-	0.80	0.59	0.68	

Table 1: Performance results of the proposed pipeline MultiReflect on the VERITE dataset. The results are shown for both the binary case (denoted as B, with labels TRUE and FALSE) and the multi-class case (denoted as M, with labels TRUE, MISCAPTIONED [MC], and OUT-OF-CONTEXT [OOC]). The drop in performance in the multi-class classification indicates that the model struggles to distinguish between MISCAPTIONED and OUT-OF-CONTEXT datapoints. The best overall accuracy and F1-scores for both binary and multi-class settings are **highlighted**. We observe that GPT-4o-mini performs best in the multi-class setting, while GPT-4V performs best in the binary classification setting.

Model	Accuracy	
	Multi-class	Binary
VERITE (Papadopoulos et al., 2024b)	0.52	0.73
RED-DOT (Papadopoulos et al., 2024a)		0.77
GPT-4V (OpenAI, 2023)	0.49	0.78
GPT-4o-mini (OpenAI, 2024)	0.50	0.72
LLaVA-CoT (11B) (Xu et al., 2024a)	0.38	0.56
DeepSeek-VL2 (4B) (Wu et al., 2024)	0.36	0.54
Gemma 3 (12B) (Team et al., 2025)	0.44	0.63

Table 2: The results show that MultiReflect with GPT-4V outperforms all baselines in binary classification. However, all MultiReflect versions underperform against the original VERITE baseline in multi-class classification.

tract and align the most relevant visual and textual features. We select DeepSeek-VL2 as a comparison due to its reliance on mixture-of-experts and good performance on reasoning benchmarks given its relatively small size.

4.2 Results

The VERITE dataset provides three classes: TRUE, MISCAPTIONED, and OUT-OF-CONTEXT. For binary classification, however, MISCAPTIONED and OUT-OF-CONTEXT are combined into a single FALSE class. We evaluate both binary and multi-class (taking into account all three classes).

Multi-class results: In the multiclass setting, our pipeline achieved the best result with GPT-4o-mini with a macro F1-score of 0.50 and accuracy of 0.50, slightly lower than the VERITE benchmark accuracy of 0.52 (see Table 2). Surprisingly, the score for the larger GPT-4V is lower, suggesting that the pipeline struggles to differentiate false subclasses. This is also shown when we look at the TRUE class, as the GPT-4V model performs the best with

a F1-score of 0.65, the highest among all classes. However, for the MISCAPTIONED class GPT-4V showed a low F1-score of 0.25, driven by a recall of 0.16, indicating difficulty in identifying MISCAPTIONED posts. The same difficulty arised for LLaVA-CoT, which only identified a single MISCAPTIONED post due to being overconfident in the verification stage. Surprisingly, Gemma 3 performs the best win identifying MISCAPTIONED posts with an F1-score of 0.49 showing its capability of using evidence critically. For OUT-OF-CONTEXT, GPT-4o-mini achieved an F1-score of 0.53, primarily due to low precision (0.43) of OUT-OF-CONTEXT class. DeepSeek-VL2, performs the worst out of the four with an F1-score of 0.36 due to misclassifying OUT-OF-CONTEXT posts. We note that these results highlight the pipelines poor capability to differentiate which modality includes the false information.

Binary results: In the binary setting, we see that GPT-4V performs the best out of the three models in all metrics achieving a F1-score of 0.74 and an accuracy of 0.78, exceeding the VERITE benchmark of 0.72 and RED DOT baseline of 0.77 (see Table 2). The TRUE class retained its F1-score of 0.65, while the combined false class achieved 0.83 as shown in Table 1. The overall result against other baselines is shown in Table 2, our model achieves the best binary results in the VERITE benchmark dataset. The open-source models (Gemma 3, LLaVA-CoT and DeepSeek-VL2) all perform worse in both classes compared to the OpenAI models. This performance gap may be attributed to less effective use of evidence in the verification process.

Type	Class	No Evidence				All Evidence			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Multi-Class	ALL	0.49	0.53	0.49	0.46	0.50	0.53	0.50	0.50
	TRUE	-	0.69	0.62	0.65	-	0.60	0.48	0.53
	MISCAPTIONED	-	0.54	0.16	0.25	-	0.56	0.34	0.42
	OUT-OF-CONTEXT	-	0.37	0.69	0.48	-	0.43	0.70	0.53
Binary	ALL	0.78	0.75	0.74	0.74	0.72	0.70	0.72	0.71
	TRUE	-	0.69	0.62	0.65	-	0.60	0.48	0.53
	FALSE	-	0.81	0.86	0.83	-	0.76	0.84	0.80

Table 3: Performance results on the VERITE dataset under **No Evidence** and **All Evidence** conditions using GPT-4V. The results are shown for both the multi-class (TRUE, MISCAPTIONED, and OUT-OF-CONTEXT) and binary (TRUE, FALSE) settings. The “ALL” row gives the overall accuracy (Acc.), while per-class rows show only precision (Prec.), recall (Rec.), and F1.

5 Ablation Study

We conduct an ablation study of our best-performing model, GPT-4V, on the benchmark to evaluate the role of evidence within the **MultiReflect** pipeline. This analysis focuses on two key questions: (1) Is any evidence necessary for effective verification? (2) Does the ranking of evidence contribute meaningfully to performance? To address the first question, we evaluate the system’s performance when no evidence is provided during the verification stage. For the second, we provide all available evidence without applying any ranking. The results demonstrate that RAG-enhanced retrieval and ranking both play a critical role in strengthening multimodal reasoning.

5.1 No evidence

This subsection analyzes the pipeline without using evidence, excluding phases 2 to 5, for both multi-class and binary settings. This means that this variation of the pipeline only checks for consistency and then goes directly into verification if the post is found to be consistent.

Multi-class results: Without evidence, the model achieves an F1-score of 0.41, which is lower than the pipeline’s 0.46, as shown in Table 3. This indicates that evidence improves multi-class verification. Specifically, for the TRUE class, the F1-score drops to 0.46 from 0.65. Interestingly, MISCAPTIONED posts perform better without evidence, achieving an F1-score of 0.29 compared to 0.25, suggesting that evidence may mislead in this category. In both evaluation scenarios, the F1-score for MISCAPTIONED posts remains consistently low, highlighting the model’s persistent difficulty in accurately distinguishing them from the other classes.

Binary results: As detailed in Table 3, without evidence, the model’s overall F1-score is 0.63, underperforming compared to 0.74 in the full pipeline. The classwise F1-scores for TRUE and FALSE drop to 0.46 and 0.79 from 0.65 and 0.83, respectively, highlighting the importance of evidence in binary settings. The larger drop in the TRUE class score highlights that evidence is crucial for reducing false negatives and confirming truthful posts, as its absence increases uncertainty.

5.2 All evidence

This subsection analyzes the pipeline without phase 2 (evidence checking), providing all retrieved evidence in the verification phase for both multi-class and binary settings.

Multi-class results: Providing all evidence does not improve the F1-score beyond 0.46, matching the regular pipeline’s performance as reflected in Table 3. This suggests that adding more evidence does not necessarily enhance the model’s accuracy. However, giving all of the evidence adds additional computational costs to the pipeline, making the regular pipeline more preferable. The classwise F1-scores for all classes are lower than in the full pipeline, except for MISCAPTIONED, which increases to 0.30 from 0.25.

Binary results: With all evidence included, the F1-score decreases to 0.72 compared to 0.74 in the full pipeline, confirming that an overload of evidence can hinder effective reasoning, as shown in Table 3. The F1-scores for TRUE and FALSE are slightly lower at 0.63 and 0.81, respectively, than those in the regular pipeline. This highlights the need for careful evidence selection methods as providing all of the retrieved evidence can make the reasoning noisy in the verification phase.

Fact-Checking System	Evidence Retrieval	Multimodal	Verification	Evidence Ranking	RAG
COSMOS (Aneja et al., 2021)	×	✓	✓	×	×
EXMULF (Amri et al., 2022)	×	✓	✓	×	×
Twitter-COMMs (Biamby et al., 2022)	×	✓	✓	×	×
MuRAG* (Chen et al., 2022)	✓ (static knowledge base)	✓	×	✓	✓
CCN (Abdelnabi et al., 2022)	✓ (internet)	✓	✓	×	×
BERT + LSTM (Hammouchi and Ghogho, 2022)	✓ (internet)	✓	✓	✓ (source credibility)	×
Self-RAG* (Asai et al., 2023)	✓ (internet + static knowledge base)	×	×	✓ (relevancy, support- edness, usefulness)	✓
ChartBERT (Akhtar et al., 2023a)	×	✓	✓	×	×
FakeNewsGPT4 (Liu et al., 2024)	×	✓	✓	×	×
RED-DOT (Papadopoulos et al., 2024a)	✓ (internet)	✓	✓	✓ (similarity)	×
MultiReflect (Ours)	✓ (internet)	✓	✓	✓	✓

Table 4: Overview of related works and associated features. We highlight that our work **MultiReflect** is the only one to utilize RAG for the multimodal verification task. The (*) refers to work in the domain of Question-Answering.

6 Related Works

In this section, we introduce several related methods and papers to our work. We additionally highlight the main feature differences between the methods in Table 4.

Automated Fact-Checking. Fact-checking methods have significantly evolved with advancements in artificial intelligence, particularly through the development of LLMs and automated systems. Early systems such as those introduced by Thorne et al. (2018) and Thorne and Vlachos (2021) relied on static knowledge bases for evidence retrieval for fact-checking and correction. However, these systems lacked the ability to update their knowledge bases dynamically, which is critical in the fast-paced information era.

Recent efforts have seen the integration of Retrieval Augmented Generation (RAG) techniques to enhance the reliability and accuracy of fact-checking systems. For instance, models such as MuRAG (Chen et al., 2022) and Self-RAG (Asai et al., 2023) have utilized not only static knowledge bases but also the internet to retrieve current and relevant information. These models enhance the fact-checking process by employing RAG for dynamic evidence retrieval, allowing for a more accurate verification of facts by evaluating various aspects of information quality. This approach significantly surpasses earlier models that relied only on static databases or lacked evidence ranking mechanisms (Gao et al., 2024).

Multimodal Fact-Checking. The need for multimodal fact-checking has grown with the increased

prevalence of complex information that spans various data types: text, image, video, and audio. Systems like COSMOS (Aneja et al., 2021), Twitter-COMMs (Biamby et al., 2022), EXMULF (Amri et al., 2022), ChartBERT (Akhtar et al., 2023a), RED-DOT and (Papadopoulos et al., 2024a) have made significant progress in tackling the challenges posed by multimodal data. However, prior works rely heavily on training, limiting usability in low-resource settings, and often focus only on intramodal relationships, overlooking nuanced cross-modal relationships.

To the best of our knowledge, our MultiReflect approach is the first to integrate evidence retrieval, multimodality, verification, evidence ranking, and RAG into a single fact-checking pipeline.

7 Conclusions

We introduce **MultiReflect**, a novel multi-modal RAG-based fact-checking pipeline. The novelty of the pipeline lies in its new evidence ranking and reflection scheme over multimodal posts. We validate the efficiency of the pipeline using a specialized multimodal fact-checking benchmark dataset VERITE. Our results show that MultiReflect underperforms in the multiclass setting but outperforms other baselines in the binary class scenario. Future works could improve this pipeline by focusing on how to better identify in which modality the error exists. Additionally, incorporating modality-specific retrieval strategies could help disentangle complex cross-modal contradictions.

8 Ethics Statement

The VERITE dataset used in this study is publicly available and specifically designed for benchmarking multimodal fact-checking systems. All annotation work was done by the study authors without crowd-sourcing. VERITE consists of fact-checked articles from Snopes and Reuters, curated by experts. Our pipeline retrieves only publicly available evidence using official APIs of search engines that comply with the Robots Exclusion Protocol.

Acknowledgment

This work has received funding from the EU H2020 program under the SoBigData++ project (grant agreement No. 871042), by the CHIST-ERA grant No. CHIST-ERA-19-XAI-010, (ETAg grant No. SLTAT21096), and partially funded by HAMISON project.

References

- Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. [Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources](#). [Preprint](#), arXiv:2112.00061.
- 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.
- Sabrine Amri, Dorsaf Sallami, and Esma Aimeur. 2022. [Exmulf: An explainable multimodal content-based fake news detection system](#). In [Foundations and Practice of Security](#), pages 177–187, Cham. Springer International Publishing.
- Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. [Cosmos: Catching out-of-context misinformation with self-supervised learning](#). [Preprint](#), arXiv:2101.06278.
- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). [Preprint](#), arXiv:2310.11511.
- Giscard Biambay, 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.
- Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. [Murag: Multimodal retrieval-augmented generator for open question answering over images and text](#). [Preprint](#), arXiv:2210.02928.
- Google Cloud. 2024. [Detect web entities and pages - cloud vision api](#).
- Community Notes. 2024. [Introduction](#).
- Eurostat. 2022. [Consumption of online news rises in popularity](#). Accessed: 2024-04-13.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. [Retrieval-augmented generation for large language models: A survey](#). [Preprint](#), arXiv:2312.10997.
- Google. 2024. [Programmable search engine - google for developers](#).
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). [Transactions of the Association for Computational Linguistics](#), 10:178–206.
- Hicham Hammouchi and Mounir Ghogho. 2022. [Evidence-aware multilingual fake news detection](#). [IEEE Access](#), 10:116808–116818.
- Uku Kangur, Roshni Chakraborty, and Rajesh Sharma. 2024. [Who checks the checkers? exploring source credibility in twitter’s community notes](#). [Preprint](#), arXiv:2406.12444.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). [Preprint](#), arXiv:2005.11401.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). [Preprint](#), arXiv:2301.12597.
- Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024. [A survey on fairness in large language models](#). [Preprint](#), arXiv:2308.10149.
- Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024. [Fakenewsgpt4: Advancing multimodal fake news detection through knowledge-augmented lvlms](#). [Preprint](#), arXiv:2403.01988.

- Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro, and Benjamin Van Durme. 2025. [Wikivideo: Article generation from multiple videos](#). Preprint, arXiv:2504.00939.
- Microsoft. 2024a. [Bing visual search v7 - microsoft learn](#).
- Microsoft. 2024b. [Bing web search v7 - microsoft learn](#).
- OpenAI. 2023. [Gpt-4 technical report](#). [arXiv](#), abs/2303.08774.
- OpenAI. 2024. [Gpt-4o system card](#). Preprint, arXiv:2410.21276.
- Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2024a. [Red-dot: Multimodal fact-checking via relevant evidence detection](#). Preprint, arXiv:2311.09939.
- Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024b. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). Preprint, arXiv:2103.00020.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. [arXiv preprint arXiv:2503.19786](#).
- James Thorne and Andreas Vlachos. 2021. [Evidence-based factual error correction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3298–3309, Online. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Wikimedia. 2024. [Api portal](#).
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. [Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding](#). Preprint, arXiv:2412.10302.
- Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024a. [Llava-o1: Let vision language models reason step-by-step](#). [arXiv preprint arXiv:2411.10440](#).
- Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024b. [Hallucination is inevitable: An innate limitation of large language models](#). Preprint, arXiv:2401.11817.

Appendix

The appendix is structured into three sections: (A) Additional Information, (B) Qualitative Examples and (C) Prompts.

A Additional Information

A.1 Limitations

There are several limitations that have an impact on the pipeline’s results. First, not all post have evidence available for them, thus reducing the quality of the verification of those posts. Future works could solve this issue by expanding the amount of evidence sources. Second, as generative models are prone to hallucinate, it might be that the model sometimes hallucinates on the given evidence - this being specifically the case when we provide all evidence. Additionally, the OpenAI API policy filters might refuse to answer some prompts. If the pipeline is to be used, we recommend always including a human in the loop and running the model several times and taking into account the standard deviation of the results. Third, there is no way to identify if an evidence is originally written by the source where it comes from. This can create problems as platforms can repost information in misleading contexts. A possible solution for this would be to keep a blacklist of uncredible sources. Fourth, the pipeline is rather costly as for one post. The costliness primarily arises from the amount of evidence (10 images and 10 texts) that is retrieved and ranked. It might require around 70-100 prompts to verify all evidences involved. Cost can be lowered by reducing the amount of evidences retrieved, but

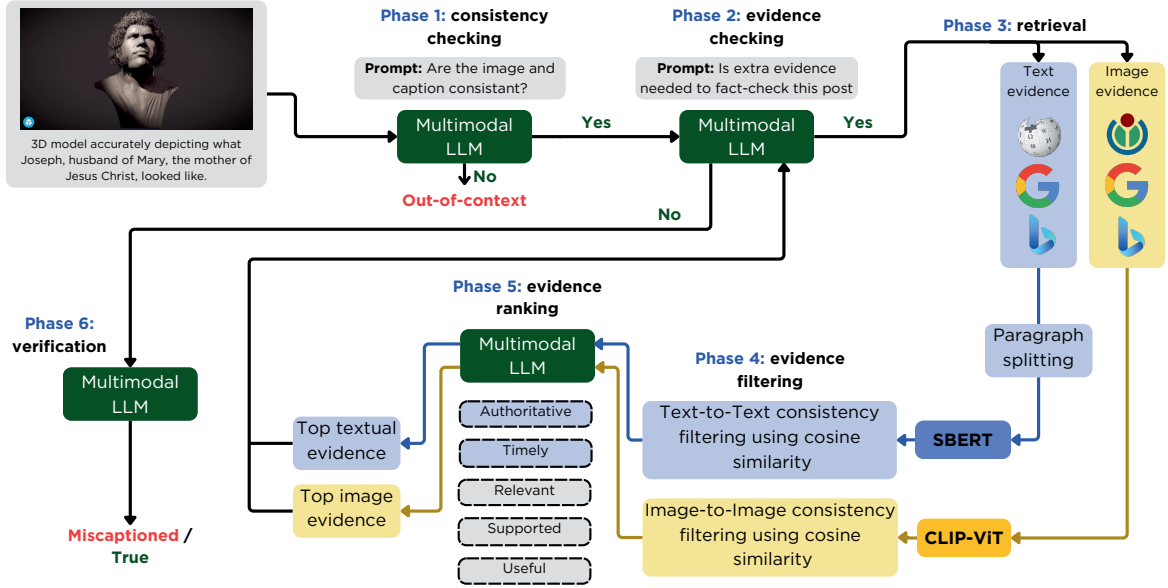


Figure 3: Overview of the **MultiReflect** Pipeline. The pipeline is processed in six steps: The *first phase* checks if the image and text are consistent. The *second phase* checks if evidence is needed for fact-checking the image-text pair. The *third phase* retrieves image and text evidences using different search APIs. The *fourth phase* filters the evidence so that both the image and text evidences are consistent to the original image and text. The *fifth phase* ranks the evidence based on five features. The top ranked evidence is extracted and if no more evidence is needed then the pipeline end with verifying the image-text pair in *phase six*. Note that we highlight procedures involving both image and text modalities in **gray/black**, procedures involving only image data in **yellow** and procedures involving only text data in **blue**.

Model	Threshold	Acc	F1	P	R
CLIP Large 336	0.28	0.633	0.591	0.474	0.784
BLIP 2 (2-7B) (Li et al., 2023)	0.13	0.369	0.500	0.341	0.930
BLIP 2 (6-7B)	0.10	0.359	0.503	0.341	0.906
BLIP 2 FLAN XL	0.10	0.358	0.505	0.349	0.966
GPT-4V (OpenAI, 2023)	N/A	0.680	0.638	0.517	0.834

Table 5: **Full result.** Performance Metrics of Different Consistency Checking Strategies (Phase 1). We rely on three validation methods: Image-to-Text consistency using CLIP, Text-to-Text consistency using BLIP (via SBERT), and Multimodal consistency using GPT-4V models.

that can have a negative effect on the performance of the pipeline. Finally, the pipeline performs sub-optimally on open-source models. LLaVA-CoT exhibits confirmation bias during verification, labeling nearly all posts as TRUE regardless of evidence. DeepSeek-VL2, on the other hand, struggles with consistency checks, resulting in low accuracy for OUT-OF-CONTEXT cases.

A.2 Consistency checking

The detailed scores for the consistency checking phase are highlighted in Table 5. The table shows that the multimodal GPT-4V surpassed all of the models in terms of accuracy. Surprisingly BLIP 2 FLAN XL got a better recall, showing its better capability in detecting consistent image-text pairs

compared to non-consistent ones.

A.3 Dataset examples

We additionally provide six example image-text pairs from the original VERITE dataset. We highlight in Table 6 all three class variants (TRUE, MISCAPTIONED, OUT-OF-CONTEXT). The TRUE variant has the correct caption together with the correct image. The MISCAPTIONED variant has the wrong caption together with the correct image. The OUT-OF-CONTEXT variant has the correct caption together with the wrong image. As demonstrated by the examples, the dataset demands complex reasoning that involves interpreting text embedded within images, recognizing visual elements, and applying external knowledge about

events or well-known individuals. This highlights the complexity of the task.

A.4 Implementation Details

The high level overview of the MultiReflect pipeline is shown in Figure 3. The figure shows all of the 6 phases, with their corresponding tasks. We outline the implementation details of the models used in the experiments. All models were initialized with their default parameters to ensure reproducibility and consistency across experiments. The experiments for Gemma 3 and LLaVA-CoT were using a 2xV100 GPU with 64 GB VRAM. For the LLaVA-CoT (11B) the model ran for approximately 20 days, while the Gemma 3 (12B) model ran for 1 week. The experiments for DeepSeek-VL2 were using a A100 GPU with 80 GB VRAM. For DeepSeek-VL2 (4.2B) the model ran for approximately 1 week. All models used the default temperature for generations. The model versions used are the following:

GPT-4V: [gpt-4-1106-vision-preview](#)⁵

GPT-4o-mini: [gpt-4o-mini-2024-07-18](#)⁶

Gemma 3: [gemma-3-12b-it](#)⁷

LLaVA-CoT: [Llama-3.2V-11B-cot](#)⁸

DeepSeek-VL2: [deepseek-vl2](#)⁹

CLIP-336: [clip-vit-large-patch14-336](#)¹⁰

SBERT: [all-mpnet-base-v2](#)¹¹

BLIP-2 2.7B: [blip2-opt-2.7b](#)¹²

BLIP-2 6.7B: [blip2-opt-6.7b](#)¹³

BLIP 2 FLAN XL: [blip2-flan-t5-xl](#)¹⁴

B Qualitative Examples

We introduce qualitative examples predicted by the MultiReflect pipeline using GPT-4V. We show examples from GPT-4V due to its largest accuracy, but also due to it giving also the reasoning for its verification label, something other models did not show in the final output. We separate these examples into two - those that do not require evidences for verification in Table 7 and those that do require

evidence in Table 8.

Without evidence: The first example shows a mother fox feeding cubs near Montreal, Canada, in 2009. However, upon analyzing the image, the pipeline identifies a golden jackal, not a fox, which is clear from its physical characteristics, thus classifying it as OUT-OF-CONTEXT. The second and third examples show known people from news stories: Justine Damond and Dmytro Vasilievich Khaladzi. The pipeline successfully identifies the people on the image together with the context of their news story. The fourth example caption claims that U.S. President Donald Trump said, "I don't care how sick you are. [...] Get out and vote" during a November 2016 campaign event. However, the image shows a similar tweet from Eric Trump in November 2020. Despite the text's alignment with the image's message, the people involved and the dates do not match, leading the pipeline to classify the caption as OUT-OF-CONTEXT. In the fifth example, a shocking image about Christmas display is presented. The pipeline argues that since the caption describes the image exactly as it is, then there is no reason to doubt the authenticity of the post.

With evidence: In the first example, the caption about Melania Trump's dress is verified by evidence confirming it is part of Alexander McQueen's Spring/Summer 2020 collection, designed with art students. The image evidence shows the author of the dress, which additionally does not contradict the textual evidence and thus allows the pipeline to verify the post as TRUE. The second example describes the London riots in August 2011. The text evidence supports the caption, confirming the locations and dates of the riots, making the caption true, despite the image not matching a specific looting event. The third example features an archaeologist excavating a woolly mammoth near Paris in 2012. Text evidence corroborates this discovery, confirming the caption as true.

⁵[openai.com/index/gpt-4v-system-card/](#)

⁶[platform.openai.com/docs/models/gpt-4o-mini](#)

⁷[huggingface.co/google/gemma-3-12b-it](#)

⁸[huggingface.co/xkev/Llama-3.2V-11B-cot](#)

⁹[huggingface.co/deepseek-ai/deepseek-vl2](#)

¹⁰[huggingface.co/openai/clip-vit-large-patch14-336](#)

¹¹[huggingface.co/sentence-transformers/all-mpnet-base-v2](#)

¹²[huggingface.co/Salesforce/blip2-opt-2.7b](#)

¹³[huggingface.co/Salesforce/blip2-opt-6.7b](#)

¹⁴[huggingface.co/Salesforce/blip2-flan-t5-xl](#)



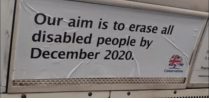






Image	Image Caption	label
	Image shows a 'We plan to cut all homeless people in half by 2025' sign created by street artist who goes by the name of Foka Wolf.	TRUE
	Image shows signs posted by the UK Conservative Party vowing 'We plan to cut all homeless people in half by 2025'	MISCAPTIONED
	Image shows a 'We plan to cut all homeless people in half by 2025' sign created by street artist who goes by the name of Foka Wolf.	OUT-OF-CONTEXT
	A sketchfab 3D rendering of the actor Andre the Giant.	TRUE
	3D model accurately depicting what Joseph, husband of Mary, the mother of Jesus Christ, looked like.	MISCAPTIONED
	A sketchfab 3D rendering of the actor Andre the Giant.	OUT-OF-CONTEXT
	Image shows grounds covered with garbage after the end of Glastonbury Music Festival in 2015.	TRUE
	Image shows grounds covered with garbage after Greta Thunberg's environmentalist speech at the Glastonbury Music Festival in June 2022.	MISCAPTIONED
	Image shows grounds covered with garbage after the end of Glastonbury Music Festival in 2015.	OUT-OF-CONTEXT

Table 6: Examples of image-caption pairs in the VERITE dataset. The dataset contains a variety of image-caption pairs that are challenging and require context understanding and reasoning. Each example has three versions. TRUE has the correct image with the correct caption. MISCAPTIONED has the correct image with the wrong caption. OUT-OF-CONTEXT has the wrong image with the correct caption.

Input image



Input caption: Image of a mother fox feeding fox cubs near Montreal, Canada in 2009.

Output: [OUT-OF-CONTEXT] The caption claims that the image depicts a mother fox feeding her cubs near Montreal, Canada in 2009. However, the image shows a golden jackal (not a fox) with its pups, which can be determined by the physical characteristics of the animals in the picture such as the shape of their ears, the color and pattern of the fur, and their general morphology. Golden jackals are not native to Canada, they are found in regions such as Africa, Southeast Europe, and parts of Asia. Therefore, the caption is out-of-context as it erroneously identifies the animal species and their location. Moreover, the date provided cannot be verified from the image alone, but the main reason for assigning an out-of-context label is the incorrect identification of the species and their geographical location. evidence not needed



Input caption: Image of Justine Damond, a white woman, was killed by a police officer named Mohamed Noor in Minneapolis in 2017.

Output: [TRUE] The image provided matches widely circulated images of Justine Damond, and the caption correctly states that she was killed by a police officer named Mohamed Noor in Minneapolis in 2017. evidence not needed



Input caption: Ukrainian strongman Dmytro Vasilievich Khaladzhi - a power-lifting champion - carrying a horse.

Output: [TRUE] The image shows Ukrainian strongman Dmytro Khaladzhi carrying a horse, consistent with the provided caption. Dmytro Khaladzhi is known for performing feats of strength, and a video of him carrying a horse has been circulated online in the past, confirming that this image is likely true to the caption's claim. Therefore, the image and caption are factually correct and in context. evidence not needed



Minnesota get out and vote!!!

5:35 AM · 11/10/20 · Twitter for iPhone

2,584 Retweets 791 Quote Tweets 14.2K Likes

Input caption: U.S. President Donald Trump said 'I don't care how sick you are. [...] Get out and vote' during a campaign event in November 2016.

Output: [OUT-OF-CONTEXT] The caption pertains to Donald Trump making a statement during a campaign event in November 2016, while the image shows a tweet from Eric Trump dated November 10, 2020, encouraging people to vote. The content of the tweet does not match the statement attributed to Donald Trump in the caption, and the dates also do not align. Therefore, the image is out of context with the caption provided. evidence not needed



Input caption: A picture of a humorous Christmas display of a doll staged to look as if the homeowner was trapped by a fallen ladder while installing Christmas decorations.

Output: [TRUE] The image shows a humorous Christmas display featuring what appears to be a doll dressed in red and blue, with its legs dangling from a gutter and a ladder fallen below, creating the illusion that someone was trapped while putting up Christmas decorations. Since the caption accurately describes what is depicted in the image, and there's no internal contradiction or reason to doubt the authenticity of the humor intended by the display, the statement is factually correct. evidence not needed

Table 7: Examples of pipeline outputs with GPT-4V that did not require evidence for verification.

Input image:



Input caption: Image of Melania Trump wore a dress designed by Alexander McQueen in collaboration with art students and was inspired by life drawings.

Output: [TRUE] The caption describes Melania Trump wearing a dress designed by Alexander McQueen, which was a collaboration with art students inspired by life drawings. The text evidence confirms that the dress is from Alexander McQueen's Spring/Summer 2020 collection, and the sketches on the dress were created during a life drawing class held at the Alexander McQueen educational space in London. The sketches were drawn by Central Saint Martins MA students.

[...] **evidence needed**

Image evidence:



Image scores: Relevancy=1, Support=0, Usefulness=0.

Text evidence:

Sketches on dress are from students. The dress worn by the first lady did come from Alexander McQueen's Spring/Summer 2020 collection, however, the drawings on the dress were not created by children who had been victims of sexual abuse as posts claim. On March 21, Alexander McQueen's official Facebook page posted that the dress features dancing girls. "The continuous, spontaneous sketches of dancing girls by Central Saint Martins MA students, created during a life drawing class held at the Alexander McQueen educational space in London this year, is the source of these motifs," the post reads, along with a video of the dress. [...]

Text scores: Relevancy=1, Support=0.5, Usefulness=1, Factuality=0.33, Reliability=1, Bias = 0.5, Timeliness=0.

Input image:



Input caption: Photograph shows London residents looting electronic devices during riots that erupted after police killed a black man in August 2011.

Output: [TRUE] The caption describes a specific event that took place in August 2011, referring to the London riots that were indeed sparked by the police killing of a black man. The external text evidence lists various locations and dates, including Enfield, Brixton, Hackney, and others, along with dates that match the August 2011 timeline. [...]

[...] **evidence needed**

Image evidence:

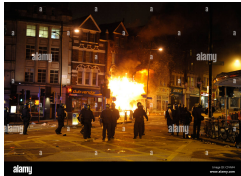


Image scores: Relevancy=1, Support=0.5, Usefulness=0.5.

Text evidence:

Luke Macgregor / X01981: Police cordon off an area in Enfield, north London, on August 7, 2011. Karel Prinsloo / AP: Youths throw bricks at police in Enfield, north London, on August 7, 2011. Dan Kitwood / Getty Images Europe: A girl is detained outside Currys electrical store in Brixton, south London, on August 8, 2011. Peter Macdiarmid / Getty Images Europe: People loot a Carhartt store in Hackney, north London, on August 8, 2011. [...]

Text scores: Relevancy=1, Support=0.5, Usefulness=1, Factuality=0.66, Reliability=1, Bias = 0.5, Timeliness=0.

Input image:



Input caption: Image shows an archaeologist excavating the remains of a woolly mammoth that was found at a quarry site near Paris, France in November 2012.

Output: [TRUE] [...]. The text evidence from BBC Newsround corroborates the scene depicted in the image by reporting that a woolly mammoth skeleton was indeed unearthed near Paris in the Changis-sur-Marne area, and the discovery occurred in November 2012. [...]

[...] **evidence needed**

Image evidence:

All image evidences filtered out.

Image scores: N/A.

Text evidence:

Woolly mammoth skeleton found near Paris, France [...]. The skeleton of a woolly mammoth has been unearthed - not in remote Siberia, but near the capital of France, Paris! It was discovered accidentally by a team digging at an ancient Roman site in the Changis-sur-Marne area. [...]

Text scores: Relevancy=1, Support=0.5, Usefulness=1, Factuality=0.66, Reliability=1, Bias = 1, Timeliness=0.

Table 8: Examples of pipeline outputs with GPT-4V that required evidence retrieval for verification. We additionally provide the scores for the top ranked evidences retrieved for these input posts.

C. Prompts

In the MultiReflect pipeline, prompts play an important role in evaluating the quality of both the original input post and the evidences retrieved. In this section, we introduce the prompts used within the pipeline. The pipeline utilizes prompts in four phases: consistency checking (phase 1), evidence checking (phase 2), evidence ranking (phase 5) and verification (phase 6). For GPT-4V and GPT-4o the images for both the post and evidences were given through the OpenAI API platform.

1. Consistency checking

For consistency checking (phase 1) we used the following prompt together with the original image.

Prompt 1: Given a caption and image, determine whether the caption matches the image or not, if yes respond `<verdict>TRUE</verdict>` else `<verdict>FALSE</verdict>`, also give the consistency score between 0 and 1 like `<score>...</score>`
Caption: {caption}
{encoded image}

2. Evidence checking

For evidence checking (phase 2), we use two different prompts. The first time this phase is initiated, we use this prompt together with the original caption and image.

Prompt 2: Given a image and caption, please make a judgment on whether finding some external documents from the web (e.g., Wikipedia) helps to decide whether the image and caption is factually correct. Please answer [Yes] or [No] and write an explanation.
Caption: {caption}
{encoded image}

If we run into phase 2 again, then during the next times we use:

Prompt 3: Given a image and caption along with some external documents (evidences). Your task is to determine whether the factuality of the image and caption can be fully verified by the evidence or if it requires further external verification. There are three cases:
- If image and caption can be verified solely with the evidences, then respond with [Continue to Use Evidence].
- If the sentence doesn't require any factual verification (e.g., a subjective sentence or a sentence about common sense), then respond with [No Retrieval].
- If additional information is needed to verify, respond with [Retrieval].
Please provide explanations for your judgments
Caption: {caption}
{encoded image}
Evidences: {evidence texts and encoded images}

3. Evidence ranking

Evidence ranking (phase 5) get the relevancy, support and usefulness scores using prompts. For each of these prompts we used two variations, one for ranking images and another for ranking texts. For relevancy, we used the following two prompts.

Prompt 4: You'll be provided with an image, along with evidence. Your job is to determine if the evidence is relevant to the determine the factual correctness of the image, and provides useful information to complete the task described in the instruction. If the evidence meets this requirement, respond with [Relevant]; otherwise, generate [Irrelevant]. Also determine the relevancy score of the evidence, on a scale of 0 to 1.
{encoded image}
Text Evidence: {evidence text}

Prompt 5: You'll be provided with a text, along with an image evidence. Your job is to determine if the evidence is relevant to the determine the factual correctness of the text, and provides useful information to complete the task described in the instruction. If the evidence meets this requirement, respond with [Relevant]; otherwise, generate [Irrelevant]. Also determine the relevancy score of the evidence, on a scale of 0 to 1.
Text: {caption}
{evidence encoded image}

For support, we used the following two prompts.

Prompt 6: You will receive an input text, input image and text evidence towards determining the factuality of the input. Your task is to evaluate if the input is fully supported by the information provided in the evidence. Use the following entailment scale to generate a score:
- [Fully supported] - All information in input is supported by the evidence, or extractions from the evidence.
- [Partially supported] - The input is supported by the evidence to some extent, but there is major information in the input that is not discussed in the evidence. For example, if the input asks about two concepts and the evidence only discusses either of them, it should be considered a [Partially supported].
- [No support / Contradictory] - The input completely ignores evidence, is unrelated to the evidence, or contradicts the evidence. This can also happen if the evidence is irrelevant to the instruction.
Make sure to not use any external information/knowledge to judge whether the input is true or not. Only check whether the input is supported by the evidence, and not whether the input follows the instructions or not. Output Entailment like [Fully supported], [Partially supported] or [No support / Contradictory]
Input text: {caption}
Input Image: {encoded image}
Text Evidence: {evidence text}

Prompt 7: You will receive an input text, input image and image evidence towards determining the factuality of the input. Your task is to evaluate if the input is fully supported by the information provided in the evidence. Use the following entailment scale to generate a score:

- [Fully supported] - All information in input is supported by the evidence, or extractions from the evidence.
- [Partially supported] - The input is supported by the evidence to some extent, but there is major information in the input that is not discussed in the evidence. For example, if the input asks about two concepts and the evidence only discusses either of them, it should be considered a [Partially supported].
- [No support / Contradictory] - The input completely ignores evidence, is unrelated to the evidence, or contradicts the evidence. This can also happen if the evidence is irrelevant to the instruction.

Make sure to not use any external information/knowledge to judge whether the input is true or not. Only check whether the input is supported by the evidence, and not whether the input follows the instructions or not.

Output Entailment on the first line and the explanation on the second line.

Input text: {caption}

Input Image: {encoded image}

Image Evidence: {evidence encoded image}

For usefulness, we used the following two prompts.

Prompt 8: Given an input text and input image along with an text evidence, rate whether the evidence appears to be a helpful and informative answer to determine the factuality of the input, from 1 (lowest) - 5 (highest). We call this score perceived utility. The detailed criterion is as follows: 5: The evidence provides a complete, highly detailed, and informative response to the factuality of the input, fully satisfying the information needs. 4: The evidence mostly fulfills the need to get the factuality of the input, while there can be some minor improvements such as discussing more detailed information, having better structure of the evidence, or improving coherence. 3: The evidence is acceptable, but some major additions or improvements are needed to satisfy factuality. 2: The evidence still addresses the main request, but it is not complete or not relevant to the input. 1: The response is barely on-topic or completely irrelevant.

Input text: {caption}

Input Image: {encoded image}

Text Evidence: {evidence text}

Prompt 9: Given an input text and input image along with an image evidence, rate whether the evidence appears to be a helpful and informative answer to determine the factuality of the input, from 1 (lowest) - 5 (highest). We call this score perceived utility. The detailed criterion is as follows: 5: The evidence provides a complete, highly detailed, and informative response to the factuality of the input, fully satisfying the information needs. 4: The evidence mostly fulfills the need to get the factuality of the input, while there can be some minor improvements such as discussing more detailed information, having better structure of the evidence, or improving coherence. 3: The evidence is acceptable, but some major additions or improvements are needed to satisfy factuality. 2: The evidence still addresses the main request, but it is not complete or not relevant to the input. 1: The response is barely on-topic or completely

Input text: {caption}

Input Image: {encoded image}

Image Evidence: {evidence encoded image}

4. Verification

During verification (phase 6), we have two different prompts - one for verifying with evidence and one without evidence. Note that the prompt here outputs true/false, but later depending on the dataset these can be renamed to actual classes. The prompt with evidence is as follows.

Prompt 10: You will receive an image and caption along with some external documents (evidences). Based on the evidences provided you need to determine factual correctness of the input image and caption. If the input image and caption are out-of-context output [OUT-OF-CONTEXT], else if factually correct output [TRUE], otherwise [FALSE]. Also output the confidence score in scale 0 to 1 for the same decision.

Caption: {caption}

{encoded image}

Evidences: {evidence texts and encoded images}

When verifying without evidence, then the pipeline uses the following prompt.

Prompt 11: You will receive an image and caption. Based on the knowledge you have, you need to determine factual correctness of the input image and caption. If the input image and caption are out-of-context output [OUT-OF-CONTEXT], else if factually correct output [TRUE], otherwise [FALSE]. Also output the confidence score in scale 0 to 1 for the same decision.

Caption: {caption}

{encoded image}