"Image, Tell me your story!" Predicting the original meta-context of visual misinformation

Anonymous ACL submission

Abstract

To assist human fact-checkers, researchers have developed automated approaches for visual misinformation detection. These methods assign veracity scores by identifying inconsistencies between the image and its caption, or by detecting forgeries in the image. However, they neglect a crucial point of the human fact-checking process: identifying the original meta-context of the image. By explaining what is *actually true* about the image, fact-checkers can better detect misinformation, focus their efforts on check-worthy visual content, engage in counter-messaging before misinformation spreads widely, and make their explanation more convincing. Here, we fill this gap by introducing the task of automated image contextualization. We create 5Pils, a dataset of 1,676 fact-checked images with question-answer pairs about their original meta-context. Annotations are based on the 5 Pillars fact-checking framework. We implement a first baseline that grounds the image in its original meta-context using the content of the image and textual evidence retrieved from the open web. Our experiments show promising results while highlighting several open challenges in retrieval and reasoning.

1 Introduction

The surge of visual misinformation poses a growing threat to our society. In 2023, more than 30% of all fact-checked claims contained images, an increasing portion of which are AI-generated (Dufour et al., 2024). Journalists and fact-checkers have developed pipelines for detecting and debunking visual misinformation (Silverman, 2013; Urbani, 2020; Khan et al., 2023a,b; Dufour et al., 2024). Image contextualization is a key component of these pipelines: it establishes the meta-context of the image by collecting evidence and answering a concise set of questions. Urbani (2020) proposes a framework, which we refer to as "5 Pillars", where information about the Provenance (Was the image used before?), Source (Who is the source?), Date (When was it taken?), Location (Where was it taken?), and Motivation (Why was it taken?) serve as the image meta-context. For example, Figure 1 shows an out-of-context image purporting to represent an Indian plane evacuating people from Kabul in 2021. After conducting a 5 Pillars contextualization, fact-checkers established that the image was taken in 2013 in a US plane in the Philippines.

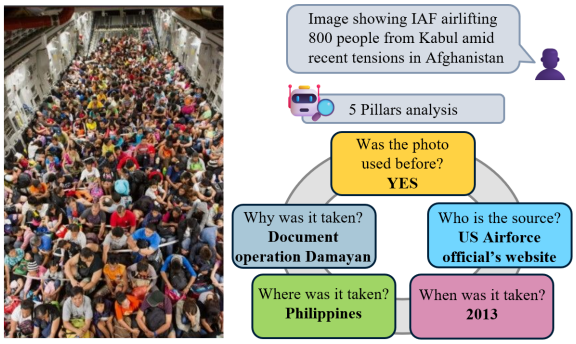

Figure 1: An out-of-context image with a false caption. The original meta-context of the image is established by answering the questions of the 5 Pillars framework.

Image contextualization benefits to several other components of the human fact-checking (FC) pipeline. (1) It helps identify inconsistencies between false claims and images, as shown in Figure 1. (2) It is a mandatory component of writing a solid debunking article. Indeed, debunking requires not only stating *why a claim is false* but also *what alternative is true instead* (Lewandowsky et al., 2020). (3) It allows writing prebunking articles (Guo et al., 2022), by informing about the image's true meta-context before misinformation spreads widely. (4) It helps assess whether the image is check-worthy. For example, images published by an unreliable source, or depicting critical events (war, crises, ...) deserve more debunking efforts.

Image contextualization is a time-consuming task that involves finding several evidence pieces, and combining various tools that require expertise (Sil-

verman, 2013). Despite its crucial position in the human FC pipeline, the automation of image contextualization has been ignored in favor of other components, predominantly manipulation (Huh et al., 2018; Liu et al., 2023c; Shao et al., 2023) and image-caption inconsistency detection (Luo et al., 2021; Abdelnabi et al., 2022; Papadopoulos et al., 2024; Qi et al., 2024). The output of these tasks is a score, indicating whether or not the content is misinformation. Fact-checkers have expressed scepticism about the real-world performance of these scores, with the risk of publishing incorrect automated verdicts (Arnold, 2020). In contrast, there is a real interest in methods that provide evidence and additional context (Arnold, 2020; Nakov et al., 2021; Schlichtkrull et al., 2023b), which includes image contextualization.

The contributions of this work are as follows: (1) We fill the gap by introducing the novel task of automated image contextualization. Given an image, a model must identify its meta-context using the image content and evidence from the open web. (2) We collect 5Pils, the first dataset with question-answer pairs about the image meta-context based on the 5 Pillars framework (Urbani, 2020). (3) We create the first baseline for image contextualization with textual evidence retrieved from the open web. Our extensive quantitative and qualitative analyses reveal that the task is challenging, even for SOTA large language models (LLMs). We make our code and dataset loaders available upon acceptance under an Apache 2.0 license. We analyze the intended uses of this work in Appendix A (Schlichtkrull et al., 2023b).

2 Related work

Visual misinformation

Existing work on visual misinformation focuses on assigning a veracity score to images. Many real-world (Zlatkova et al., 2019; Suryavardan et al., 2023; Papadopoulos et al., 2024) and synthetic (Jaiswal et al., 2017; Müller-Budack et al., 2020; Luo et al., 2021; Biamby et al., 2022; Papadopoulos et al., 2023a) datasets have been introduced. Separate methods are used for *out-of-context* images, where a score is assigned based on image-caption inconsistencies (Khattar et al., 2019; Kou et al., 2020; Huang et al., 2022), and *manipulated* and *fake* images, where forgeries are detected in the image (Huh et al., 2018; Liu et al., 2023c; Shao et al., 2023). Some works rely on external evidence (Zlatkova et al., 2019; Abdelnabi et al., 2022; Pham et al., 2023; Hu et al., 2023a; Papadopoulos et al., 2023b; Yuan et al., 2023; Zhang et al., 2023; Qi et al., 2024), while others use only the image and the caption (Nakamura et al., 2020; Luo et al., 2021; Aneja et al., 2023; Papadopoulos et al., 2024; Huang et al., 2022). Akhtar et al. (2023) provide a comprehensive survey of the field. Unlike these works, we use the image and evidence not to provide a veracity score, but to ground the image in its original meta-context.

Question Answering for Fact-Checking

Representing automated FC (AFC) as a question answering (QA) task has attracted recent interest, especially for text claims that require several reasoning steps (Chen et al., 2022; Ousidhoum et al., 2022; Yang et al., 2022; Pan et al., 2023; Schlichtkrull et al., 2023a). By decomposing the reasoning, the verdict becomes more human-interpretable (Yang et al., 2022; Rani et al., 2023). Rani et al. (2023) and Chakraborty et al. (2023) introduce a framework for decomposing a text claim in atomic statements, which they call the 5W (who, what, when, where, and why). Unlike our work, these approaches do not ask questions about images, nor do they try to identify the meta-context of the image.

3 The 5 Pillars of Image Verification

The 5 Pillars of image verification is a framework proposed by Urbani (2020), which summarizes years of insights on journalistic best practice in image verification (Silverman, 2013; Urbani, 2020; Mossou and Higgins, 2021; Khan et al., 2023a). It divides image contextualization into 5 components, as shown in Figure 1: the Provenance, the Source, the Date, the Location, and the Motivation.

Provenance answers the question "*Was the image used before?*" (Khan et al., 2023a). Identifying prior uses of the image is essential for answering the subsequent questions. Hence, it is considered the most important pillar (Urbani, 2020). To answer this question, fact-checkers rely mainly on reverse image search (RIS) engines. These engines take images as input and return (partially) matching images on the Internet (Urbani, 2020).

The **Source** is the person who captured or created the image, to be distinguished from the person who shared the visual misinformation online. Establishing the Source allows to evaluate the credibility of

the original image. A combination of social media profiling, interviews, and RIS allow fact-checkers to answer this question.

The **Date**, i.e., "*When was the image taken?*", and the **Location**, i.e., "*Where was the image taken?*", position the image in time and space and highlight any mismatch with the claimed context. The tasks associated with these two pillars are also known as chronolocation and geolocation (Mossou and Higgins, 2021; Khan et al., 2023a). In the simplest case, the image contains the Date in EXIF format. However, most social media platforms remove EXIF data from uploaded images (Urbani, 2020). RIS results can also help to find the Date and the Location. If that is not enough, fact-checkers can define keywords to query search engines and retrieve (partially) matching images, a process detailed in Appendix C. In other cases, they search for cues in the image, for example, landmarks, road signs, clothing. Additional tools include street imagery, sun angle analysis, and historical weather records (Urbani, 2020; Mossou and Higgins, 2021).

The **Motivation**, asks the question "*Why was the image taken?*". In most cases, the Motivation is to report on an event. However, some sources may have a particular agenda, for example if they are political activists. In other cases, a fake image used for malicious purposes might originate from the harmless work of an artist. The source's Motivation is different from the claimant's intent, which is explored in Da et al. (2021).

Figure 1 shows the 5 Pillars analysis of an out-of-context image. Examples with manipulated and fake images are shown in Appendix B. Given its popularity and comprehensive documentation (Urbani, 2020; Mossou and Higgins, 2021; Khan et al., 2023a), we use the 5 Pillars framework to define QA pairs for automated image contextualization.

4 Task definition and evaluation

The task is to provide answers for each of the 5 Pillars using the image content itself and evidence retrieved from the open web. Formally, a model has to generate answer A given a question Q, an image I, and evidence items E retrieved from the open web W using a retriever R,

$$A = \underset{a}{argmax}\ P\left(a \mid Q, I, E\right), \quad E = R(W).$$

We impose two restrictions on the open web retriever. (1) It cannot return evidence published after the FC article from which we collected the image and annotations, to avoid temporal leaks (Glockner et al., 2022). (2) It cannot return FC articles published by other FC organizations.

The answer to **Provenance** ("Was the image used before?") is either "Yes" when (partially) matching images are retrieved, and "No" or "Unknown" otherwise, the latter two being difficult to distinguish in practice. The answer is automatically derived from the retrieved results. Therefore, the Provenance evaluates the ability of retrievers to obtain evidence that contain previous versions of the same image. We define the Provenance score as the recall on images with ground truth answer "Yes". We do not penalize nor reward retrievers for collecting matching images when the ground truth is "No" or "Unknown".

Source and **Motivation** expect free-form answers, evaluated using the RougeL (Lin, 2004) and Meteor (Banerjee and Lavie, 2005) metrics.

The expected answers for **Date** are actual timestamps. We evaluate the answers using the Exact Match (EM) and Δ, a custom metric ranging from 1 for predictions that are close in time to the ground truth to 0 for those that are far. If the ground truth answer contains more than one date for the image, for example if it is a composite image, we compute Δ based on the smallest time difference that can be obtained by matching pairs of predicted and ground truth answers. Formally Δ is defined as

$$\Delta = \frac{1}{N} \sum_{(i,j) \in M} \frac{1}{1 + d_{i,j}},$$

where N is the number of ground truth elements, M is the set of matched prediction-ground truth pairs (i, j), and $d_{i,j}$ is the optimal distance between i and j, obtained with the Jonker-Volgenant algorithm (Jonker and Volgenant, 1987; Crouse, 2016). Years serve as the distance unit.

Location is evaluated with RougeL and Meteor. The subset of images where ground truth locations can be mapped to coordinates using GeoNames[1] is further evaluated in a spatial dimension. We employ variants of Δ, namely, Coordinates Δ (COΔ), with thousands kilometers as distance unit, and Hierarchical Location Δ (HLΔ), using the distance between the prediction and ground truth in terms of their GeoNames hierarchies.[2] For example, the distance for the pair (USA, Chicago) is 2, based on the hierarchy *USA-Illinois-Chicago*.

[1] geonames.org

[2] geonames.org/export/place-hierarchy.html

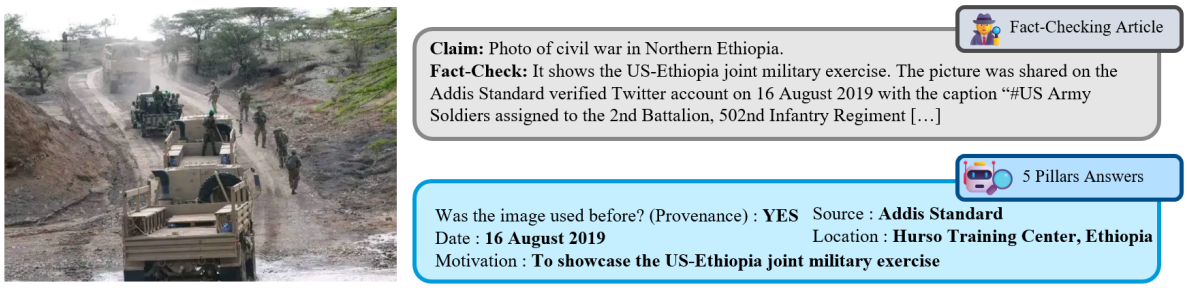

Figure 2: An annotated image of the 5Pils dataset. Above: the FC article. Below: the extracted 5 Pillars answers.

5 The 5Pils Dataset

The 5Pils dataset consists of 1,676 images annotated with 5 Pillars answers. We collect metadata for analysis purposes, including the *type of image*, the *verification strategy*, and the *verification tools*.

5.1 Data Collection

We selected FC organizations that met the following criteria : they write some of their articles in English, include the verified image in the article, do not systematically tag the images with labels, for example, marking the image with a "Fake" stamp, and are verified signatories of the International Fact-Checking Network (IFCN)[3] since Fall 2023, which guarantees that the images are verified according to the highest standards. None of the existing visual misinformation datasets met these combined criteria. However, three IFCN members did: Factly, PesaCheck, and 211Check, based in India, Kenya, and South Sudan, respectively. Hence, we created a new dataset with articles from these organizations. We collected all URLs archived on the Wayback Machine between January 2019 and December 2023.[4] We kept URLs that contain the keywords *photo*, *image*, or *picture*. This results in a collection of 3,424 online articles. We scraped the *title*, the *publication date*, the *text content*, and the fact-checked *image URL*. We downloaded the images and cleaned them by cropping social media sidebars. We applied a second round of filtering by removing articles written in French, Kannada, and Telugu, articles that are not verifying images despite the keyword match, articles duplicated from one organization to another, and articles with FC information embedded in the image. The resulting unlabeled dataset contains 1,692 FC articles.

[3] ifcncodeofprinciples.poynter.org/signatories
[4] archive.org/help/wayback_api.php

5.2 Data Annotation

We make the assumption that each article explicitly discusses one or more of the 5 Pillars. Hence, the annotation process becomes a text extraction task. Given its relative simplicity, we decided to automate the process using GPT4 (OpenAI, 2023). We tasked GPT4 to extract the 5 Pillars answers and relevant metadata as a json file using a custom prompt detailed in Appendix D. An annotated image is illustrated in Figure 2. In total, only 16 articles (<1%) did not contain any 5 Pillars answers, confirming our initial assumption. We recruited three students to evaluate the quality of the GPT4 annotations for a random sample of 50 images. After reading the corresponding FC article, they assigned a label to each of the GPT4 extracted answers: "Correct" if GPT4 detects the presence or absence of the answer in the article, "Missing" if GPT4 does not provide an answer while it was available in the article, and "Incorrect" if GPT4 provides a wrong answer. Taking their majority vote, annotators found 100, 100, 98, 96, and 94% of the answers to be "Correct" for Provenance, Source, Date, Location, and Motivation, respectively, the remaining cases being "Missing" answers. Restricting to unanimous vote for "Correct" answers, those percentages decrease to 94, 82, 90, 82, 76%. For the rest, at least one annotator has chosen the "Missing" (2, 10, 10, 12, 20%) or "Incorrect" (4, 8, 0, 6, 4%) label. "Missing" are preferable over "Incorrect" labels as they lower the recall without impacting the precision of the annotations. Overall, the high values of majority and unanimous vote for the "Correct" label confirm the quality of the automated annotations. To evaluate inter-annotator agreement, we use Randolph's κ (Randolph, 2005), a variant of Fleiss' κ (Fleiss, 1971) that is better suited for annotations that are heavily skewed towards one label (Warrens, 2010; Schlichtkrull et al., 2023a), "Correct" in this case. The Randolph's κ for each

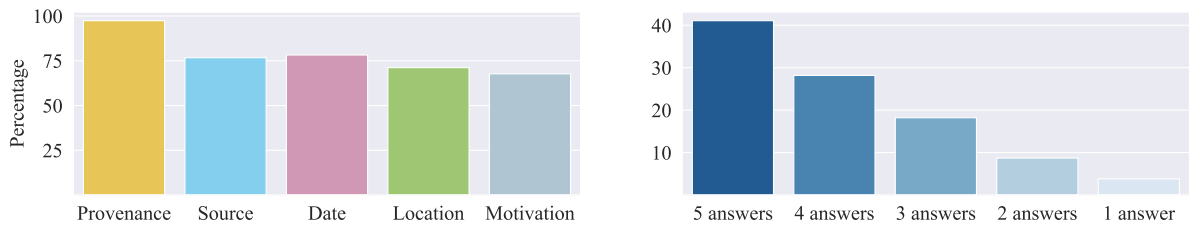

Figure 3: Left: Images with answers per pillar (%). Right: Images with n answered pillars (n=1,..,5) (%).

pillar is 0.94, 0.82, 0.87, 0.82, 0.76. Those scores show a strong inter-annotator agreement.

5.3 Dataset Statistics

5Pils contains 1,676 annotated images. This makes it slightly larger than other real-world visual misinformation datasets (Zlatkova et al., 2019; Papadopoulos et al., 2024). We split the dataset into a train, validation, and test set (60%, 10%, 30%). The split is performed along the time dimension to limit temporal leakage (Glockner et al., 2022; Schlichtkrull et al., 2023a). The sequential end dates are May 31st 2022, Sep. 20th 2022, and Dec. 28th 2023. Figure 3 shows the 5Pils answers statistics. All pillars are present in 68 % or more of the FC articles. Furthermore, answers are often present in combination, with 87% of the images having 3 or more answered pillars. These high values confirm the prevalence of image contextualization and the 5 Pillars approach in human FC practices. Appendix E reports the co-occurence of 5 Pillars answers. We searched for keywords in the FC articles referring to common strategies and tools for image contextualization. 86% of the FC articles explicitly refer to RIS, 1.94% to geolocation techniques, and 1.9% to keyword search. Google Image Search is the most popular RIS engine (36.7%), followed by Yandex Images (5.2%), and Bing Visual Search (3.8%). Additional analysis about the type of misinformation, and the spatial and temporal distribution of images is provided in Appendix F, G and H, respectively.

6 Experiments

6.1 Baseline model

To our knowledge, there is no prior method for automated image contextualization. Therefore, we introduce a first baseline for the task, illustrated in Figure 4. The baseline is a pipeline based on the best practices of human fact-checkers and the existing AFC literature (Akhtar et al., 2023).

Manipulation detection To detect images that have been manipulated, we fine-tune a vision transformer (ViT) (Dosovitskiy et al., 2021) on the train set. The input is the image and the target is the *type of image* metadata (manipulated, non-manipulated).

Evidence retrieval Collecting evidence from the open web requires a retriever. Given its popularity among human fact-checkers, we rely on the Google RIS engine.[5] We provide the image as input and obtain up to 37 web-pages, 6.84 on average, that used (partially) matching images. Each web-page acts as a single text evidence. We scrape their content with Trafilatura (Barbaresi, 2021).[6]

Evidence Ranking We select the top k=3 text evidence, which corresponds on average to selecting half the evidence matching a given image. We sort the evidence by the cosine similarity between the embedding of the image to verify and the text evidence embedding computed with CLIP (Radford et al., 2021).

Original identification This step is applied to manipulated images to identify the original, unaltered image among the RIS results. The intuition is that the alterations might have removed key elements to identify the original meta-context. As a simple heuristic, we select the retrieved image with the oldest publication date.

Answer generation We generate zero-shot answers with the following LLMs: Llama2 7B-Chat (Touvron et al., 2023), Llava 1.5 7B (Liu et al., 2023a,b), and GPT4 (OpenAI, 2023). We try different input modalities: only the image, only the text evidence, or both. Llama2 is only used with text evidence. For the multimodal setting, we experiment with few-shot predictions using 1 or 2 **demonstrations**. The prompt template and implementation details are provided in Appendix I and J, respectively.

[5] cloud.google.com/vision/docs/detecting-web
[6] trafilatura.readthedocs.io

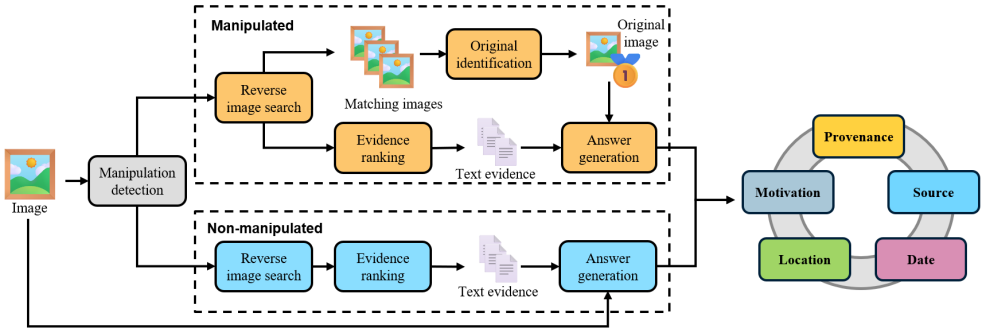

Figure 4: Our baseline for image contextualization. The retriever is a RIS engine that searches web-pages containing previous versions of the image. The top k evidence and the image are provided as input to a QA model to answer the 5 Pillars questions. Manipulated images go through the additional step of identifying the original unaltered image.

6.2 Main Results

We evaluate our baseline on the 5Pils test set. The **Provenance** score is 66.2%, meaning that the Google RIS engine retrieved scrapable web-pages with (partially) matching images for less than 2 out of 3 images for which matching images are known to be available. We attribute this low score to the use of a single RIS engine. In the absence of retrieved results, human fact-checkers typically use additional RIS engines or rely on keyword search, which we do not incorporate in this baseline.

Table 1 shows the results obtained for the **Source**, **Date**, **Location**, and **Motivation**. The best models achieve a promising performance, especially for Location and Motivation, with RougeL and Meteor close to or above 20%. Source performance is much lower. Predicting the Source can rarely be done with the image alone. Hence, it depends on the quality of the evidence ranking and retrieval. The latter being limited by the use of a single RIS engine. Obtaining a high EM for Date is challenging, especially when a specific day or month is expected. The best EM improves from 7.58 to 13.89 and 19.78 if we evaluate the match at the month and year level, respectively. There is more potential to improve Δ. The best value is 39.42%, or an average distance of one year and a half between the ground truth and the prediction. The best OCΔ and HLΔ scores for Location are promising too, indicating an average distance of a thousand kilometers and an average hierarchical distance of 1.5, respectively. These results confirm that image contextualization can be automated to some extent. At the same time, they inform us that a simple baseline that only considers RIS engines as open web retrievers and one-step answer generation with LLMs is not sufficient to solve the task.

6.3 Quantitative analysis

How do different LLMs compare on the task? For image input, GPT4 slightly outperforms Llava. The difference between Llama2 and GPT4 is smaller for text input. Llama2 even outperforms GPT4 for Location. GPT4 and Llava achieve similar performance with multimodal input. Llava and Llama2 score better Δs than GPT4 because they always provide an estimate while GPT4 often refrains from answering.

What is the preferred input modality? For Source and Motivation, a multimodal input gives better results. For Date and Location, a multimodal input is equal or worse to text evidence, implying that the multimodal LLMs have limited abilities to reason about the dates and locations of images, and their visual understanding can even override correct information present in the text evidence.

What is the impact of demonstrations? Demonstrations improve the performance for the Source with GPT4, but decrease it with Llava. They help Llava for Location and Motivation by providing additional guidance on the expected type of answer. Their impact for Date is negative for both models. A manual investigation reveals that the models replace some correct zero-shot Date answers by one of the demonstration's answer.

How performant is the evidence ranking? We use as target the ranking that sorts evidence according to their n-gram overlap with the ground truth answer. We evaluate the ranking with the normalized discounted cumulative gain (nDCG) metric (Järvelin and Kekäläinen, 2002). We compare the CLIP ranking with a baseline that sorts evidence by chronological order of publication (Time ranking). The CLIP ranking outperforms the Time ranking for all pillars, as shown in Table 2.

		Source		Date		Location				Motivation	
Method	**N**	**RougeL**	**Meteor**	**EM**	**Δ**	**RougeL**	**Meteor**	**HLΔ**	**COΔ**	**RougeL**	**Meteor**
Image only											
Llava	0	$1.16_{\pm 0.2}$	$0.67_{\pm 0.1}$	$0.00_{\pm 0.0}$	$13.49_{\pm 0.3}$	$5.79_{\pm 0.2}$	$3.79_{\pm 0.2}$	$12.74_{\pm 0.5}$	$16.99_{\pm 0.8}$	$7.12_{\pm 0.3}$	$3.33_{\pm 0.2}$
GPT4	0	$2.78_{\pm 0.0}$	$2.37_{\pm 0.1}$	$0.17_{\pm 0.1}$	$5.68_{\pm 2.2}$	$15.73_{\pm 0.3}$	$15.68_{\pm 0.2}$	$22.15_{\pm 0.5}$	$27.31_{\pm 0.6}$	$14.81_{\pm 0.1}$	$11.08_{\pm 0.7}$
Text evidence only											
Llama2	0	$2.83_{\pm 0.0}$	$4.47_{\pm 0.0}$	$6.06_{\pm 0.0}$	$37.73_{\pm 0.0}$	$\mathbf{28.65}_{\pm 0.0}$	$\mathbf{29.65}_{\pm 0.0}$	$39.96_{\pm 0.0}$	$44.81_{\pm 0.0}$	$14.89_{\pm 0.0}$	$15.65_{\pm 0.0}$
GPT4	0	$4.18_{\pm 0.0}$	$3.03_{\pm 0.0}$	$\mathbf{7.58}_{\pm 0.0}$	$23.71_{\pm 0.2}$	$24.95_{\pm 3.1}$	$23.63_{\pm 3.0}$	$28.04_{\pm 3.1}$	$31.35_{\pm 2.9}$	$15.23_{\pm 0.2}$	$12.58_{\pm 0.0}$
Image + Text evidence											
Llava	0	$6.73_{\pm 0.1}$	$5.00_{\pm 0.1}$	$5.06_{\pm 0.4}$	$\mathbf{39.42}_{\pm 0.1}$	$25.52_{\pm 0.3}$	$19.48_{\pm 0.3}$	$33.08_{\pm 0.4}$	$40.32_{\pm 0.5}$	$8.71_{\pm 0.1}$	$4.21_{\pm 0.1}$
	1	$6.69_{\pm 0.2}$	$5.08_{\pm 0.1}$	$4.22_{\pm 0.3}$	$34.96_{\pm 0.2}$	$25.63_{\pm 0.6}$	$21.53_{\pm 0.6}$	$37.06_{\pm 0.7}$	$44.92_{\pm 0.5}$	$19.25_{\pm 0.1}$	$14.09_{\pm 0.2}$
	2	$5.96_{\pm 0.2}$	$4.65_{\pm 0.1}$	$4.86_{\pm 0.4}$	$37.67_{\pm 0.1}$	$26.14_{\pm 0.5}$	$21.79_{\pm 0.3}$	$\mathbf{40.27}_{\pm 0.4}$	$\mathbf{49.52}_{\pm 0.6}$	$19.33_{\pm 0.3}$	$13.36_{\pm 0.2}$
GPT4	0	$6.54_{\pm 0.3}$	$5.31_{\pm 0.4}$	$6.31_{\pm 0.2}$	$19.92_{\pm 0.7}$	$27.63_{\pm 0.6}$	$26.71_{\pm 0.4}$	$32.20_{\pm 0.8}$	$36.82_{\pm 0.9}$	$18.58_{\pm 0.3}$	$15.53_{\pm 0.1}$
	1	$7.19_{\pm 0.2}$	$5.61_{\pm 0.2}$	$6.23_{\pm 0.1}$	$21.32_{\pm 1.4}$	$28.15_{\pm 0.5}$	$26.47_{\pm 0.6}$	$31.59_{\pm 0.3}$	$36.22_{\pm 0.4}$	$19.17_{\pm 0.1}$	$16.58_{\pm 0.0}$
	2	$\mathbf{8.06}_{\pm 0.1}$	$\mathbf{6.46}_{\pm 0.2}$	$5.56_{\pm 0.3}$	$19.00_{\pm 0.1}$	$27.34_{\pm 1.5}$	$26.34_{\pm 1.6}$	$30.57_{\pm 1.3}$	$34.81_{\pm 1.5}$	$\mathbf{19.96}_{\pm 0.2}$	$\mathbf{17.02}_{\pm 0.0}$

Table 1: Zero- and few-shot results for answer generation (%). **N** is the number of demonstrations. Reported results are averages over 3 iterations with standard deviations. The best scores are marked in **bold**.

	Source	**Date**	**Location**	**Motivation**
Time	40.15	61.9	55.75	64.18
CLIP	**43.55**	**62.17**	**61.4**	**71.92**

Table 2: nDCG of different ranking methods (%). Time is the chronological ranking by publication date.

	Source		Date		Location				Motivation	
Method	**R**	**M**	**EM**	**Δ**	**R**	**M**	**HLΔ**	**COΔ**	**R**	**M**
No detector	6.3	5	8.3	21.8	42.6	39.7	**43.1**	**49.1**	20	18.5
Detector	6	4.7	8.3	23	44.7	39.8	**43.1**	**49.1**	20.8	19
Perfect detector	6.3	4.7	**10**	**24**	42.9	39.6	**43.1**	**49.1**	21.6	19.7
Oracle	**9.3**	**6.5**	**10**	22.7	**45.8**	**40.4**	40.6	47.6	**22.2**	**20.4**

Table 3: Zero-shot multimodal GPT4 results for the manipulated images. R is RougeL and M is Meteor.

What is the impact of the separated pipeline for manipulated images? Applying the fine-tuned ViT and the publication date heuristic, we retrieve a candidate unaltered version for 58 images. With multimodal zero-shot GPT4, little to no gains are observed over a pipeline that ignores these steps (*No detector*), as shown in Table 3. The results are slightly better with a perfect accuracy manipulation detector (*Perfect detector*), except for Source and Location. The RIS results do not often contain the unaltered image, partly explaining these minor performance gains. As an upper bound, we consider an *oracle* which provides the original versions of the image reported in the FC article, if available online, and observe a larger improvement, especially for Source and Location. This indicates that although the current pipeline is imperfect, it is meaningful to develop methods that replace a manipulated image with the original, unaltered version. Future work should consider alternatives to RIS for retrieving the original version, or image editing models to remove manipulated content, obtaining an image close to the unaltered version.

6.4 Qualitative analysis

We randomly sample 100 test images and manually categorize the predictions of multimodal zero-shot GPT4. Four examples are provided in Figure 5. 31.9% of the answers are (partially) correct. We classify the other answers into six error categories. The distribution of the correct and error categories is shown in Figure 6. We break down this distribution per pillar in Appendix K. 20.4% of the answers accurately state that not enough information is available in the image and text evidence, for example, in the 3rd image of Figure 5, the Source cannot be identified. For 10.2% of the answers, the response is the same, but sufficient information was actually available. This problem arises for the 4th image, where useful evidence is directly available but unexploited. In the 3rd image, while the exact date is not given, the information on the event ("Coronation of Jean-Bedel Bokassa") should be sufficient to predict the date based on world knowledge. For 17.9% of the answers, the response is wrong and based on irrelevant evidence. Wrong answers based on wrongly interpreted image content are comparatively lower, at 4.2%. 13.7% of the answers are not generated because of the GPT4 content filter. This mainly prevents GPT4 from contextualizing images of conflicts and violent incidents. Given the significant proportion of these images in FC articles, this constitutes a serious limitation to the use of GPT4. Lastly, 1.8% of the answers contain misinformation elements.

From this analysis, we conclude that increasing the number of evidence retrieved from the open web constitutes the first direction to improve performance. The lack of evidence constitutes a regular challenge for human fact-checkers too. In the

Image	Text evidence	Image contextualization
A collapsed building following an earthquake in Karnah tehsil […]	Date : 2023-02-06 […] earthquake in southeastern Turkey […] kills two Armenians in Aleppo [...] कश्मीर में भूकंप के बाद क्षतिग्रस्त हुई इमारत की नहीं है ये तस्वीर […]. Author: ABP Live (This photo is not of the building damaged after the earthquake in Kashmir […]) Title: The M7.8 Turkey Earthquake Happened While Some People Were Sleeping […]	Was the image used before? Yes Source: MPTL | GT: Anadolu's journalist Firat Ozdemir Date: 2023-02-06 Location: Maraş, Adana, Diyarbakır, Hatay, Urfa, Malatya, Kilis | GT: Kahramanmaras Motivation: show destruction caused by the 7.7 magnitude earthquake in Turkey | GT: To report on the earthquake in […] Turkey
Ethiopian National Defense Forces […] Tigray region in October 2022	Ethiopian government has a special responsibility to de-escalate Tigray war. Author: David Thomas. Date: 2022-10-24. […] October 24th, 2022, Opinion by David Thomas Image : Amanuel Sileshi/AFP After a five-month truce, fighting has resumed in Tigray – timenews […] Ethiopian National Defense Force (ENDF) soldiers board a truck in Wichale, Ethiopia on December 13, 2021.	Was the image used before? Yes Source: ENDF | GT: AFP Photographer Amanuel Sileshi Date: 2021-12-14 | GT: 2021-12-13 Location: Wichale | GT: Wuchale, Amhara regional state Motivation: To show Ethiopian soldiers boarding a truck in Wichale, Ethiopia | GT: To report on the conflict in the Amhara region, Ethiopia
Kenyan opposition leader Raila Odinga being crowned a Nigerian elder	Instagram.com […] "Coronation of Jean-Bedel Bokassa as Emperor of the Central African Empire, 1977 […] Title: Emperor Jean-Bedel Bokassa , seated on his tasteless golden throne for his coronation, 1977 […] Date: 2018-07-12 […] […] If being asked to write music for the coronation of a king is an honour, then doing it for an emperor […]	Was the image used before? Yes Source: Cannot be determined | GT: Pierre Guillaud Date: No specific date | GT: 1977-12-04 Location: Central African Empire | GT: Bangui, Central African Republic
EAF attacks on Tigray cities and weapon warehouses	Date : 2019-03-07 […] two AIM-9X Air-to-Air missiles, completes refueling prior to a training mission over the W-291 training area in southern California, March 6 […] (U.S Marine Corps photo by Sgt. Dominic Romero) Sitename: © 2024 Alamy Limited Title: An F/A-18 Hornet with Marine Fighter Attack Squadron (VMFA) […] training mission over the W-291 training area in southern California, March 6	Was the image used before? Yes Source: Sgt. Dominic Romero | GT: Alamy Date: Not enough information | GT: 2019-03-06 Location: In the air, over the ocean | GT: Southern California, USA Motivation: aviation-related purposes, possibly to showcase the aircraft and its capabilities | GT: to show a U.S. Marine Corps F/A-18 Hornet fight aircraft […] during a training mission

Figure 5: Images from the 5Pils test set with zero-shot multimodal GPT4 predictions. The middle column shows excerpts of the evidence texts with relevant snippets. The right column shows correct, partially correct, and wrong predictions. GT is the ground truth answer.

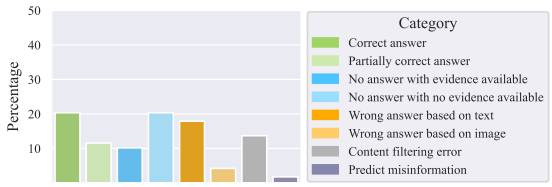

Figure 6: Qualitative analysis of zero-shot multimodal GPT4 answers (%).

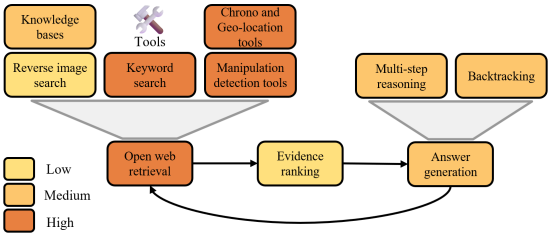

Figure 7: Open challenges in image contextualization, ranked by level of complexity.

absence of RIS results, they turn to their world knowledge and keyword search to gather evidence, as illustrated in Appendix C. For the Date and Location pillars, this can be complemented by chrono- and geolocation tools.

Planning and combining the output of those tools constitute an interesting research direction for multimodal LLMs. As the number of evidence grows, the reasoning can be improved by breaking it down in several steps, interleaving evidence retrieval and reasoning, and implementing backtracking mechanisms (Hu et al., 2023b). Figure 7 shows open challenges for the task, ranked by estimated level of complexity.

7 Conclusion

In this work, we fill an important gap in the field of visual misinformation by introducing the task of automated image contextualization. We create 5Pils, a dataset of 1,676 fact-checked images with meta-context QA pairs based on the 5 Pillars framework. We define a first baseline for the task, which combines the image with textual evidence retrieved from the open web. Despite achieving promising performance, several challenges remain. Directions for future work include better open web retrievers that combine multiple tools, improved answer generation based on multi-step reasoning, and better methods for removing image manipulations.

8 Limitations

The 5Pils dataset has multiple limitations.

Firstly, we restricted data collection to articles written in English. While the 5Pils images are spread all over the world, this restriction results in underrepresentation of certain regions, such as Latin America and Oceania. On the other hand, our dataset contains many images from Sub-Saharan Africa, a region underrepresented in previous misinformation datasets, compared to North America, South Asia, and China (Zlatkova et al., 2019; Hu et al., 2023a; Suryavardan et al., 2023).

Secondly, while we implemented a filter to remove evidence from FC articles, we cannot ensure a perfect recall, especially if the articles were not published by FC websites. In the 1st example of Figure 5, one of the evidence is a FC article that passed our filter.

Thirdly, the dataset contains a majority of images that have effectively been used for misinformation purposes. This is a common limitation when relying on real-world FC articles (Zlatkova et al., 2019; Schlichtkrull et al., 2023a), as human fact-checkers invest more time in writing articles about false claims than true claims.

Lastly, the metrics for Source are slightly pessimistic with regard to the model's performance. If the prediction is more specific than the ground truth, no or partial credit is given. On the 4th example of Figure 5, GPT4 predicts accurately the photographer's name, while the FC article only reports the website where the image can be found. This results in a score of 0, while the prediction is actually better than the human-written ground truth.

Regarding our baseline, we find two limitations.

Firstly, not all RIS results can be scraped from the open web. Evidence from social media platforms is typically not accessible. In some cases, image contextualization requires to access such social media posts. This constitutes an important difference with human fact-checkers, who can access social media posts from their browser.

Secondly, our baseline does not assess the reliability of the retrieved evidence. Hence, there is no guarantee that the information they provide is accurate, and different evidence might contradict each other. Assessing reliability and using it to filter and rank text evidence constitutes an interesting direction for future work.

9 Ethics statement

Automated image contextualization is designed as a new tool to help fact-checkers verify the vast quantities of visual misinformation circulating on the internet. Automated approaches should respect the same code of practice as professional fact-checkers, which includes the respect of online privacy. In particular for the Source pillar, techniques for automated profiling on social media should not be used for automated image contextualization. Care should also be taken when publishing the results, as they could put the Source at risk.

The 5Pils dataset was collected from publicly available FC organizations websites. Due to the real-world nature of the data, events covered include wars and conflicts. As a result, some images contain graphic violent content. When collecting the data, we decided not to filter out images with violent content, in order to cover the actual distribution of images that our target users, professional fact-checkers, would want to provide as input. Given the graphic nature of some images, we do not release them directly. Instead, we do publicly release the URLs of the FC articles and extracted evidence, as well as the script that allows to collect and process the images and the evidence text.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14940–14949.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2023. Cosmos: Catching out-of-context image misuse using self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14084–14092.

Phoebe Arnold. 2020. The challenges of online fact checking: how technology can (and can't) help. Accessed: 2023-09-16.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Pro-*

ceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

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.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Megha Chakraborty, Khushbu Pahwa, Anku Rani, Shreyas Chatterjee, Dwip Dalal, Harshit Dave, Ritvik G, Preethi Gurumurthy, Adarsh Mahor, Samahriti Mukherjee, Aditya Pakala, Ishan Paul, Janvita Reddy, Arghya Sarkar, Kinjal Sensharma, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY3M: A benchmark for multimodal fact verification with explainability through 5W question-answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15282–15322, Singapore. Association for Computational Linguistics.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David F. Crouse. 2016. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696.

Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. 2021. Edited media understanding frames: Reasoning about the intent and implications of visual misinformation. 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 2026–2039, Online. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Dudfield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, et al. 2024. Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild. *arXiv preprint arXiv:2405.11697*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023a. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2901–2912, New York, NY, USA. Association for Computing Machinery.

Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David Ross, Cordelia Schmid, and Alireza Fathi. 2023b. Avis: Autonomous visual information seeking with large language model agent. In *Advances in Neural Information Processing Systems*, volume 36, pages 867–878. Curran Associates, Inc.

Mingzhen Huang, Shan Jia, Ming-Ching Chang, and Siwei Lyu. 2022. Text-image de-contextualization detection using vision-language models. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8967–8971.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117.

Ayush Jaiswal, Ekraam Sabir, Wael AbdAlmageed, and Premkumar Natarajan. 2017. Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1465–1471.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

R Jonker and A Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Sohail Ahmed Khan, Jan Gunnar Furuly, Henrik Brattli Vold, Rano Tahseen, and Duc-Tien Dang-Nguyen. 2023a. Online multimedia verification with computational tools and osint: Russia-ukraine conflict case studies. *arXiv preprint arXiv:2310.01978*.

Sohail Ahmed Khan, Ghazaal Sheikhi, Andreas L. Opdahl, Fazle Rabbi, Sergej Stoppel, Christoph Trattner, and Duc-Tien Dang-Nguyen. 2023b. Visual user-generated content verification in journalism: An overview. *IEEE Access*, 11:6748–6769.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.

Ziyi Kou, Daniel Yue Zhang, Lanyu Shang, and Dong Wang. 2020. Exfaux: A weakly supervised approach to explainable fauxtography detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 631–636.

Stephan Lewandowsky, John Cook, Ullrich Ecker, Dolores Albarracín, Panayiota Kendeou, Eryn J Newman, Gordon Pennycook, Ethan Porter, David G Rand, David N Rapp, Jason Reifler, Jon Roozenbeek, Philipp Schmid, Colleen M Seifert, Gale M Sinatra, Briony Swire-Thomposon, Sander van der Linden, Thomas J Wood, and Maria S Zaragoza. 2020. The debunking handbook 2020. Accessed: 2023-09-19.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Huan Liu, Zichang Tan, Qiang Chen, Yunchao Wei, Yao Zhao, and Jingdong Wang. 2023c. Unified frequency-assisted transformer framework for detecting and grounding multi-modal manipulation. *arXiv preprint arXiv:2309.09667*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Annique Mossou and Ross Higgins. 2021. A beginner's guide to social media verification. Accessed: 2023-09-15.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 16–25, New York, NY, USA. Association for Computing Machinery.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. 2023a. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, MAD '23, page 36–44, New York, NY, USA. Association for Computing Machinery.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2023b. Red-dot: Multimodal fact-checking via relevant evidence detection. *arXiv preprint arXiv:2311.09939*.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4.

Kha-Luan Pham, Manh-Thien Nguyen, Anh-Duy Tran, Minh-Son Dao, and Duc-Tien Dang-Nguyen. 2023. Detecting cheapfakes using self-query adaptive-context learning. In *Proceedings of the 4th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 60–63.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. *arXiv preprint arXiv:2403.03170*.

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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Justus J. Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. In *Joensuu Learning and Instruction Symposium*. ERIC.

Anku Rani, S.M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY-5WQA: 5W aspect-based fact verification through question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10421–10440, Toronto, Canada. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023a. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167. Curran Associates, Inc.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023b. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Craig Silverman. 2013. Verification handbook. Accessed: 2023-09-15.

S Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. 2023. Factify 2: A multimodal fake news and satire news dataset. *arXiv preprint arXiv:2304.03897*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shaydanay Urbani. 2020. Verifying online information. Accessed: 2023-09-15.

Matthijs J Warrens. 2010. Inequalities between multirater kappas. *Advances in data analysis and classification*, 4:271–286.

Jing Yang, Didier Vega-Oliveros, Taís Seibt, and Anderson Rocha. 2022. Explainable fact-checking through question answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8952–8956. IEEE.

Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 2023. Support or refute: Analyzing the stance of evidence to detect out-of-context mis- and disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4268–4280, Singapore. Association for Computational Linguistics.

Fanrui Zhang, Jiawei Liu, Qiang Zhang, Esther Sun, Jingyi Xie, and Zheng-Jun Zha. 2023. Ecenet: Explainable and context-enhanced network for mutimodal fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 1231–1240, New York, NY, USA. Association for Computing Machinery.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

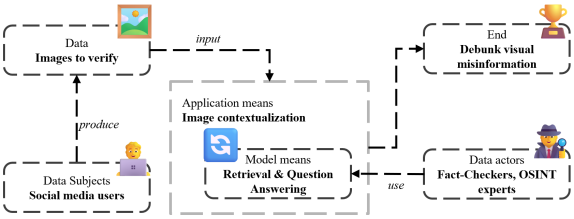

Figure 8: Image contextualization narrative diagram. Figure based on Schlichtkrull et al. (2023b).

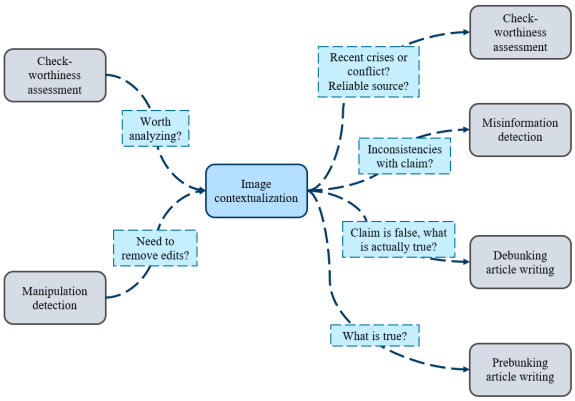

Figure 9: Connections of image contextualization with other components of the human fact-checking pipeline.

A Analysis of the FC narrative

Following the recommendations of Schlichtkrull et al. (2023b), we conduct an analysis of the FC narrative introduced in this paper. Figure 8 provides a diagram of the different epistemic elements of our narrative. We envision fact-checkers as the primary data actors that will use our automated retrieval and QA pipeline (model means) to predict the original meta-context of images (application means) and debunk visual misinformation (ends). Other potential data actors include Open Source Intelligence (OSINT) experts. The data subjects are primarily social media users who will share image content that might be out-of-context, manipulated, or fake. Figure 9 shows the interactions of image contextualization with other components of the human FC pipeline. We identify two components at the input of image contextualization. (1) Assessing check-worthiness saves computational resources by filtering out content that is unlikely to be visual misinformation. (2) Detecting manipulations informs whether any edits need to be removed to obtain the original image. For the output, we identify four components, which are covered in Section 1. In this work, we integrate manipulation detection in our automated baseline, while other tasks at the input and output are expected to be performed by human fact-checkers. Integrating image contextualization with other automated FC tasks is an interesting direction for future work.

(a) Manipulated image

(b) Original image

(c) Fake image

Figure 10: A manipulated image with edited television background (a), and the original version (b). A digital artwork of a giant skeleton (c).

B Contextualization of manipulated and fake images

We provide examples of manipulated and fake images in Figure 10. The manipulated image (a) claimed that footballer Cristiano Ronaldo was showing support for the Argentine football team.[7] By collecting RIS results, it was first observed that the image had been manipulated (b). Additional meta-context elements informed the fact-checkers that the image was taken prior in time, when the footballer showed support to the Portugal team, and originated from his official social media account. Image (c) claimed to show the discovery of giant human skeletons.[8] By searching the Source and Motivation pillars, it was discovered that the image was created by an artist for a digital artwork competition.

C Image contextualization with keyword search

It can happen that RIS engines do not return any results for a given image. In those cases, human fact-checkers use other methods to retrieve evidence, including keyword search. We provide two examples in Figure 11, corresponding to the 2nd and 4th images of Figure 5.

Left image: The image contains visual cues,

[7] factly.in/this-post-shares-an-edited-image-to-claim-that-cristiano-ronaldo-is-supporting-argentina

[8] factly.in/a-digitally-created-artwork-falsely-shared-as-a-real-picture-of-a-giant-human-skeleton

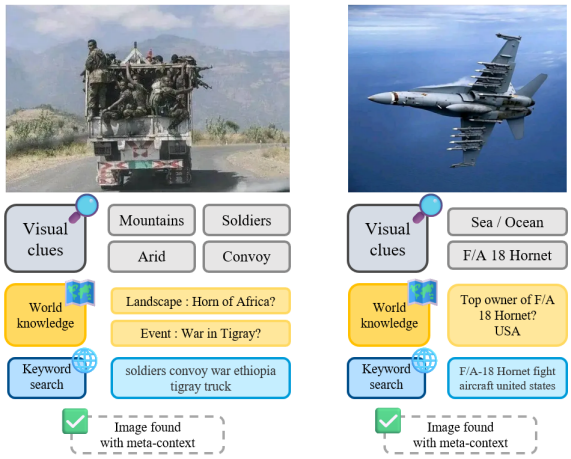

Figure 11: Contextualization with keyword search.

namely, a military convoy, a semi-arid landscape, and mountain chains in the background. Assuming sufficient world knowledge, we know that this landscape is characteristic of the Horn of Africa, and the main recent conflict in the area was the war in Tigray. Hence, we make the hypothesis that the image is related to the war in Tigray. We run the web search query "soldiers convoy war ethiopia tigray truck". The 10th result is the same image hosted on the platform Getty, captioned with all relevant meta-context.

Right image: the plane is recognizable as a Boeing F/A-18 Hornet fighter aircraft. Searching Wikipedia for information, we find that the top owner of F/A-18 is the US Airforce. We run the following web search query : "F/A-18 Hornet fight aircraft united states", and retrieve the target image among the first results on a webpage that contains all the meta-context information.

D Automated annotations prompt

Figure 12 shows the instructions given to GPT4 to extract the 5 Pillars answers from FC articles. The prompt was improved iteratively over batches of 10 articles, until a satisfactory performance was achieved. The prompt instructs GPT4 to create a json file containing various fields. If the information is not available, the standard answer should be "Not enough information". Figure 13 reports the output for the image of Figure 1. The *claimed location*, *claimed date*, *claimant*, and *claimant motivation* field are requested to make sure that GPT4 clearly distinguishes the claimed context of the image from its original meta-context. The *verification strategy* was initially extracted by querying GPT4.

Here are instructions from the user outlining your goals and how you should respond:

Fact Parser specializes in creating structured JSON reports for analyzing viral images, focusing on detailed fact-checking. The JSON report includes the following key entries: "URL", "claim", "publication date", "verdict", "was the photo used before?", "claimed location" (physical place), "real location" (physical place), "claimed date", "real date", "claimant" (the person making the current claim about the image), "source" (the original author of the image, do not invent if not specified),"claimant motivation" (the reason behind the claimant's use or portrayal of the image), "motivation" (why the original author captured the image), "type of image" (True, Out-of-Context, Fake, Manipulated), and "verification strategy". When specific details are not available in a fact-checking article, Fact Parser will note "Not Enough Information" for that entry. Responses are presented in a formal, technical style, ensuring accuracy and clarity.

Figure 12: Prompt for GPT4 annotations of FC articles.


```
{
    "URL": "https://factly.in/2013-evacuation-image-from-typhoon
            -hit-philippines-is-passed-off-as-iaf-airlifting-800-people-from-kabul ",
    "claim": "The image shows 800 people airlifted by the Indian Air Force from Kabul.",
    "publication date": "2021-08-17T12:00:55+00:00 ",
    "verdict": "Misleading ",
    "was the photo used before? ": "Yes ",
    "claimed location": "Kabul, Afghanistan ",
    "real location": "Tacloban city, Philippines ",
    "claimed date": "2021 ",
    "real date": "2013 ",
    "claimant": "Unspecified social media user ",
    "source": "US Airforce's official website ",
    "claimant motivation": "To attribute the evacuation efforts to the Indian Air Force ",
    "motivation": "To document the US Airforce\u2019s operation Damayan
            evacuation in  2013 ",
    "type of image": "Out-of-Context ",
    "verification strategy": "The fact-checking article performed a reverse image search
            which led to the real source of the image - US Airforce\u2019s official
            website from 2013."
}
```


Figure 13: Example of GPT4 annotations.

Later on, it was replaced by a keyword matching over the original article. GPT4 sometimes extracts a list of strings, for example when an image is composite. In those cases, we merge the list items into a string. The data labeling was performed using the Azure OpenAI service with API version *2023-10-01-preview*. The content filter was turned off to allow GPT4 to label FC articles that discuss conflicts and other events that involve violence.

E Co-occurrence of 5 pillar answers

Figure 14 shows the percentage of images for which two pillars are answered together. The percentages are coherent with the total occurrence of each pillar in the dataset. For example, the Date-Source is higher than the Date-Location co-occurrence, because there are mo images with Source answers than images with Location answers. No specific over- or under-representation of a pair is observed.

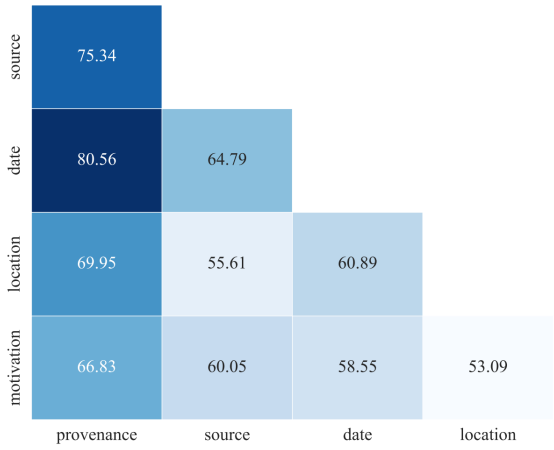

Figure 14: Co-occurrence of 5 Pillars answers in the dataset (%).

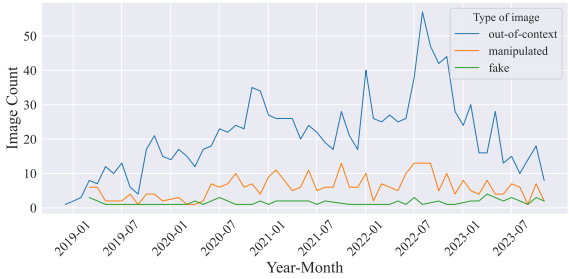

Figure 15: Time evolution of the type of misinformation in the dataset.

F Type of misinformation over time

Figure 15 displays the type of images verified by the FC organizations over time. Despite the rise of generative AI technologies, Fake images still represent a stable and low share of the verified content. Out-of-context consistently remains the most popular type of visual misinformation. The peak in out-of-context images in the summer of 2022 corresponds to the Kenyan general elections and the end of the conflict in Tigray.

G Spatial analysis

We look deeper into the Location pillar and investigate whether there are any noticeable geographic patterns in the way images from one location are used in a claim about another location. Using Nominatim,[9] we obtain coordinates for most of the claimed and real locations of the images in the dataset. Figures 16 shows the spatial distribution of all images that could be mapped to Nominatim coordinates. Large dots at the geographic center of a country indicates that the location is the country itself. The dataset covers a large diversity of real locations. East Africa and South Asia are predominant. This is expected because the FC organizations from which we collected the data are located in India, Kenya, and South Sudan. Despite that, Europe, and North America, and other parts of Africa get important coverage too. Interestingly some countries have few or no claims about them, but their images are nevertheless frequently used out-of-context. This includes China, Syria, Irak, and Spain. In a second analysis, we study the dispersion of real locations. To do this, we looked at one specific world event, where the claimed locations are concentrated in a limited geographic area. We plotted all verified images claiming to be associated with the war in Tigray, a conflict that took place between 2020 and 2022 in the Tigray region of Ethiopia, close to Eritrea. The resulting map is shown in Figure 17. While the claimed locations are concentrated in Northern Ethiopia and Eritrea as expected, only a part of the real locations are located there too, while the rest is spread out over the globe. Images can be found as far away as the United States. Those are mainly images of drones and aircraft training exercises. This large geographical dispersion illustrates the challenge and importance of predicting the Location pillar.

[9] geopy.readthedocs.io

H Temporal analysis

We study the distribution of the real image dates for a given claimed date. We report the distributions for images claimed to be from 2020, 2021, 2022, and 2023 in Figure 18. Most real dates are concentrated in the 3 to 4 years that include and precede the claimed date. A downward trend is then observed going back in time. This tells us that visual misinformation is often based on recent images. This information could be used as a heuristic to guide the prediction of the Date pillar.

I Few-shot prompt template

Figure 19 shows the few-shot prompt template provided to GPT4 for multimodal answer generation. Demonstration and evidence are numbered. The zero-shot prompt follows the same structure, except that the demonstrations are removed. Different instructions are provided when the input modality is the image ("You are given an image. Your task is to answer a question about the image.") or the text evidence ("You are given online articles that used a certain image. Your task is to answer a ques-

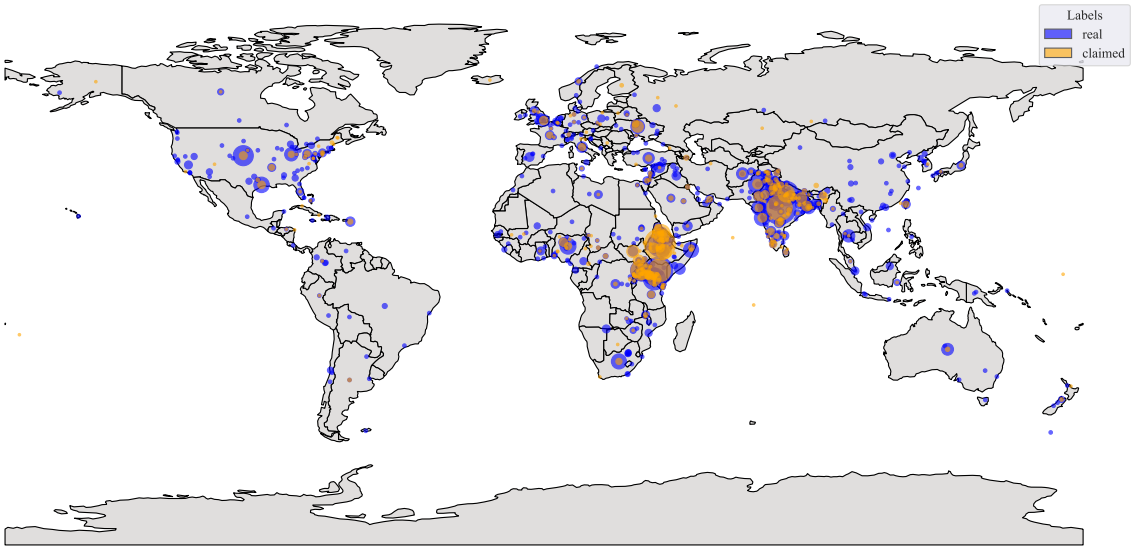

Figure 16: **Claimed** and **real** locations of 5Pils images. The size of the circle is proportional to the number of images with these coordinates.

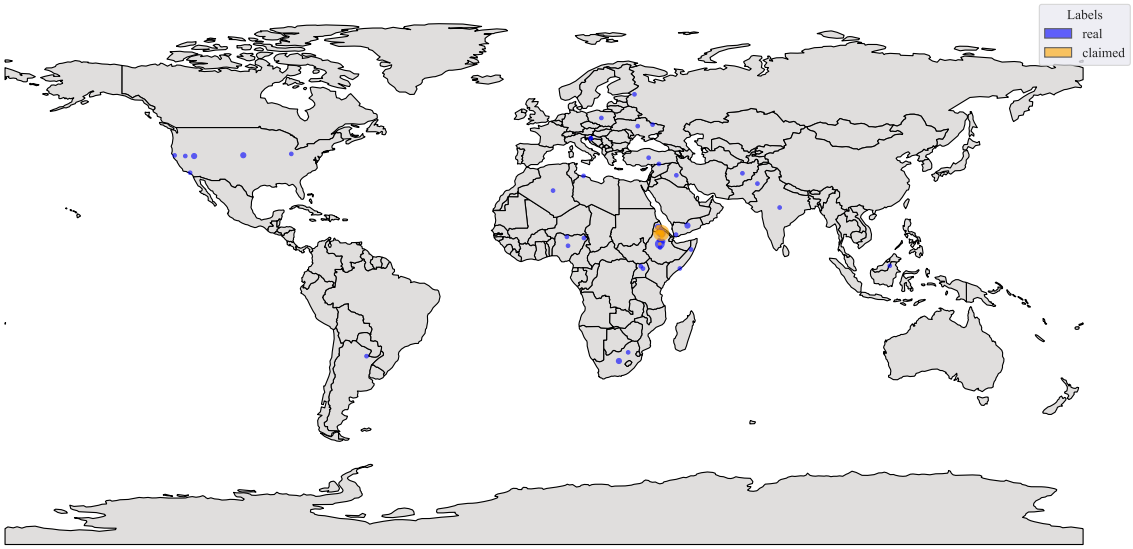

Figure 17: **Claimed** and **real** locations of images with the conflict in Tigray as claimed context. The size of the circle is proportional to the number of images with these coordinates.

Figure 18: The distribution of real dates over time for images with a given claimed date, in years. The orange bar indicates images for which the claimed date matches the real date.

Demonstration 1

You are given online articles that used a certain image. Your task is to answer a question about the image.

Evidence 1 [Text]
[…]

Question : […]

Answer : […]

Demonstration 2

[…]

You are given an image and online articles that used that image. Your task is to answer a question about the image using the image and the articles.

Evidence 1 [Text]

[…]

Question : […]

Answer :

Figure 19: Few-shot prompt template for GPT4 with multimodal input.

tion about the image.”). Adjustments are made to the prompts for Llama2 and Llava to include the special tokens expected by those LLMs. For Llava, every prompt starts with “USER: ” and “Answer: ” at the end of the GPT4 prompt is replaced by “ASSISTANT: ”. For Llama2, every prompt starts with “<s>[INST] «SYS» You are given [...] Your task is to answer a question about the image. «/SYS»”, and ends with “[/INST]”.

J Implementation details

Manipulation detection The ViT model is fine-tuned for 10 epochs on the train set with a learning rate of 2e-4.[10] It achieves an F1 score of 51% on the validation set, and 40% on the test set.

Evidence retrieval We set the maximum number of URLs to return by the Google RIS engine to 50. We scrape the following fields of the web-pages using Trafilatura: the *title*, *description*, *author*, *hostname*, *sitename*, and *publication date*. We also retrieve the *image* and its *caption* if available using a custom script. In total, we retrieved and scraped 2284 web-pages for the test set. 64.1% of the evidence is written in English.

Evidence ranking Embeddings are computed with a pre-trained multilingual CLIP model (Radford et al., 2021; Carlsson et al., 2022).[11] Due to budget constraints regarding the use of GPT4, we fixed the maximum number of evidence to use in the prompt k to 3. We report in Table 4 the results obtained on the validation set for different values of k with zero-shot multimodal Llava. The results suggest that increasing k does not always improve the performance, as shown by the Location scores that decline when k increases from 5 to 10. We conclude that better performance can be obtained by optimizing k for each pillar separately.

Answer generation Demonstrations are selected with a nearest neighbor strategy (Liu et al., 2022). Given a image from the test set, we compute the similarity with images in the train set using CLIP, and select the N nearest neighbors as demonstrations. For all LLMs, we set the temperature to 0.2

[10] huggingface.co/google/vit-base-patch16-224

[11] huggingface.co/M-CLIP/XLM-Roberta-Large-Vit-L-14

	Source		Date		Location				Motivation	
k	RougeL	Meteor	EM	Δ	RougeL	Meteor	HLΔ	COΔ	RougeL	Meteor
1	**6.5**$_{\pm 0.2}$	**4.2**$_{\pm 0.2}$	2.1$_{\pm 0.6}$	42.7$_{\pm 0.2}$	27.3$_{\pm 0.2}$	20.0$_{\pm 0.2}$	33.0$_{\pm 0.6}$	39.9$_{\pm 0.4}$	6.2$_{\pm 0.4}$	3.2$_{\pm 0.4}$
3	4.5$_{\pm 0.2}$	2.4$_{\pm 0.2}$	4.1$_{\pm 0.3}$	**42.9**$_{\pm 0.7}$	28.7$_{\pm 0.3}$	20.9$_{\pm 0.3}$	33.7$_{\pm 0.4}$	40.5$_{\pm 0.5}$	6.7$_{\pm 0.3}$	3.4$_{\pm 0.3}$
5	5.3$_{\pm 0.2}$	3.3$_{\pm 0.2}$	2.4$_{\pm 0.3}$	42.7$_{\pm 0.2}$	**28.9**$_{\pm 0.9}$	**21.6**$_{\pm 0.9}$	**35.6**$_{\pm 0.3}$	**42.6**$_{\pm 0.2}$	6.6$_{\pm 0.6}$	3.3$_{\pm 0.6}$
10	5.9$_{\pm 0.5}$	3.9$_{\pm 0.5}$	**4.6**$_{\pm 0.3}$	42.1$_{\pm 0.7}$	26.1$_{\pm 0.4}$	19.6$_{\pm 0.4}$	31.6$_{\pm 0.3}$	38.8$_{\pm 0.3}$	**7.7**$_{\pm 0.0}$	**3.7**$_{\pm 0.0}$

Table 4: Zero-shot multimodal Llava results on the validation set for different values of k, the number of text evidence provided in the prompt (%). Reported results are averages over 3 iterations with standard deviations. The best scores are marked in **bold**.

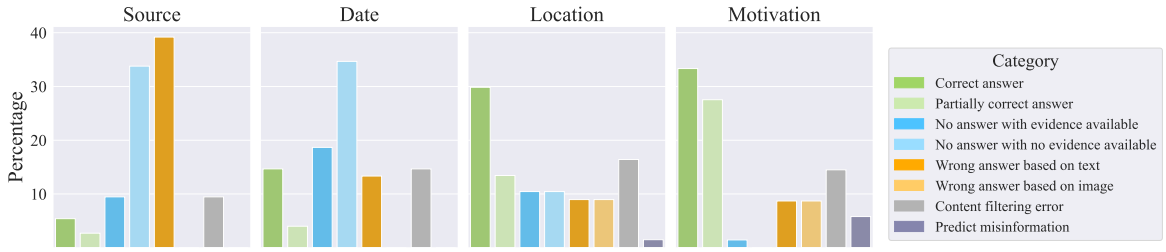

Figure 20: Qualitative analysis per pillar of zero-shot multimodal GPT4 answers (%).

to allow some diversity while remaining close to a deterministic output. Experiments with Llama2 (*meta-llama/Llama-2-7b-chat-hf*) and Llava (*llava-hf/llava-1.5-7b-hf*) are run on a single A100 GPU. For GPT4 (*gpt-4-1106-vision-preview*), we conduct our experiments through the Azure OpenAI service with API version *2023-10-01-preview*. The following questions are used as part of the prompts: "Who is the source/author of the image? Answer with one or more person or entities in a few words." for Source, "When was the image taken? Answer with one or more dates in a few words." for Date, "Where was the image taken? Answer with one or more locations in a few words." for Location, "Why was the image taken? Answer in a few words." for Motivation.

K Qualitative analysis per pillar

We break down the qualitative analysis of zero-shot multimodal GPT4 per pillar in Figure 20. Location and Motivation follow similar distributions, and contain most of the (partially) correct answers. They also concentrate all wrong answers based on the image content. Unlike Location, an answer is almost always provided for Motivation. For Source, GPT4 tends to provide answers that are incorrect and based on incorrect text evidence. Comparatively, it refrains more often from answering for Date, even when the image and the text evidence provide cues about the correct time period.
In the absence of relevant text evidence, GPT4 provides an answer based on the image content for Location more often than for Date. This is counter-intuitive, as one would expect the visual content to provide some cues for the Location of the image of course, but also for its Date, for example, by recognizing celebrities, or identifying the conflict or event depicted based on flags and clothes. This apparently weaker temporal knowledge of GPT4, compared to its geographical knowledge, deserves further investigation in future work.