Evidence Triangulation for Multimodal Fact-Checking in the Wild

Stefanos-Iordanis Papadopoulos[1,2], Zacharias Chrysidis[1], Christos Koutlis[1], Symeon Papadopoulos[1], and Panagiotis C. Petrantonakis[2]

[1] Information Technologies Institute @ CERTH, Thessaloniki, Greece
[2] Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece
{stefpapad,zchrysid,ckoutlis,papadop}@iti.gr, ppetrant@ece.auth.gr

Abstract. The proliferation of multimedia content on social platforms has fueled multimodal misinformation, where images are used to reinforce false claims. Consequently, Multimodal Fact-Checking (MFC) has emerged as an increasingly important research area. However, current progress is hindered by a reliance on synthetic training data and curated benchmarks that fail to capture the complexity of in-the-wild data. Furthermore, existing detection models rely on restricted intra-modality consistency or unconstrained all-to-all fusion, failing to capture nuanced relations between posts and external evidence. To address these limitations, we introduce X-POSE, a benchmark of real-world, community-annotated multimodal posts from X (formerly Twitter), augmented with full-length news articles retrieved via VLM-optimized search. Additionally, we propose TRENT, a novel MFC model that performs evidence triangulation using three parallel cross-attention streams alongside a relational fusion mechanism that explicitly models entailment and contradiction. Extensive evaluations demonstrate that TRENT consistently outperforms state-of-the-art specialized models and commercial VLMs. The code, prompt templates, and dataset are available at `https://github.com/stevejpapad/evidence-triangulation`.

Keywords: Automated Fact-Checking · Misinformation Detection · Multimodal Deep Learning · Evidence Retrieval · Crowdsourced Dataset

1 Introduction

Digital information ecosystems are increasingly defined by multimodal content. Videos, images, text, and synthetic media enable richer communication and expression, but also introduce new forms and avenues for large-scale deception. Beyond edited and AI-generated images, a key challenge is multimodal misinformation, in which images are paired with false claims to increase their perceived credibility [29] and wide-spread diffusion [21]. In response, researchers have been developing automated Multimodal Fact-Checking (MFC) systems to retrieve relevant information from the Web to cross-examine and verify the veracity of multimodal claims [2].

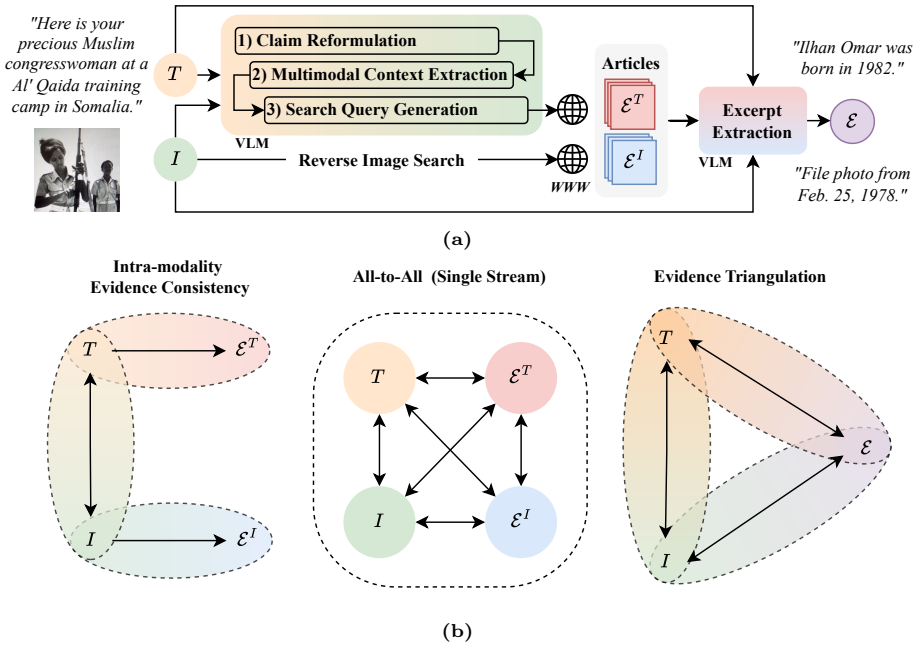


Fig. 1: (a) Proposed evidence collection pipeline: A VLM reformulates a multimodal post (I, T) into search queries to retrieve relevant articles $(\mathcal{E}^T, \mathcal{E}^I)$ and extract relevant excerpts $(\mathcal{E})$. (b) Conceptual comparison of MFC architectures: Evidence Triangulation leverages dedicated streams to explicitly isolate the three-way relationship between image, text, and evidence, instead of relying on intra-modality or all-to-all interactions.

To date, developing robust MFC models for image-text verification is constrained by the scarcity of high-quality data. As a result, recent work relies on weakly annotated or synthetic datasets [3, 4, 23, 26, 27, 38], or benchmarks sourced directly from fact-checking organizations [9, 31, 37]. The former introduces label noise, and synthetic data fail to capture the visual, linguistic, and cultural nuances of *in-the-wild* misinformation. Conversely, although datasets derived from fact-checking sources offer high-quality annotations, they provide clean, curated, and self-contained claims that do not reflect the noisy, contextual way users communicate on social media platforms.

To address this, we introduce X-POSE, a new in-the-wild, evidence-enhanced dataset of multimodal user posts from X (formerly Twitter), annotated through the Community Notes system[3], enabling both training and evaluation of MFC models—unlike prior work that used it only for evaluation [50]. We leverage VLMs to refine search queries, use both text-based and reverse image search APIs to retrieve full-length news articles as potential evidence, and apply VLMs to extract the article excerpts most relevant for supporting or refuting each

[3] `https://communitynotes.x.com/guide`

claim. Moreover, we integrate source credibility to enhance evidence quality and community ratings to assess consensus for more reliable evaluation.

As shown in Figure 1a, a post falsely claims a "Muslim congresswoman" was photographed training at an "Al Qaeda camp," exploiting harmful religious stereotypes to incite distrust and hostility. The VLM extracts a structured claim and contextual metadata (e.g., named entities, OCR text, image descriptions, and the "5Ws": Who, What, Where, When, Why), then generates search queries that, together with reverse image search, return relevant web articles. Finally, the VLM identifies the most pertinent evidence excerpts— revealing that the inferred individual was born in 1982, while the photo was taken in 1978—providing the key information needed for the MFC system to dispute the claim.

On the modeling side, existing methods fall into three categories: (1) *Intra-modality evidence consistency* models, which use separate streams to examine text-text evidence, image-image evidence, and the internal consistency of the image-text pair [1,55]; (2) *All-to-All* models, which jointly attend over all modalities and evidence [32,33]; and (3) *VLM-based* methods [7,49]. However, the first ignores cross-modal evidence relations, the second can struggle to identify subtle relations, while the third can suffer from high computational complexity.

To address these limitations, we introduce TRENT, an architecture that, as illustrated in Figure 1b, models text-to-all evidence, image-to-all evidence, and image-to-text alignment using three parallel cross-attention streams. The resulting outputs are processed by a relational fusion mechanism designed to capture entailment and contradiction. A comparative evaluation against commercial VLMs (including Grok 4, Claude Sonnet 4.6, Gemini 3, and GPT-5.4) and state-of-the-art (SotA) specialized MFC methods trained on X-POSE shows that TRENT consistently outperforms competing methods. Extensive ablation studies demonstrate that each component—evidence source credibility, evidence reranking, VLM-assisted excerpt extraction, and especially triangulated cross-attention and relational fusion—contributes significantly to overall performance.

2 Background and Related Work

2.1 Problem Formulation

Let $\mathcal{D} = \{(I_n, T_n, \mathcal{E}_n^I, \mathcal{E}_n^T, y_n)\}_{n=1}^N$ be a dataset of N multimodal instances. Each instance consists of a multimodal pair (I_n, T_n), where $I_n \in \mathcal{I}$ is an image and $T_n \in \mathcal{T}$ is its associated text (user post). To verify the veracity of the pair, we incorporate two sets of external evidence:

- $\mathcal{E}_n^I = \{E_{n,1}^I, E_{n,2}^I, \ldots, E_{n,M}^I\}$ represents external information retrieved via reverse-image search using I_n.
- $\mathcal{E}_n^T = \{E_{n,1}^T, E_{n,2}^T, \ldots, E_{n,M}^T\}$ represents external information retrieved using the text T_n as a query.

Each evidence set contains up to M items (e.g., full-length news articles or article excerpts). The ground-truth label is denoted by $y_n \in \{0, 1\}$, where $y_n = 0$

Table 1: Comparison of multimodal fact-checking datasets.

Dataset	Samples	Train	Eval.	Claims	Annotation
NewsCLIPpings+ [1]	85K	✓	✓(Synth.)	Synthetic	Automated
MMFakeBench [22]	11K	–	✓(Synth.)	Synthetic	Automated
FACTIFY [26]	50K	✓	✓	News Handles	Automated
VERITE [31]	1K	–	✓	Self-contained	Fact-checks
VeriTaS [37]	25K	–	✓	Self-contained	Fact-checks
M4FC [16]	7K	✓(Limited)	✓	Self-contained	Fact-checks
AVerImaTeC [9]	1.3K	✓(Limited)	✓	Self-contained	Fact-checks
X-FACTA [50]	2.4K	–	✓	In-the-wild	Crowdsourced
X-POSE (Ours)	5.7K	✓	✓	In-the-wild	Crowdsourced

represents a truthful claim and $y_n = 1$ indicates misinformation. The task of MFC is to learn a predictive function $f : (\mathcal{I}, \mathcal{T}, \mathcal{E}^I, \mathcal{E}^T) \rightarrow \hat{y}$, which minimizes the discrepancy between the predicted verdict $\hat{y}$ and the ground-truth y.

2.2 Multimodal Fact-Checking Datasets

Due to the limited scale and event diversity of early MFC datasets [5, 6, 18, 56], subsequent works turned to scalable synthetic data generation via *decontextualization* [3,4,23], *named-entity manipulation* [27,30,38], VLMs [34], or automated weak supervision [28]. To bridge the resulting distribution gap between algorithmically generated data and real-world misinformation, recent works leverage real-world evaluation benchmarks such as *COSMOS* [3] and *VERITE* [31]. However, these benchmarks comprise self-contained claims curated by fact-checkers rather than uncurated user posts, restricting their suitability for studying misinformation in the wild.

Beyond internal image-text consistency, robust multimodal fact-checking typically necessitates external information to cross-examine claims. However, existing evidence-enhanced datasets suffer from significant limitations in scale, balance, and realism. The NewsCLIPpings+ dataset [1] relies on synthetic decontextualization of news media and restricts textual evidence to article titles and brief captions. This setup frequently surfaces original or near-identical articles during evidence retrieval, causing data leakage [17,33]. Similarly, FACTIFY [26] restricts its scope to official news handles on Twitter and relies on heuristic annotations, thus allowing models to exploit shortcuts and reach an inflated 99–100% accuracy on the *Refute* class [11, 44], limiting its utility and realism.

Conversely, although *AVerImaTeC* [9] and *M4FC* [16] provide curated multimodal claims from fact-checked articles, their limited scale and class imbalances (e.g., only 17 'Supported' claims in *AVerImaTeC*, 292 in *M4FC*) prevent effective training. Similarly, while *VeriTaS* [37] expands MFC beyond image-text pairs, it is intended for evaluation and comprises curated self-contained claims from fact-checking sites rather than user-generated content.

2.3 Crowdsourced Fact-Checking

Recent work has examined the effectiveness of crowdsourced fact-checking systems such as X's Community Notes, which append clarifying annotations to potentially misleading posts. Adding a note has been shown to significantly reduce user engagement with misleading content [41]; furthermore, posts labeled as misleading are much more likely to be voluntarily deleted by their authors [15]. Notes citing neutral, high-quality sources also receive higher helpfulness scores, suggesting that contributors actively prioritize factual accuracy [19]. Beyond human curation, recent research has explored generating notes with LLMs and predicting their subsequent helpfulness ratings [14,39,48,51]. Community Notes have also served as specialized benchmarks for the intent-aware classification of AI-generated images [40] and multimodal fact-checking [50]. However, they have not yet been utilized for both training and evaluating MFC models.

In contrast, as shown in Table 1, the proposed X-POSE dataset comprises uncurated, user-generated content of sufficient scale and balance to enable both the training and evaluation of MFC models under realistic, in-the-wild settings.

2.4 Evidence-Based Detection

Existing MFC methods can be classified into three paradigms. The first assesses intra-modality consistency (e.g., text-to-text and image-to-image evidence examination) alongside internal image-text consistency, leveraging memory networks, stance extraction modules, or named-entity co-occurrence features [1, 54, 55]. The second paradigm models all-to-all relationships across the claim, image, and retrieved evidence blocks using dedicated attention- and Transformer-based architectures [30,32,33]. However, approaches that restrict comparisons to intra-modality consistency often overlook vital cross-modal evidence relations, whereas full all-to-all architectures can capture broad salient cues but fail to isolate the subtle relations of entailment and contradiction. More recently, a third paradigm explores VLMs and LLM-based agentic frameworks to rerank retrieved evidence and generate final verifications [7, 20, 35, 45, 49, 52].

In this study, we synthesize these paradigms by leveraging VLMs for claim reformulation, evidence collection, and filtering, coupled with TRENT, a specialized, lightweight MFC architecture that explicitly models the relationships between internal image-text pairs, image-to-all evidence, and text-to-all evidence.

3 Construction of X-POSE

Data Collection and Annotation. To construct X-POSE (*X Posts with Evidence*), we source from the X Community Notes archive. From the *Notes* file, we define the *Truthful* class using notes marked as *factually correct*, *not satire*, *not outdated*, and citing *trustworthy sources*. For the *Misinformation* class, we retain notes marked as *missing important context* or containing *factual errors* that still cite *trustworthy sources*. We exclude cases involving manipulated or AI-generated

images to focus on misinformation involving authentic but miscaptioned or out-of-context images. While identifying synthetic media is increasingly relevant to multimodal misinformation [10], it often demands distinct visual forensic analysis [13, 24] that falls outside the scope of this study.

We collect the associated posts, retaining only multimodal samples that contain both text and images. The final X-POSE comprises 5,704 unique image-text pairs, balanced across factually correct (2,881) and misinformation (2,823) samples, spanning 2017–2025. Averaging the outputs of two language detection tools—Fast-LangDetect and Lingua—we estimate that the majority of posts are in English (84.8%), followed by French (3.5%), Portuguese (2.1%), Spanish (1.8%), and German (1.2%). Using the IPTC taxonomy[4] with Gemma 3, we find X-POSE is primarily composed of political content (24.8%), followed by arts and media (19.7%), conflict and war (15.8%), crime and justice (7.6%), science and technology (6.8%), and society (5.7%), with the remaining 11 categories (e.g., health, economy, sports) each accounting for less than 5%, reflecting the dataset's topical diversity.

Consensus-based Data Filtering. Crowdsourced annotations can be noisy due to annotator disagreements regarding note phrasing, source credibility, or subjective interpretations. To mitigate this, we filter examples using a *helpfulness* score, which serves as a proxy for community consensus. For a given sample n, let η_n^+ and η_n^- denote the number of HELPFUL and NOT HELPFUL votes, respectively, aggregated from the raw Community Notes `ratings` files. An item's helpfulness score $h_n \in [0, 100]$ is defined as:

$$h_n = \begin{cases} \frac{\eta_n^+}{\eta_n^+ + \eta_n^-} \times 100, & \text{if } \eta_n^+ + \eta_n^- > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We retain samples where $h_n \geq \theta$ to construct high-agreement evaluation subsets. For simplicity, we refer to these as the $h \geq \theta$ subsets.

Evidence Collection. Our evidence collection pipeline is grounded in the workflows of professional fact-checkers, who (i) disambiguate claims, (ii) contextualize them by identifying relevant entities and circumstances, and (iii) formulate targeted queries to search for corroborating or contradicting evidence [25, 43, 47]. We model these steps using a three-stage VLM-assisted pipeline[5] (1) **Claim Reformulation (*Prompt 1*)** transforms raw, noisy user posts into clear, self-contained textual claims. (2) **Multimodal Context Extraction (*Prompt 2*)** extracts structured metadata, including named entities, image descriptions, OCR text, and the "5Ws". (3) **Search Query Generation (*Prompt 3*)** produces concise queries for search APIs, conditioned on the reformulated claim, image, and extracted context. We utilize Gemma 3 due to its consistent performance and lack of the restrictive safety filters—such as those regarding pub-

[4] `https://www.iptc.org/std/NewsCodes/treeview`

[5] See `src/vlm_prompts.py` in the codebase for all prompt templates.

lic figures or sensitive political discourse—frequently encountered in large proprietary models. We submit the generated queries and images to the Google Search and Lens APIs via ScrapingDog, retrieving up to 10 results per query. We fetch the pages, extract clean body text and discard cookie-consent popups and JavaScript-rendering errors.

Evidence Refinement. To reduce unreliable content and prevent information leakage (e.g., retrieving the associated Community Note), we exclude major social media platforms (X, Facebook, Instagram, YouTube, and TikTok) and use the MBFC[6] taxonomy to assess source credibility as a proxy for evidence quality and remove low-credibility outlets. From 44,559 retrieved pages, we remove 6,366 social media items and 1,404 low-credibility articles, while 18,536 remain unrated. The final collection comprises articles from 10,296 unique domains, with the most frequent sources including Wikipedia (1,958), BBC (704), The Guardian (680), CNN (507), Al Jazeera (403), and NPR (341). Furthermore, because 49.9% of these collected articles postdate the original user post, we conduct additional experiments under a past-only evidence setting to more accurately simulate the fact-checking of novel claims without future data leakage.

Excerpt Extraction. Information relevant to a specific claim is typically concentrated in localized segments of full-length articles [46]. Accordingly, we utilize a VLM, MiniCPM-V 2.6[7] [53], to perform cross-modal reasoning between the multimodal claim (text and image) and each retrieved article. Using *Prompt 4*, the VLM extracts *verbatim excerpts* from the text that are directly relevant to the claim. This process yielded 22,226 evidence excerpts (median length: 61 words). Of the collected posts, 4,785 are associated with at least one relevant excerpt, while 919 yielded no relevant evidence. The number of excerpts per post ranges up to a maximum of 16, with a median of 4; 79.4% originate from text-based retrieval and 20.6% from image-based retrieval.

4 The TRENT Architecture

4.1 Multimodal Representation

We employ CLIP ViT-L/14 [36] as a pre-trained encoder to extract image embeddings $\mathbf{i} \in \mathbb{R}^d$ and text embeddings $\mathbf{t} \in \mathbb{R}^d$, where $d = 768$ denotes the embedding dimension. Similarly, the top-M retrieved evidence excerpts are encoded into matrices $\mathbf{e}^I, \mathbf{e}^T \in \mathbb{R}^{d \times M}$ representing image- and text-based evidence, respectively, or concatenated into a single matrix $\mathbf{e} \in \mathbb{R}^{d \times 2M}$ when combined.

4.2 Evidence Reranking

Search engine APIs already include mechanisms to retrieve and rank potentially relevant content based on a given image-text pair. Nevertheless, as prior studies

[6] `https://github.com/drmikecrowe/mbfcext`
[7] `https://huggingface.co/openbmb/MiniCPM-V-2_6`

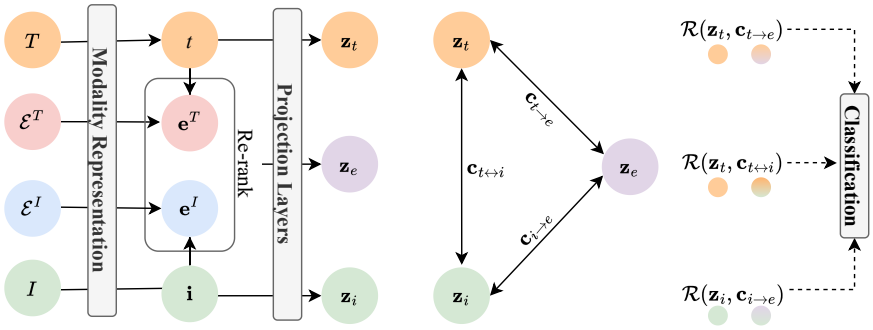


Fig. 2: Overview of TRENT: Input text (T), image (I), and evidence excerpts ($\mathcal{E}$) are projected to embeddings ($\mathbf{z}_t, \mathbf{z}_i, \mathbf{z}_e$) and processed by three parallel cross-attention streams. The resulting contextualized representations ($\mathbf{c}_{t\to e}, \mathbf{c}_{t\to i}, \mathbf{c}_{i\leftrightarrow e}$) are integrated via relational fusion $\mathcal{R}(\cdot)$ to capture entailment and contradiction for classification.

have shown [32], it is often beneficial to apply a reranking step. For each post n, we rerank retrieved evidence items using cosine similarity in the embedding space. Specifically, we compare the claim text embedding $\mathbf{t}_n$ against all text-retrieved evidence embeddings $\mathbf{e}_n^T$, and the claim visual embedding $\mathbf{i}_n$ against all image-retrieved evidence embeddings $\mathbf{e}_n^I$. Each evidence set contains up to 10 items. We then select the top-M evidence items with the highest similarity scores. If fewer than M candidates are available, we pad with zero vectors.

4.3 Evidence Triangulation

TRENT employs three parallel cross-attention transformer blocks $\mathrm{C}(\cdot)$ and relational fusion $\mathcal{R}(\cdot)$ to capture interactions across images, texts, and evidence.

First, raw embeddings are projected into a shared latent space of dimension $d_c = 512$ using modality-specific learnable linear projections:

$$\mathbf{z}_i = \mathbf{W}_i\ \mathbf{i}, \quad \mathbf{z}_t = \mathbf{W}_t\ \mathbf{t}, \quad \mathbf{z}_e = \mathbf{W}_e\ \mathbf{e} \tag{2}$$

where $\mathbf{W}_i, \mathbf{W}_t, \mathbf{W}_e \in \mathbb{R}^{d_c \times d}$ for image, text, and evidence, respectively.

Given query $\mathbf{z}_q$ and key-value $\mathbf{z}_{kv}$ vectors, a $\mathrm{C}(\cdot)$ block is defined as:

$$\mathbf{a} = \mathrm{LN}(\mathbf{z}_q + \mathrm{MHA}(\mathbf{z}_q, \mathbf{z}_{kv}, \mathbf{z}_{kv})) \tag{3}$$

$$\mathrm{C}(\mathbf{z}_q, \mathbf{z}_{kv}) = \mathrm{LN}(\mathbf{a} + \mathrm{FFN}(\mathbf{a})) \tag{4}$$

where $\mathrm{MHA}(\cdot)$ denotes multi-head attention with $h = 8$ heads and a key padding mask to ignore zero-padded entries, $\mathrm{FFN}(\mathbf{z}) = \mathbf{W}_2(\mathrm{GELU}(\mathbf{W}_1\mathbf{z}))$ is a Feed-Forward Network with $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_c \times d_c}$, and Layer Normalization (LN) is applied after each residual connection.

We implement three parallel verification streams to examine the relationship between: (i) the visual content and all external evidence ($\mathbf{c}_{i\to e}$), (ii) the textual claim and all external evidence ($\mathbf{c}_{t\to e}$), and (iii) the internal cross-modal consistency between the paired text and image ($\mathbf{c}_{t\leftrightarrow i}$). Formally expressed:

$$\mathbf{c}_{i\to e} = \mathrm{C}(\mathbf{z}_i, \mathbf{z}_e),\ \mathbf{c}_{t\to e} = \mathrm{C}(\mathbf{z}_t, \mathbf{z}_e),\ \mathbf{c}_{t\leftrightarrow i} = \mathrm{C}(\mathbf{z}_t, \mathbf{z}_i) \tag{5}$$

Relational Fusion. To capture entailment and contradiction relations between the image-text pair and the external evidence, we employ a relational fusion function $\mathcal{R}(\cdot)$ inspired by standard formulations in natural language inference (NLI) [12], but adapted for MFC:

$$\mathcal{R}(\mathbf{z}_1, \mathbf{z}_2) = [\mathbf{z}_1; \mathbf{z}_2; |\mathbf{z}_1 - \mathbf{z}_2|; \mathbf{z}_1 \odot \mathbf{z}_2] \tag{6}$$

where $[;]$ denotes concatenation and $\odot$ denotes the element-wise product. Applying $\mathcal{R}$ to each stream and concatenating yields the final representation:

$$\mathbf{z}_r = [\mathcal{R}(\mathbf{z}_i, \mathbf{c}_{i\to e}); \mathcal{R}(\mathbf{z}_t, \mathbf{c}_{t\to e}); \mathcal{R}(\mathbf{z}_t, \mathbf{c}_{t\leftrightarrow i})] \tag{7}$$

with $\mathbf{z_r} \in \mathbb{R}^{12d_c}$; $4d_c$ per stream.

Classification. The final veracity score $\hat{y}$ is obtained by a linear classification layer followed by a sigmoid activation: $\hat{y} = \sigma(\mathbf{W}_r\ \mathbf{z}_r + b)$, with $\mathbf{W}_r \in \mathbb{R}^{1\times 12d_c}$. The model is trained using the binary cross-entropy loss.

5 Experimental Setup

Implementation Details. TRENT and all competing methods are trained and evaluated on X-POSE under two experimental setups. The primary setup uses a random split into 4,588 training, 559 validation, and 557 test samples. Performance is measured using macro F1 on the unfiltered test set and on high-agreement subsets ($h \geq 80\%, h \geq 90\%$). Applying thresholds $\theta \in \{10, 20, \ldots, 90\}$ on this random test set results in pruned evaluation subsets of 494, 475, 448, 402, 362, 315, 269, 200, and 120 samples, respectively. To evaluate temporal generalization, we introduce a chronological data split consisting of 4,563 training posts spanning 2017–11/2024 (with 95% in 2023–2024), 570 validation posts from 11/2024–01/2025, and 571 test posts from 01/2025–04/2025. Across both setups, the validation sets are used for hyperparameter tuning over the number of evidence items $M \in \{1, 3, 5, 10\}$ and learning rate $\eta \in \{10^{-4}, 5 \times 10^{-5}\}$. All models are optimized with a batch size of 512 using Adam for up to 50 epochs, with early stopping after 10 epochs of validation stagnation. Random seeds are set to 0 across PyTorch, Python, and NumPy for reproducibility.

Competing Methods. As no prior work has evaluated on X-POSE, we implement a broad set of MFC models: **CCN** [1], **ERIC-FND** [8], and **ECENet** [55] which leverage intra-modality evidence consistency, as well as **DT-Transformer**

[30], **RED-DOT** [32], **MUSE**, and **AITR** [33] that model all-to-all interactions within a single stream. For a fair comparison, all replicated methods utilize a CLIP ViT-L/14 backbone and follow the training protocols and official repositories of the original papers. Additionally, we evaluate several open-source and commercial VLMs for zero-shot detection on X-POSE, using *Prompt 6* to incorporate the image-text pair alongside the extracted evidence excerpts: MiniCPM [53], Gemma 3 and 4, GPT-5 Mini and 5.4 Mini, Gemini 2.5 Flash and 3 Flash, Grok 4 Fast, and Claude 4.6 Sonnet. Except for MiniCPM (which is run locally), all VLMs are accessed via the OpenRouter.ai API. Finally, we evaluate TRENT on NewsCLIPpings+ and VERITE, comparing it against optimized detectors **RED-DOT** and **AITR**, as well as VLM-based detectors **SNIFFER** [35], **TRUST-VL** (LLaVA-1.5) [52], **E2LVLM** (Qwen2-VL) [49], **DEFAME** (GPT-4o) [7], and **MAD-Sherlock** (GPT-4o) [20], reporting performance from their original papers.

6 Results

6.1 Comparative Results

Table 2 presents the performance of TRENT on X-POSE, against a wide range of MFC detectors. Our model consistently outperforms all baselines across the standard random split, achieving macro-F1 scores of 63.10%, 67.36%, and 70.83% on the full test set and high-agreement subsets ($h \geq 80, 90$). Under a past-only evidence setting—only allowing evidence published before the user post—TRENT's performance drops to 62.22% on the random split and 60.14% on the chronological split. This decrease is expected given the significantly more challenging and realistic constraints of these settings; however, TRENT still consistently outperforms all competing methods.

Crucially, this competitive efficiency stems from architectural design rather than scale. TRENT comprises only 5.13M trained parameters, significantly fewer than specialized methods like RED-DOT (16.9M) or AITR (18.7M), and orders of magnitude smaller than the tens or hundreds of billions of parameters of VLMs. Additionally, excluding the shared upfront costs of evidence excerpt extraction and embedding generation (which are identical across all models and performed only once), TRENT requires a mere 6.6–32.6 MFLOPs (scaling from 2 to 20 evidence documents) and is trained in just 21–30 seconds on a single RTX 3060 GPU. Furthermore, TRENT performs inference on the test set (557 samples) in less than 1 second; in stark contrast, evaluating the same inference set took 11m for GPT-5.4 Mini, 26m for Claude Sonnet 4.6, 33m for Gemma 4, and up to 56m for Gemini 3 Flash.

Within specialized MFC methods, all-to-all attention-based models (AITR, RED-DOT, DT-Transformer) tend to outperform intra-modality consistency models (CCN, ERIC-FND, ECENet). The latter restrict examination across modalities, missing key interactions between the claim and image evidence, and the image with textual evidence. Yet, while all-to-all architectures attend to all inputs simultaneously, their single-stream design can still overlook fine-grained

Table 2: Performance comparison on X-POSE. We report Macro-F1 across the random data split (including full test set and high-agreement subsets), alongside a past-only evidence setting evaluated on the full ($h \geq 0\%$) random and chronological data splits.

	Random Split			**Chrono. Split**	
				Past-Only Evidence	
Methods	**F1**	**F1** ($h \geq 80\%$)	**F1** ($h \geq 90\%$)	**F1**	**F1**
MiniCPM	56.85	55.59	56.55	53.67	56.48
Gemma 3	54.48	56.16	54.92	54.02	52.81
Gemma 4	60.86	<u>65.91</u>	<u>68.33</u>	<u>60.31</u>	<u>58.67</u>
GPT-5 Mini	56.36	57.12	58.74	56.38	55.21
GPT-5.4 Mini	57.48	57.37	57.91	54.73	55.49
Gemini 2.5 Flash	56.65	62.49	63.96	56.18	54.42
Gemini 3 Flash	61.33	63.32	64.91	59.93	56.86
Grok 4 Fast	57.67	62.99	64.91	57.54	57.67
Claude Sonnet 4.6	57.15	61.28	65.43	54.17	56.78
CCN	58.61	61.45	58.52	56.89	54.33
ERIC-FND	55.56	57.33	61.03	57.27	55.22
ECENet	57.92	61.20	61.16	56.91	56.67
MUSE	53.36	55.36	57.27	53.76	49.37
DT-Transformer	57.99	60.78	62.25	57.43	57.94
RED-DOT	<u>61.56</u>	62.00	63.17	59.69	56.23
AITR	58.46	64.02	63.78	57.99	56.77
TRENT	**63.10**	**67.36**	**70.83**	**62.22**	**60.14**

dependencies. In contrast, TRENT explicitly models text → all evidence, image → all evidence, and internal image-text interactions through three dedicated cross-attention streams, combined with relational fusion, thereby better isolating subtle supporting or contradicting signals.

Among the evaluated VLMs, Gemma 4 and Gemini 3 yield the highest performance behind TRENT. We observe improvements within these model families, with Gemma 4 significantly outperforming Gemma 3, and Gemini 3 surpassing Gemini 2.5. While these gains may stem from enhanced reasoning capabilities, they could also be partially attributed to data contamination, as these large models could have encountered some of the target posts, events, or associated source articles during pre-training. Nevertheless, this risk is mitigated on the chronological split, which features minimal overlap with the training data of these VLMs due to its post-January 2025 timeline.

Moreover, as shown in Table 3, TRENT outperforms RED-DOT, E2LVLM, TRUST-VL, SNIFFER, DEFAME, and MAD-Sherlock across NewsCLIPpings+ and VERITE, while achieving competitive performance with AITR. However, specialized MFC models like RED-DOT and AITR struggle on X-POSE because they were developed on out-of-context benchmarks prone to retrieval shortcuts (Section 2.2), where the similarity-based baseline MUSE attains high scores on

Table 3: Comparison on NewsCLIPpings+ and VERITE. The top two rows (shown in gray) denote methods relying on retrieval shortcuts.

Method	NewsCLIPpings+	VERITE
MUSE (MLP)	90.0	80.5
AITR (w/ MUSE)	93.3	81.0
RED-DOT	90.3	76.9
AITR (w/o MUSE)	89.4	71.0
SNIFFER	88.4	74.0
E2LVLM (Qwen2-VL)	89.9	74.4
TRUST-VL (LLaVA-1.5)	90.4	73.6
DEFAME (GPT-4o)	–	78.4
MAD-Sherlock (GPT-4o)	<u>90.8</u>	<u>79.5</u>
TRENT	**92.5**	**79.6**

NewsCLIPpings (90.0%) and VERITE (80.5%) due to leakage [33]. Similarly, when MUSE shortcuts are ablated, AITR's performance significantly decreases on VERITE, from 81.0 to 71.0%. In contrast, MUSE drops to near-random on X-POSE (53.36% on the full test set and 49.37% on the chronological split). This sharp degradation indicates that X-POSE avoids such retrieval shortcuts.

6.2 Ablation Study

Table 4 presents ablations analyzing the training data (Ablation 1), inputs (Ablations 2–10), and architectural components (Ablations 11–16) of TRENT.

Scale vs. Data Quality. All models are trained on the full training set and evaluated on the test set and two high-agreement subsets ($h \geq 80$, $h \geq 90$). Training only on high-agreement data (Ablation 1) improves label reliability but reduces the dataset to 1,765 ($h \geq 80$) and 1,015 ($h \geq 90$) samples, causing a clear performance drop. For TRENT, scale outweighs consensus during training.

Modality and Evidence Representation. Ablations 2-4 modify evidence representations. Using full-article embeddings from a long-context LLM (Ablation 2; 4096-dim Qwen GTE-7B[8] with *Prompt 5*) substantially degrades performance, suggesting verification cues are localized and diluted in full-document embeddings. Replacing excerpts with VLM summaries (Ablation 3; MiniCPM with *Prompt 5*) also reduces performance, likely due to loss of fine-grained details. Substituting the CLIP-aligned text encoder with multi-E5 (Ablation 4) yields no gains, likely due to visual-textual misalignment.

Evidence Filtering and Reranking. Allowing low-credibility and social media sources in the evidence pool reduces performance (Ablation 5), validating

[8] `https://huggingface.co/Alibaba-NLP/gte-Qwen1.5-7B-instruct`

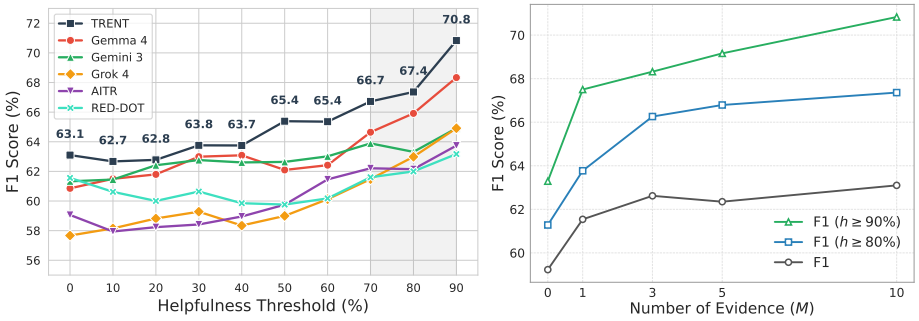
Table 4: Ablation analysis of TRENT's components.

#	**Ablation**	**F1**	**F1** ($h \geq 80\%$)	**F1** ($h \geq 90\%$)
1	Trained on filtered data	-	64.91	66.43
2	Full article features	58.88	60.33	62.48
3	Relevant summaries	60.33	65.99	67.20
4	Multi-E5	60.17	62.93	68.01
5	−Evidence quality filter	61.35	66.05	68.19
6	−Evidence reranking	61.37	65.95	69.99
7	$-\mathcal{I}$	57.62	58.42	58.32
8	$-\mathcal{E}$	59.23	61.28	63.29
9	$-\mathcal{E}^I$	61.75	61.48	64.04
10	$-\mathcal{E}^T$	59.52	66.13	68.32
11	Intra-modality Consistency − $\mathcal{R}$	59.46	61.34	61.90
12	Intra-modality Consistency	58.83	62.25	62.13
13	$-\mathcal{R}$	60.41	65.00	63.17
14	$-\mathbf{c_{t\to e}}$	58.94	61.49	63.08
15	$-\mathbf{c_{i\to e}}$	61.88	62.50	64.96
16	$-\mathbf{c}_{t\leftrightarrow i}$	58.50	62.34	65.83
	TRENT	**63.10**	**67.36**	**70.83**

the role of credibility-based filtering in mitigating the impact of noisy or biased evidence. Similarly, disabling reranking (Ablation 6) results in a slight performance decline on high-agreement subsets. This indicates that while the evidence retrieval and excerpt extraction pipeline yields reasonably relevant candidates, reranking provides a beneficial refinement step.

Contribution of Modalities and External Evidence. Removing images (Ablation 7) drastically degrades performance, confirming the contribution of the visual signals and that X-POSE is not dominated by unimodal shortcuts. Removing all external evidence ($-\mathcal{E}$, Ablation 8) causes a significant drop, underscoring the necessity of external information. Removing image-retrieved evidence ($-\mathcal{E}^I$, Ablation 9) is more detrimental than removing text-retrieved evidence ($-\mathcal{E}^T$, Ablation 10), suggesting that reverse-image search often provides crucial historical context, or provenance, missing from the user post.

Evidence Triangulation and Relational Fusion. Ablations 11-12 restrict processing to image → image-retrieved evidence and text → text-retrieved evidence, mirroring prior intra-modality evidence consistency architectures [1, 55], and significantly reducing performance. Replacing relational fusion $\mathcal{R}$ (Ablations 11 and 13) with feature concatenation similarly weakens performance, underscoring its importance. Moreover, removing any individual stream from the three-stream architecture ($\mathbf{c_{t\to e}}$, $\mathbf{c_{i\to e}}$, $\mathbf{c}_{t\leftrightarrow i}$) (Ablations 14-16) significantly reduces performance. Together, these findings validate TRENT's design.



(a) **(b)**

Fig. 3: (a) Performance of TRENT and baselines across helpfulness thresholds θ. (b) Impact of evidence quantity (M) on TRENT.

6.3 Further Analysis

Figure 3a evaluates TRENT and five of the top performing methods across varying helpfulness thresholds $\theta \in \{10, 20, \ldots, 90\}$. Performance remains static under low agreement ($\theta \leq 60\%$, slope $\beta \approx 0.02$), suggesting that simple majorities (50–60%) may suffer from user polarization. Conversely, beyond $\theta = 70\%$, methods exhibit a positive trend in macro-F1 ($\beta \approx 0.11$), indicating that super-majority agreement yields cleaner, higher-fidelity labels. However, enforcing these higher thresholds significantly reduces available sample volume, introducing a trade-off between annotation quality and dataset scale.

Figure 3b illustrates the impact of evidence quantity M on TRENT. Performance improves markedly from $M = 0$ to $M = 1$ and continues to rise, peaking at $M = 10$. This indicates that access to multiple evidence excerpts facilitates cross-source referencing, resolving ambiguities that a single excerpt ($M = 1$) may miss. Furthermore, it demonstrates that TRENT can successfully isolate relevant signals from increasingly noisy, high-dimensional inputs.

Figure 4 presents two inference examples that illustrate the importance of consensus-based and evidence-quality filtering. In (a), the retrieved evidence is credible, but the annotation is incorrect, as indicated by the very low helpfulness score (3.6%), since the post is humorous rather than misinformation. This motivates our consensus-based filtering, which isolates higher-fidelity evaluation labels. In (b), despite a correct annotation (100% helpfulness), the evidence originates from an unreliable, highly biased source[9] with low factual reporting, which provides misleading information that superficially supports the claim. This highlights the necessity of assessing evidence credibility (Section 3).

[9] `https://mediabiasfactcheck.com/pacific-pundit/`

(a) Post: "*It looks like they froze him in an ice cube*"
Community Note: "*This image is dated July 5th, 2019...*"
Label: Misinformation
Prediction: Factually Correct
Helpfulness: 3.6% (7 Helpful, 185 Not-helpful)
Excerpt: "*The President delivered his speech behind a [...] bulletproof barrier.*" [Source: CNN]

(b) Post: "*Columbia Univ 'smart' kids spell 'Palestine' incorrectly...*"
Community Note: "*This is not Columbia... shows Tabaret Hall at U. Ottawa.*
Label: Misinformation
Prediction: Factually Correct
Helpfulness: 100% (6 Helpful, 0 Not-helpful)
Excerpt: "*[...] at Columbia, [...], they can't even spell Palestine.*" [Source: Pacific Pundit]

Fig. 4: Qualitative examples of low consensus and evidence credibility.

7 Conclusion

In this study, we introduced X-POSE, a novel *in-the-wild* evidence-enhanced dataset for training and evaluating MFC models, comprising user posts annotated through X's Community Notes and augmented with VLM-refined external evidence. To address the limitations of prior MFC architectures, we proposed TRENT, which performs explicit evidence triangulation with relational fusion across text, image, and retrieved sources; it outperforms specialized MFC models and commercial VLMs, highlighting its efficiency and effectiveness.

Despite these improvements, overall performance remains limited, underscoring the difficulty of in-the-wild detection and the realism of X-POSE. This leaves ample room for further research. First, X-POSE can be used to develop and evaluate novel methods for in-the-wild MFC, while also exploring explainability [47] by comparing model explanations against community notes. Second, because nearly 50% of the retrieved articles in X-POSE currently lack credibility ratings, further research is required on evidence quality [17] and source credibility [42]. Finally, while our consensus-based filtering improves crowdsourced annotation quality by removing low-agreement entries, it introduces a quality-scale trade-off. To address this, X-POSE can be scaled up as new community notes are released, while expanding beyond X to other platforms like TikTok, Facebook, and Instagram to strengthen future research on cross-platform MFC.

Acknowledgements

This work is partially funded by Horizon Europe projects AI-CODE and ELLIOT under grant agreement no. 101135437 and 101214398, respectively.

References

1. Abdelnabi, S., Hasan, R., Fritz, M.: 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. pp. 14940–14949 (2022)
2. Akhtar, M., Schlichtkrull, M., Guo, Z., Cocarascu, O., Simperl, E., Vlachos, A.: Multimodal automated fact-checking: A survey. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 5430–5448 (2023)
3. Aneja, S., Bregler, C., Niessner, M.: Cosmos: Catching out-of-context image misuse using self-supervised learning. Proceedings of the AAAI Conference on Artificial Intelligence **37**(12), 14084–14092 (Jun 2023). https://doi.org/10.1609/aaai.v37i12.26648
4. Biamby, G., Luo, G., Darrell, T., Rohrbach, A.: 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. pp. 1530–1549 (2022)
5. Boididou, C., Andreadou, K., Papadopoulos, S., Dang Nguyen, D.T., Boato, G., Riegler, M., Kompatsiaris, Y., et al.: Verifying multimedia use at mediaeval 2015. In: MediaEval 2015, vol. 1436. CEUR-WS (2015)
6. Boididou, C., Middleton, S.E., Jin, Z., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Kompatsiaris, Y.: Verifying information with multimedia content on twitter: a comparative study of automated approaches. Multimedia tools and applications **77**, 15545–15571 (2018). https://doi.org/10.1007/s11042-017-5132-9
7. Braun, T., Rothermel, M., Rohrbach, M., Rohrbach, A.: Defame: Dynamic evidence-based fact-checking with multimodal experts. In: International Conference on Machine Learning. pp. 5383–5417. PMLR (2025)
8. Cao, B., Wu, Q., Cao, J., Liu, B., Gui, J.: External reliable information-enhanced multimodal contrastive learning for fake news detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 31–39 (2025)
9. Cao, R., Ding, Z., Guo, Z., Schlichtkrull, M., Vlachos, A.: Averimatec: A dataset for automatic verification of image-text claims with evidence from the web. Advances in Neural Information Processing Systems **38** (2026)
10. Chrysidis, Z., Papadopoulos, S.I., Papadopoulos, S.: The synthetic media shift: Tracking the rise, virality, and detectability of ai-generated multimodal misinformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8626–8635 (2026)
11. Chrysidis, Z., Papadopoulos, S.I., Papadopoulos, S., Petrantonakis, P.: Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking. In: Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation. pp. 73–81 (2024)
12. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 670–680 (2017)
13. Deng, J., Lin, C., Zhao, Z., Liu, S., Peng, Z., Wang, Q., Shen, C.: A survey of defenses against ai-generated visual media: Detection, disruption, and authentication. ACM Computing Surveys **58**(5), 1–35 (2025)
14. Franzmeyer, T., Shtedritski, A., Albanie, S., Torr, P., Henriques, J.F., Foerster, J.: Hellofresh: Llm evalutions on streams of real-world human editorial actions

across x community notes and wikipedia edits. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 12702–12716 (2024)
15. Gao, Y., Zhang, M.M., Rui, H.: Can crowdchecking curb misinformation? evidence from community notes. Information Systems Research (2025)
16. Geng, J., Tonglet, J., Gurevych, I.: M4fc: a multimodal, multilingual, multicultural, multitask real-world fact-checking dataset. arXiv preprint arXiv:2510.23508 (2025)
17. Glockner, M., Hou, Y., Gurevych, I.: Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 5916–5936 (2022)
18. Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q.: Novel visual and statistical image features for microblogs news verification. IEEE transactions on multimedia **19**(3), 598–608 (2016)
19. Kangur, U., Chakraborty, R., Sharma, R.: Who checks the checkers? exploring source credibility in twitter's community notes. Journal of Computational Social Science **9**(1), 24 (2026)
20. Lakara, K., Sock, J., Rupprecht, C., Torr, P., Collomosse, J., de Witt, C.S.: Mad-sherlock: Multi-agent debates for out-of-context misinformation detection (2024)
21. Li, Y., Xie, Y.: Is a picture worth a thousand words? an empirical study of image content and social media engagement. Journal of marketing research **57**(1), 1–19 (2020)
22. Liu, X., Li, Z., Li, P., Huang, H., Xia, S., Cui, X., Huang, L., Deng, W., He, Z.: Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In: International Conference on Learning Representations. vol. 2025, pp. 86327–86352 (2025)
23. Luo, G., Darrell, T., Rohrbach, A.: Newsclippings: Automatic generation of out-of-context multimodal media. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6801–6817 (2021)
24. Mehrjardi, F.Z., Latif, A.M., Zarchi, M.S., Sheikhpour, R.: A survey on deep learning-based image forgery detection. Pattern Recognition **144**, 109778 (2023)
25. Micallef, N., Armacost, V., Memon, N., Patil, S.: True or false: Studying the work practices of professional fact-checkers. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW1), 1–44 (2022)
26. Mishra, S., Suryavardan, S., Bhaskar, A., Chopra, P., Reganti, A.N., Patwa, P., Das, A., Chakraborty, T., Sheth, A.P., Ekbal, A., et al.: Factify: A multi-modal fact verification dataset. In: DE-FACTIFY@ AAAI. p. np (2022)
27. Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., Ewerth, R.: Multimodal analytics for real-world news using measures of cross-modal entity consistency. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. pp. 16–25 (2020). https://doi.org/10.1145/3372278.3390670
28. Nakamura, K., Levy, S., Wang, W.Y.: Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 6149–6157 (2020)
29. Newman, E.J., Zhang, L.: How non-probative photos shape belief. –Stephan Lewandowsky, Cognitive Science p. 90 (2020)
30. Papadopoulos, S.I., Koutlis, C., Papadopoulos, S., Petrantonakis, P.: Synthetic misinformers: Generating and combating multimodal misinformation. In: Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation. pp. 36–44 (2023). https://doi.org/10.1145/3592572.3592842
31. Papadopoulos, S.I., Koutlis, C., Papadopoulos, S., Petrantonakis, P.C.: Verite: a robust benchmark for multimodal misinformation detection accounting for uni-

modal bias. International Journal of Multimedia Information Retrieval **13**(1), 4 (2024)
32. Papadopoulos, S.I., Koutlis, C., Papadopoulos, S., Petrantonakis, P.C.: Red-dot: Multimodal fact-checking via relevant evidence detection. IEEE Transactions on Computational Social Systems (2025)
33. Papadopoulos, S.I., Koutlis, C., Papadopoulos, S., Petrantonakis, P.C.: Similarity over factuality: Are we making progress on multimodal out-of-context misinformation detection? In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 5570–5579 (February 2025)
34. Papadopoulos, S.I., Koutlis, C., Papadopoulos, S., Petrantonakis, P.C.: Latent multimodal reconstruction for misinformation detection. arXiv preprint arXiv:2504.06010 (2025)
35. Qi, P., Yan, Z., Hsu, W., Lee, M.L.: Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13052–13062 (2024)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
37. Rothermel, M., Kornmann, M., Rohrbach, M., Rohrbach, A.: Veritas: The first dynamic benchmark for multimodal automated fact-checking. arXiv preprint arXiv:2601.08611 (2026)
38. Sabir, E., AbdAlmageed, W., Wu, Y., Natarajan, P.: Deep multimodal image-repurposing detection. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1337–1345 (2018). https://doi.org/10.1145/3240508.3240707
39. Singh, S., Wu, J., Churina, S., Jaidka, K.: On the limitations of llm-synthesized social media misinformation moderation. In: I Can't Believe It's Not Better: Challenges in Applied Deep Learning
40. Skoularikis, A., Papadopoulos, S.I., Papadopoulos, S., Petrantonakis, P.C.: 'humor, art, or misinformation?": A multimodal dataset for intent-aware synthetic image detection. In: Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web. pp. 95–104 (2025)
41. Slaughter, I., Peytavin, A., Ugander, J., Saveski, M.: Community notes reduce engagement with and diffusion of false information online. Proceedings of the National Academy of Sciences **122**(38), e2503413122 (2025)
42. Srba, I., Razuvayevskaya, O., Leite, J.A., Moro, R., Schlicht, I.B., Tonelli, S., García, F.M., Lottmann, S.B., Teyssou, D., Porcellini, V., et al.: A survey on automatic credibility assessment using textual credibility signals in the era of large language models. ACM Transactions on Intelligent Systems and Technology (2025)
43. Sundriyal, M., Chakraborty, T., Nakov, P.: From chaos to clarity: Claim normalization to empower fact-checking. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 6594–6609 (2023)
44. Suryavardan, S., Mishra, S., Chakraborty, M., Patwa, P., Rani, A., Chadha, A., Reganti, A., Das, A., Sheth, A., Chinnakotla, M., et al.: Findings of factify 2: multimodal fake news detection. arXiv preprint arXiv:2307.10475 (2023)
45. Tahmasebi, S., Müller-Budack, E., Ewerth, R.: Multimodal misinformation detection using large vision-language models. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. p. 2189–2199.

CIKM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3627673.3679826
46. Vladika, J., Matthes, F.: Scientific fact-checking: A survey of resources and approaches. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 6215–6230 (2023)
47. Warren, G., Shklovski, I., Augenstein, I.: Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–21 (2025)
48. Wu, J., Fu, Z., Wang, H., Li, F., Guo, J., Nakov, P., Kan, M.Y.: Beyond the crowd: Llm-augmented community notes for governing health misinformation. arXiv preprint arXiv:2510.11423 (2025)
49. Wu, J., Fu, Y., Yu, N., Fu, G.: E2lvlm:evidence-enhanced large vision-language model for multimodal out-of-context misinformation detection (2025), https://arxiv.org/abs/2502.10455
50. Xiao, Y., Han, Z., Wang, Y., Jiang, H.: Xfacta: Contemporary, real-world dataset and evaluation for multimodal misinformation detection with multimodal llms. arXiv preprint arXiv:2508.09999 (2025)
51. Xing, R., Nakov, P., Baldwin, T., Lau, J.H.: Communitynotes: A dataset for exploring the helpfulness of fact-checking explanations. arXiv preprint arXiv:2510.24810 (2025)
52. Yan, Z., Qi, P., Hsu, W., Lee, M.L.: Trust-vl: An explainable news assistant for general multimodal misinformation detection. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 5588–5604 (2025)
53. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. Nat Commun 16, 5509 (2025) (2025)
54. Yuan, X., Guo, J., Qiu, W., Huang, Z., Li, S.: 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. pp. 4268–4280 (2023)
55. Zhang, F., Liu, J., Zhang, Q., Sun, E., Xie, J., Zha, Z.J.: Ecenet: explainable and context-enhanced network for muti-modal fact verification. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1231–1240 (2023)
56. Zlatkova, D., Nakov, P., Koychev, I.: 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). pp. 2099–2108 (2019)