RED-DOT: Multimodal Fact-Checking via Relevant Evidence Detection

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis, *Senior Member, IEEE*

***Abstract*—Online misinformation is often multimodal in nature, i.e., caused by misleading associations between texts and accompanying images. To support the fact-checking process, researchers have been recently developing automatic multimodal methods that gather and analyze external information, evidence, related to the image–text pairs under examination. However, prior works incorrectly assumed that all external information collected from the Web is relevant. In this study, we introduce a "relevant evidence detection" (RED) module to discern whether each piece of evidence is relevant, to support or refute the claim. Specifically, we develop the "relevant evidence detection directed transformer" (RED-DOT) and explore multiple architectural variants (e.g., single or dual-stage) and mechanisms (e.g., "guided attention"). Extensive ablation and comparative experiments demonstrate that RED-DOT outperforms the state-of-the-art (SotA), achieving up to 33.7% accuracy improvement on the VERITE benchmark. Furthermore, our evidence reranking and element-wise modality fusion led to RED-DOT surpassing the SotA on NewsCLIPpings+ by up to 3% without the need for numerous evidence or multiple backbone encoders. We release our code at: https://github.com/stevejpapad/relevant-evidence-detection.**

***Index Terms*—Deep learning, misinformation detection, multimodal fact-checking, multimodal learning.**

I. Introduction

THE dissemination of misinformation has greatly intensified in the digital age with the advent of the internet and social media platforms [1] and entails numerous adverse outcomes [2], [3], [4], [5]. Recent advances in generative AI, such as large language models and image synthesis models such as stable diffusion have made it easier to generate hyper-realistic misinformation [6]. However, decontextualization, where images are misused to support false narratives, remains a major challenge in multimodal misinformation detection [7], [8]. For instance, Fig. 1 provides a real-world example, where a legitimate image is used as evidence to promote a false narrative, linking 5G technology to the transmission of COVID-19 when in reality, the image captures an antisurveillance protest in Hong Kong.[1]

Recent years have witnessed a surge in research attention towards automated multimodal misinformation detection [9] and especially out-of-context (OOC) images [7], [8]. Due to the scarcity of large-scale annotated datasets for OOC detection, researchers have increasingly turned to algorithmically generated datasets [7], [8], [10] which have facilitated the development of numerous OOC detection methods [11], [12], [13], [14], [15]. Nevertheless, solely relying on the analysis of an image-text pair is not always sufficient for detecting misinformation. More often than not, the incorporation and cross-examination of external information, i.e., evidence, is necessary [16]. To this end, researchers have explored methods for automated and evidence-based fact-checking, initially focusing on text-based approaches [17], [18], [19], [20] and, more recently, venturing into the realm of multimodal fact-checking (MFC) [21], [22], [23].

MFC methods utilize multiple pretrained and fine-tuned encoders, such as CLIP, ResNet, and BERT, to extract features and make use of attention-based networks [23] or coarse- and fine-grained attention mechanisms [24] to integrate external evidence, or use entity co-occurrence and semantic clustering of external evidence to assess stance toward the claim [25]. Nevertheless, current methods operated under the assumption that all collected external items from the Web were relevant and only addressed the task of verdict prediction.

Motivation: We aim to simulate a more realistic scenario where the MFC model has to distinguish between relevant and irrelevant information collected from the Web with the goal of enhancing verdict prediction accuracy.

Contributions: We incorporate relevant evidence detection (RED), as part of the MFC process. Specifically, we propose the relevant evidence detection directed transformer (*RED-DOT*) which comprises: "evidence reranking", "modality fusion", "verdict prediction", and "RED" modules. "Evidence reranking" is a preprocessing step that ranks evidence

[1]https://www.snopes.com/fact-check/5g-tower-torn-down-china-covid

Received 7 March 2024; revised 24 January 2025; accepted 17 March 2025. Date of publication 3 April 2025; date of current version 3 December 2025. This work was supported by the project "vera.ai: VERification Assisted by Artificial Intelligence" under Grant 101070093. *(Corresponding author: Stefanos-Iordanis Papadopoulos.)*

Stefanos-Iordanis Papadopoulos is with the Information Technologies Institute, Centre for Research & Technology-Hellas, 57001 Thessaloniki, Greece, and also affiliated with the Department of Electrical & Computer Engineering, Aristotle University of Thessaloniki (ECE-AUTh), 54124 Thessaloniki, Greece (e-mail: stefpapad@iti.gr).

Christos Koutlis and Symeon Papadopoulos are with the Information Technologies Institute, Centre for Research & Technology-Hellas, 57001 Thessaloniki, Greece (e-mail: ckoutlis@iti.gr; papadop@iti.gr).

Panagiotis C. Petrantonakis is with the Department of Electrical & Computer Engineering, Aristotle University of Thessaloniki (ECE-AUTh), 54124 Thessaloniki, Greece (e-mail: ppetrant@ece.auth.gr).

Digital Object Identifier 10.1109/TCSS.2025.3553939

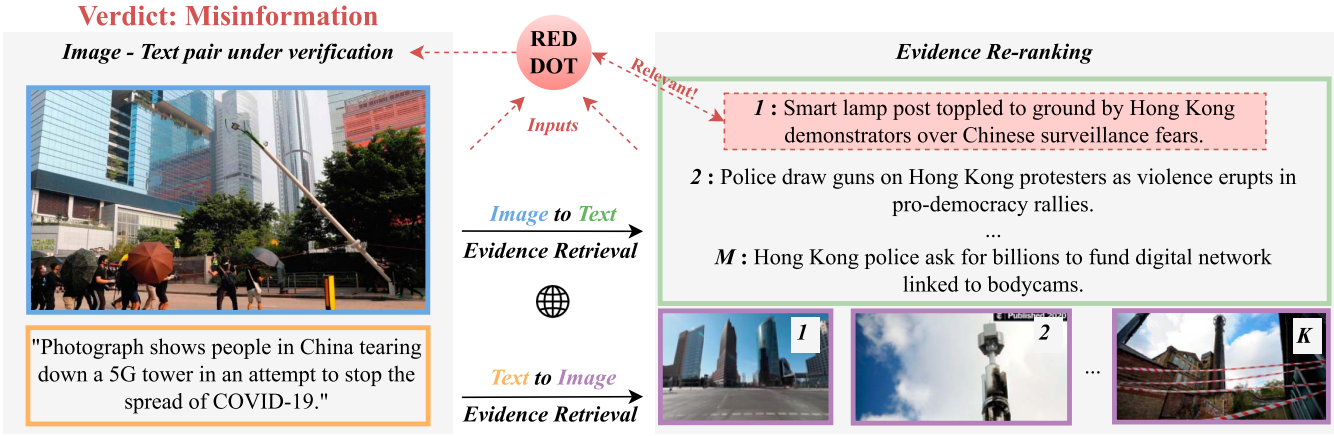

Fig. 1. Image–text pair under verification with external information (evidence), both images and texts, collected from the web. The proposed framework retrieves and reranks the evidence while RED-DOT determines which pieces of information are most relevant to (support or refute) the image–text pair and then uses those to determine the pair's veracity.

based on intramodal similarity while generating hard negative "irrelevant" samples to improve robustness. "Modality fusion" leverages element-wise operations between the two modalities to cross-check their relation. For feature extraction and evidence reranking we evaluate four vision-language models: CLIP ViT B/32, CLIP L/14 [26], ALBEF [27], and BLIP2 [28]. A Transformer encoder is used to further process the inputs. Then, "Verdict Prediction" classifies each item as either "true" or "misinformation" while "RED" examines all provided pieces of information, determines their relevance to the image–text pair, and uses the relevant ones to assess the pair's veracity. We explore numerous variants of RED-DOT, with different architectures (e.g., single or dual stage) and mechanisms (e.g., guided attention). RED-DOT is optimized via multitask learning on the NewsCLIPpings+ dataset [23] while its performance is also evaluated on the VERITE benchmark [29] comprising real-world multimodal misinformation.

Findings: We conduct extensive ablation and comparative experiments, revealing that: 1) evidence reranking delivers superior results by selecting the single most relevant piece of evidence per modality, as additional information often introduces noise and diminishes performance; 2) RED-DOT improves upon prior methods without external evidence by up to 33.7% and surpasses the RED-DOT-Baseline (without RED) by 8.9%; 3) RED-DOT outperforms the state-of-the-art on NewsCLIPpings+ by up to 3%, without relying on multiple backbone encoders, extensive evidence, or additional features.

II. Related Work

In recent years, automated misinformation detection has gained significant research attention, with an ongoing exploration of methods for identifying false information across textual [30], visual [31], and multimodal formats [32]. While several datasets have been developed for text-based fact-checking [17], [18], [19], [20], Glockner et al. [33] showed that most employ "leaked evidence" from fact-checked articles. This leads to an unrealistic setting that is not applicable for early detection and emerging misinformation detection. In the context of multimodal misinformation, the majority of research only considers the image-text pairs [7], [8], [10], [13], [29], [34], [35], [36], [37], [38], [39], [40] and do not incorporate external information or evidence; with few notable exceptions [21], [22], [23].

MOCHEG was developed as an end-to-end dataset for MFC, including evidence retrieval, verdict prediction, and explanation generation [21]. Top-k text evidence was retrieved using similarities from S-BERT and then reranked by a pretrained BERT model while image evidence was retrieved based on cross-modal similarities from CLIP. Both S-BERT and CLIP were fine-tuned with contrastive learning. However, while MOCHEG provides both text and image evidence, it only accommodates textual claims, not multimodal misinformation, so it can not be used for MFC. Moreover, the provided textual evidence can be considered "leaked" since they are collected solely from fact-check articles. In contrast, FACTIFY2 addresses multimodal evidence entailment as part of MFC, meaning the examination of whether an article and an image entails the information to support or refute a claim [22]. Recent studies relied on text summarization or top-k sentence retrieval from articles as part of the "evidence retrieval" process [41]. However, FACTIFY2 only provides a single article and image as evidence and incorporates fact-checking articles for the "refute" class, thus suffers from "leaked evidence". As a result, even simple baseline models can reach unrealistically high scores of 99%–100% in this class [41]; limiting the practical usefulness of the dataset. On the other hand, Abdelnabi et al. [23] use the NewsCLIPpings dataset [8] and augment it with external information collected from Google API. We refer to this dataset as NewsCLIPpings+ for simplicity. NewsCLIPpings is an algorithmically created dataset, where image-text pairs are decontextualized using the multimodal encoder CLIP, as well as scene and person matching computer vision models. Since the "falsified" pairs in NewsCLIPpings are algorithmically created and decontextualized -individual modalities are legitimate but their combination is falsified- it secures the absence of "leaked

evidence". These observations have been empirically supported by recent research, which found high rates of leaked evidence in both MOCHEG and FACTIFY (the "refute" class), and significantly lower rates in NewsCLIPpings+ [42].

Recent works have utilized NewsCLIPpings+ for MFC. The consistency checking network (CCN) [23] employs two attention-based memory networks for visual and textual reasoning, with ResNet152 and BERT, respectively, as well as a fine-tuned CLIP (ViT B/32) component for additional feature extraction. The stance extraction network (SEN) leverages a similar architecture to CCN but with ResNet50 (pretrained on Places365), ResNet152 (pretrained on ImageNet), and S-BERT as the backbone encoders [25]. Additionally, it extracts the "stance" of external evidence toward the image-text pair and calculates the "support-refutation score" based on the co-occurrence of named entities between the claim and the textual evidence. Finally, the explainable and context-enhanced network (ECENet) employs ResNet50, BERT, and CLIP ViT-B/32 for feature extraction on images, texts, evidence, and named entities within a "coarse- and fine-grained attention" network for intramodal and cross-modal feature reasoning [24]. Nevertheless, while the aforementioned methods demonstrated promising results on NewsCLIPpings+, they operated under the assumption that all collected information from the Web is relevant, which is not necessarily valid.

One limitation of NewsCLIPpings is that it relies on algorithmically created misinformation. This introduces uncertainty regarding the ability of models trained on it to generalize effectively to real-world multimodal misinformation. For this reason, researchers have been experimenting with training multimodal misinformation detection models on algorithmically created datasets such as NewsCLIPpings and MEIR [35] and evaluating their generalizability to real-world misinformation [10]. Due to unimodal biases in widely adopted benchmarks, such as VMU Twitter [43] and COSMOS [7], we recently developed VERITE, an evaluation benchmark for real-world multimodal misinformation detection that accounts for unimodal biases [29]. However, VERITE and other similar evaluation datasets have not yet been used for MFC since they do not provide external evidence.

In light of these considerations, we introduce a new architecture that incorporates "relevant evidence detection" (RED) as a means of improving verdict prediction accuracy. We opt for training our model on NewsCLIPpings+ while also evaluating its generalizability to VERITE, after augmenting it through evidence collection from the Web.

III. Methodology

A. Problem Formulation

We define the tasks of multimodal fact-checking (MFC) and relevant evidence detection (RED) as follows: given a dataset $(I_i^v, T_i^v, \mathcal{I}_i^e, \mathcal{T}_i^e, y_i^v, \boldsymbol{y}_i^e)_{i=1}^N$, where (I_i^v, T_i^v) represents the image-text pair under verification (pair, from now on), $\mathcal{I}_i^e = [I_{i1}^e, I_{i2}^e, \ldots, I_{i2K}^e]$ and $\mathcal{T}_i^e = [T_{i1}^e, T_{i2}^e, \ldots, T_{i2M}^e]$ represent an array of image evidence with $2 \cdot K$ elements (K relevant + K irrelevant) and text evidence with $2 \cdot M$ elements (M relevant + M irrelevant), respectively, $y_i^v \in 0, 1$ denotes the overall verdict label, truthful (0) or misinformation (1) and $\boldsymbol{y}_i^e = (y_{i1}^e, y_{i2}^e, \ldots, y_{i2\cdot(M+K)}^e)$ is an array of binary labels of $2 \cdot (M + K)$ size denoting whether each piece of evidence is relevant (1) or irrelevant (0) to the pair; with the primary objective of training a classifier $f : (\mathcal{I}^v, \mathcal{T}^v, \mathcal{I}^e, \mathcal{T}^e) \rightarrow (\hat{y}^v, \hat{\boldsymbol{y}}^e)$. During training, balanced values for M and K are employed for both relevant and irrelevant evidence to mitigate class imbalance, which could otherwise lead the model to overlook the minority labels.

B. Evidence Retrieval and Reranking

We implement a preprocessing stage to rerank collected evidence from the Web and to retrieve irrelevant evidence for hard negative sampling. To this end, we leverage the NewsCLIPpings+ dataset [23] comprising 85 360 image–text pairs, balanced in terms of truthful and misinformation pairs. The authors use the text T_i^v to retrieve visual evidence $\mathcal{I}_i^e$ and the image I_i^v to retrieve textual evidence $\mathcal{T}_i^e$. Each pair is associated with up to 19 textual evidence and up to 10 visual evidence, with a total of 146 032 text evidence and 736 731 image evidence; collected via Google API. These items are not "robust evidence" in the sense that they necessarily support or refute the claim but rather "potential evidence", nevertheless, we use the term "evidence" for consistency with most papers in the field. While some level of relevance between the collected evidence and the pair is expected, we undertake the task of reranking all gathered evidence to heighten the likelihood of their relevance.

Following [23], we use the pretrained CLIP ViT B/32 [26] as the backbone encoder to extract visual $F_I \in R^{\dim \times 1}$ and textual features $F_T \in R^{\dim \times 1}$ with dimensionality $\dim = 512$, for $\mathcal{I}^v, \mathcal{I}^e$ and $\mathcal{T}^v, \mathcal{T}^e$. Additionally, we experiment with ALBEF Base [27], BLIP2 ViT L [28], and CLIP ViT L/14 with $\dim = 768$, all taken from the LAVIS[2] library. We signify the ranked or "relevant" evidence with the superscript "+", and rank them based on the intramodal cosine similarity sim as follows:

$$\begin{aligned} \mathcal{I}_i^+ &= \underset{I_j^e \in \mathcal{I}_i^e}{\text{argsort}}\, \text{sim}(F_{I_i^v}, F_{I_j^e}) \\ \mathcal{T}_i^+ &= \underset{T_j^e \in \mathcal{T}_i^e}{\text{argsort}}\, \text{sim}(F_{T_i^v}, F_{T_j^e}). \end{aligned} \tag{1}$$

For the "irrelevant" evidence class, denoted by the superscript "-", we leverage hard negative sampling, instead of depending on random sampling. More specifically, we calculate the most similar item based on text–text similarity, fetch its ranked evidence and employ them as $\mathcal{I}_i^-$, and use image–image similarity for $\mathcal{T}_i^-$. We incorporate the "opposite" modality in similarity calculations to mimic the evidence retrieval process of NewsCLIPpings+. The process of hard negative sampling "irrelevant" evidence is illustrated in Fig. 2 and can be expressed as follows:

$$\begin{aligned} \mathcal{I}_i^- &= \mathcal{I}_j^+, \text{ where } j = \underset{T_j^v \in \mathcal{T}^v, j \neq i}{\text{argmax}}\, \text{sim}(F_{T_i^v}, F_{T_j^v}) \\ \mathcal{T}_i^- &= \mathcal{T}_j^+, \text{ where } j = \underset{I_j^v \in \mathcal{I}^v, j \neq i}{\text{argmax}}\, \text{sim}(F_{I_i^v}, F_{I_j^v}). \end{aligned} \tag{2}$$

[2]https://github.com/salesforce/LAVIS

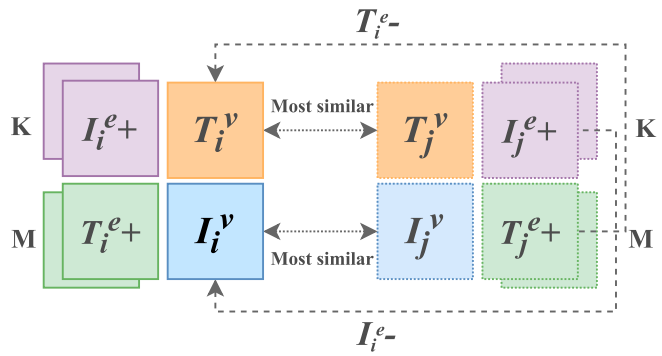

Fig. 2. Visualization of (2). Hard negative sampling for retrieving "irrelevant" evidence.

We make use of efficient indexing (Meta's FAISS[3]) to improve scalability and speed.

After retrieving relevant evidence $\mathcal{E}^+ = (\mathcal{T}^+, \mathcal{I}^+)$ and irrelevant $\mathcal{E}^- = (\mathcal{T}^-, \mathcal{I}^-)$, we also define $\boldsymbol{y}^e+ = [1, 1, \ldots, 1]$ and $\boldsymbol{y}^e- = [0, 0, \ldots, 0]$, each having size M + K. To avoid overfitting based on positional patterns during training, we combine all evidence $\mathcal{E}' = \mathcal{E}^+ \cup \mathcal{E}^-$ and labels $\boldsymbol{y'}^e = [1, 1, ...1, 0, 0, ..., 0]$, shuffle their positions with P representing the permutation positions of the elements and the final evidence and labels become: $\mathcal{E} = \mathcal{E}'[P]$ and $\boldsymbol{y}^e = \boldsymbol{y'}^e[P]$. Evidence $\mathcal{E}$ should consist of $2 \cdot (M + K)$ items. For instance, setting $M = 1$ signifies our expectation of having one relevant and one irrelevant textual evidence. If this requirement is not met, we randomly sample additional items from the dataset.

C. Modality Fusion Module

To enhance the cross-examination of image–text pairs and accentuate specific consistencies or inconsistencies, we utilize a range of multimodal fusion operations. In addition to simple concatenation of the two modalities' features, denoted as $[F_{I^v}; F_{T^v}]$, we also employ "addition", "subtraction" and "multiplication". Our rationale is that "addition" emphasizes complementarity, "subtraction" accentuates differences and "multiplication" underscores shared aspects. We express the modality fusion module as follows:

$$F^v = [F_{I^v}; F_{T^v}; F_{I^v} + F_{T^v}; F_{I^v} - F_{T^v}; F_{I^v} * F_{T^v}] \quad (3)$$

with $F^v \in R^{\dim \times 5}$. Prior research in multimodal fusion has explored different element-wise operations, including multiplication [44] and outer product [45]. However, to the best of our knowledge, the combination of multiple fusion operations remains unexplored, especially for MFC. Furthermore, we employ a Transformer encoder $D(\cdot)$ to further process the fused modalities. Within $D(\cdot)$, we introduce a trainable classification token (CLS) that acts as a global representation for all inputs. For instance, in the RED-DOT-Baseline experiments that do not make use of "irrelevant evidence", $D(\cdot)$ is expressed as follows:

$$[\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F^v}}, \mathbf{d}_{\mathbf{F}_{\mathcal{E}+}}] = D([\text{CLS}; F^v; F_{\mathcal{E}+}]). \quad (4)$$

[3]https://ai.meta.com/tools/faiss

D. Verdict Prediction Module

The processed classification token $\mathbf{d}_{\text{CLS}}$ is further used to obtain the predicted verdict $\hat{y}^v$ as follows:

$$\hat{y}^v = \mathbf{W}_1 \cdot \text{GELU}(\mathbf{W}_0 \cdot \text{LN}(\mathbf{d}_{\text{CLS}})) \quad (5)$$

where LN stands for Layer Normalization, $\mathbf{W}_0 \in \mathbb{R}^{(\dim/2)\times \dim}$ is a GELU activated fully connected layer and $\mathbf{W}_1 \in \mathbb{R}^{1\times(\dim/2)}$ is the final verdict classification layer. The network is optimized for binary classification based on the binary cross entropy (BCE) loss function L^v, after applying the sigmoid activation function on $\hat{y}^v$.

E. RED

The RED module is responsible for discerning which evidence in $\mathcal{E}$ are relevant or irrelevant to the claim under verification. Given that our work introduces and explores the RED task for the first time, we deem important to investigate diverse architectural approaches which can also serve as "strong baselines" for future research in this field. All methods utilize Transformer $D(\cdot)$ as shown in (6) with $\mathbf{d}_{F_\mathcal{E}}$ as shown in (7) and (5) to predict $\hat{y_v}$. Moreover, all methods make use of multitask learning to predict both binary $\hat{y}^v$ and multilabel $\hat{\boldsymbol{y}}^e$ and are optimized based on $L = L^v + L^e$ with L^e being the average BCE loss for multilabel classification, after applying the sigmoid activation function on $\hat{\boldsymbol{y}}^e$. An overview of the RED-DOT architecture can be seen in Fig. 3.

$$[\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F^v}}, \mathbf{d}_{\mathbf{F}_\mathcal{E}}] = D([\text{CLS}; F^v; F_\mathcal{E}]) \quad (6)$$

$$\mathbf{d}_{F_\mathcal{E}} = [\mathbf{d}_{\mathbf{F}_{\mathcal{E}_1}}, \ldots, \mathbf{d}_{\mathbf{F}_{\mathcal{E}_{(\mathbf{M}+\mathbf{K})\cdot\mathbf{2}}}}]. \quad (7)$$

1) Single-Stage Learning (SSL): Employs (6) and (8) to predict $\hat{\boldsymbol{y}}^e$ in a single stage

$$\hat{y}^e = [\mathbf{W}_3 \cdot \text{GELU}(\mathbf{W}2 \cdot \text{LN}(\mathbf{d}_{\mathbf{F}_{\mathcal{E}_i}}))], \quad \text{for } i = 1 \text{ to } 2 \cdot (M + K) \quad (8)$$

with $\mathbf{W}_2 \in \mathbb{R}^{(\dim/2)\times \dim}$ is a GELU activated fully connected layer and $\mathbf{W}_3 \in \mathbb{R}^{1\times(\dim/2)}$.

2) Single-Stage Learning With Guided Attention (SSL + GA): Similar to SSL but instead of (8), we apply a type of "guided attention" where the L_v is directly applied to the attention weights. Consider the vector $\mathbf{d} = [\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F^v}}, \mathbf{d}_{\mathbf{F}_\mathcal{E}}]$, to calculate the attention scores, we use (9) and (10) to predict $\hat{\boldsymbol{y}}^e$

$$\mathbf{a} = \frac{\mathbf{d} \cdot \mathbf{d}^\mathbf{T}}{\dim} \quad (9)$$

$$\hat{y}^e = \mathbf{a}[< \text{CLS} >][-2 \cdot (M + K) :] \quad (10)$$

where $< \text{CLS} >$ represents the position of the classification token, providing a global representation of attention and $-2 \cdot (M + K)$ denotes the last $2 \cdot (M + K)$ items in $\mathbf{d}$, corresponding to the evidence.

3) Dual-Stage Learning (DSL): First employs (6) and (8). Afterwards evidence $\mathcal{E}$ are masked by $\text{MASK} \in \{0, 1\}^{\dim \times 2\cdot(M+K)}$ where 0 s denote predicted irrelevant and 1 s predicted relevant evidence. During training, we employ teacher enforcing. In the second stage, we apply

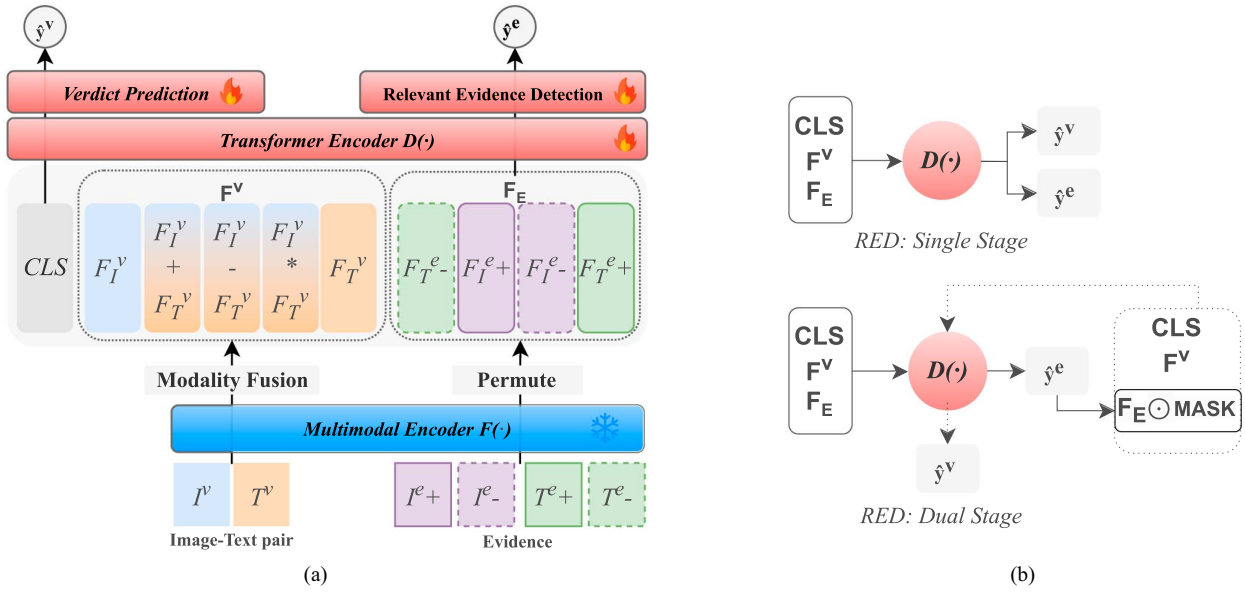

Fig. 3. (a) Overview of the proposed architecture utilizing a pretrained backbone encoder $F(\cdot)$ for feature extraction, the "modality fusion" module for combining the image and text embeddings, while external evidence is randomly permuted during training to mitigate positional biases. A Transformer encoder $D(\cdot)$ processes all inputs along with a CLS token. The "verdict prediction" network uses the transformed CLS token to classify the verdict as either "true" or "misinformation", while the relevant evidence detection (RED) network classifies each input evidence as either "relevant" or "irrelevant". (b) High-level overview of the single and dual stage RED variants. Dotted lines represent the second stage in dual-stage learning (DSL).

MASK onto the evidence and reprocess with Transformer $D(\cdot)$ with (11) and then use (5) to predict $\hat{y}^v$

$$[\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F}^v}, \mathbf{d}_{\mathbf{F}_\mathcal{E}}] = D([\text{CLS}; F^v; F_\mathcal{E} \odot \text{MASK}]). \quad (11)$$

4) Dual-Stage Learning With Guided Attention (DSL + GA): Similar to DSL, but in the first stage it utilizes guided attention [as in (9) and (10)] instead of (8) to predict $\hat{\boldsymbol{y}}^e$. The second stage remains identical to DSL.

5) Dual-Stage Learning With Two Transformers (DSL + D2): Similar to DSL, but in the second stage we employ a second identical transformer encoder $D2(\cdot)$ with $[\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F}^v}, \mathbf{d}_{\mathbf{F}_\mathcal{E}}] = D2([\text{CLS}; F^v; F_\mathcal{E} \odot \text{MASK}])$. The first stage remains identical to DSL.

IV. Experimental Setup

A. Evaluation Protocol

Our goal is to examine the generalizability of RED-DOT to real-world misinformation. For this reason, we train RED-DOT on the NewsCLIPpings+ dataset and evaluate its performance on the out-of-context (OOC) pairs from the VERITE benchmark. To gather "external evidence" via Google API for VERITE, we followed the process outlined [23]; which we make publicly available to facilitate future research and ensure fair comparability across studies.

Our evaluation protocol departs from "in distribution validation" (ID-V) which involves validation and check-pointing on NewsCLIPpings and final testing on VERITE [29]. Instead, we make use of "out-of-distribution cross validation" (OOD-CV) with k-fold cross-validation and check-pointing directly on a VERITE fold (k-fold = 3), while training RED-DOT on the NewsCLIPpings+ training set. Partially inspired by [46], we hypothesize that due to NewsCLIPpings+ and VERITE following different distributions -the first comprising algorithmically created samples and the latter real-world misinformation- OOD-CV can capture a version of the model with improved generalizability onto the "out-of-distribution" VERITE. We report "accuracy" on NewsCLIPpings+ based on ID-V and "true versus OOC accuracy" and the standard deviation on VERITE based on OOD-CV.

During the evaluation of RED-DOT on VERITE we make no assumptions about the relevance of the retrieved evidence, we solely employ "M,K" items retrieved by Google API and reranked by our pipeline, without artificially injecting "negative evidence". The model has to determine which items are relevant while producing the final verdict, thus simulating automated fact-checking "in the wild".

B. Competing Methods

On NewsCLIPpings+, we compare RED-DOT against CCN [23], ECENet [24], and SEN [25] which incorporate external evidence. For more information about these models, see Section II. Moreover, we include methods that do not incorporate external evidence, namely: CLIP [47] and self-supervised distilled learning (SSDL) [13] and the detector transformer (DT) [10]. On VERITE, we compare *RED-DOT* against the DT [29] employing features from CLIP ViT-L/14 after being trained on "R-NESt + CHASMA-D + NC-t2t"; a combination of three different algorithmically created datasets with 437 673 samples in total. To ensure comparability, we faithfully replicate the DT and train it on NewsCLIPpings+ without external evidence. Within our framework, the DT can be expressed as

TABLE I
VARIANTS OF RED-DOT (W/ CLIP VIT B/32) TRAINED UNDER THE OOD-CV PROTOCOL WITH M TEXTS AND K IMAGES AS EVIDENCE

Method	M,K=1 (ID-V)	M,K=1	M,K=2	M,K=4
Baseline	69.7	70.7 (1.5)	65.9 (2.3)	66.4 (2.3)
SSL	70.0	71.8 (0.1)	72.1 (3.7)	67.2 (1.8)
SSL + GA	71.2	72.6 (1.7)	71.9 (3.0)	65.8 (2.7)
DSL + D2	69.1	71.3 (0.5)	66.5 (2.6)	65.3 (0.7)
DSL + GA	71.3	73.6 (1.7)	69.9 (2.5)	62.7 (1.7)
DSL	70.9	73.9 (0.5)	70.0 (2.4)	64.7 (1.6)
Mean	70.4	**72.3** (1.0)	69.4 (2.8)	65.3 (1.8)

Note: We report Accuracy on VERITE and the standard deviation (in parentheses) for OOD-CV (k-fold = 3). For $M, K = 1$ we also report detection Accuracy under ID-V. Underline denotes the best performance per method. The bold denotes the best mean performance. The underline denotes the best performance per method.

TABLE II
ABLATION OF RED-DOT (W/ CLIP VIT B/32) WITH DIFFERENT MODALITY FUSION COMBINATIONS

	$F_{Iv}; F_{Tv}$	F^v	F^v - '−'	F^v - '+'	F^v - '*'
Baseline	66.7 (0.9)	70.7 (1.5)	68.3 (1.2)	70.3 (1.7)	71.6 (4.1)
SSL	71.6 (1.2)	71.8 (0.1)	73.2 (1.2)	73.9 (1.1)	72.9 (0.3)
SSL + GA	70.6 (0.8)	72.6 (1.7)	72.0 (1.5)	71.6 (2.4)	70.5 (2.1)
DSL + D2	68.8 (1.6)	71.3 (0.5)	70.0 (0.8)	69.8 (1.7)	69.3 (1.7)
DSL + GA	69.8 (1.1)	73.6 (1.7)	71.2 (2.0)	73.6 (1.7)	72.8 (1.1)
DSL	71.0 (0.4)	73.9 (0.5)	71.5 (2.2)	72.9 (1.1)	74.2 (1.5)
Mean	69.8 (1.0)	**72.3** (1.0)	71.0 (1.5)	72.0 (1.6)	71.8 (1.8)

Note: F^v followed by "-" signifies removal of one fusion operation. We report Accuracy on VERITE and the standard deviation (in parentheses). The bold denotes the best mean performance. The underline denotes the best performance per method.

$D([\text{CLS}; F_{I^v}; F_{T^v}])$ with features from CLIP ViT-L/14 and is trained under the ID-V protocol.

C. Implementation Details

We train RED-DOT for a maximum of 100 epochs with early stopping at 10 epochs, using the Adam optimizer and a learning rate of $lr = 1e-4$. When tuning hyperparameters, we consider the following configurations: $t \in \{4, 6\}$ transformer layers, $z \in \{128, 2048\}$ for the dimension of the feed-forward network, and $h \in \{2, 8\}$ for the number of attention heads. The dropout rate is set at 0.1 and the batch size to 512. For faithful reproduction of DT, we follow the implementation details of [29] which uses $lr = 5e-5$ and $t \in \{1, 4\}$. Finally, to ensure experiment reproducibility, we use a constant random seed (0) for PyTorch, Python random, and NumPy.

V. EXPERIMENTAL RESULTS

A. Out-of-Distribution Cross-Validation

First, we compare the performance of the two evaluation protocols, ID-V and OOD-CV, as described in Section IV-A. The first two columns of Table I illustrate the performance of RED-DOT variants, using $M = 1$ text and $K = 1$ image as evidence, when trained and validated under each protocol. OOD-CV consistently outperforms ID-V, achieving a mean accuracy of 72.3% compared with 70.4%. These results indicate that OOD-CV is more effective at capturing a model's ability to generalize from algorithmically generated samples (NewsCLIPpings+) to real-world out-of-context misinformation (VERITE). Consequently, we adopt OOD-CV for all subsequent experiments to ensure robust and representative performance on the VERITE dataset.

B. Impact of Multiple Ranked Evidence

Next, we examine the effect of incorporating varying amounts of external evidence. In Table I, we observe that the best performance is yielded with $M, K = 1$, with an average accuracy of 72.3%, while the performance tends to decline with the inclusion of additional evidence. As discussed in Section III-B, "evidence" in NewsCLIPpings+ refers to external information retrieved from search engines rather than being curated or annotated by professional fact-checkers. Consequently, not all retrieved Web items are genuinely relevant. Therefore, setting "M,K+ > 1" introduces additional items in $\mathcal{E}+$ that are often less relevant and less informative, but are nonetheless labeled as relevant ($\boldsymbol{y}^e+$), despite not necessarily being so, thereby introducing noise that confuses the network and degrades its performance. This is exemplified in Fig. 1, where only the top-ranked text evidence is genuinely relevant to support the image-text pair under verification, while other items pertains to unrelated events. We observe a similar outcome with RED-DOT-Baseline, which does not utilize irrelevant evidence, yet still shows a decline in accuracy as additional items are introduced in $\mathcal{E}+$. These results suggest that the evidence re-ranking process described in Section III-B offers advantages over utilizing all collected evidence from the Web, improving detection accuracy while also reducing computational complexity by minimizing the number of items the network must process. We employ $M, K = 1$ in all the following experiments.

C. Ablation on Modality Fusion

Here, we conduct an ablation study to assess the significance of each operation within the modality fusion module for RED-DOT variants, as shown in Table II. Our findings indicate that the simple concatenation $(F_{I^v}; F_{T^v})$ results in the lowest average performance (69.8%) across all RED-DOT variants while employing all fusion operations (F^v) produces the highest (72.3%). However, there are individual instances where alternative fusion operations demonstrate superior performance. For instance, RED-DOT-DSL attains the highest overall accuracy while utilizing $F^v -$ '*' (removing the multiplication operation), yielding 74.2% and surpassing the 73.9% accuracy achieved with F^v. Notably, RED-DOT-SSL yields 73.9% without "addition" ($F^v -$ '+') and 71.8% with F^v; a +2.9% improvement. Based on these outcomes, we can conclude that employing F^v can lead to high performance, but if optimal performance is required, it is advisable to experiment with multiple multimodal fusion configurations.

D. Comparative Study: NewsCLIPpings

We conduct a comparative study between RED-DOT variants and current state-of-the-art methods on NewsCLIPpings+ and VERITE to evaluate: 1) the performance of RED-DOT variants

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART

Method	Encoder	M,K+	M,K−	NewsCLIPpings	VERITE
CLIP [8]	CLIP ViT B/32 (fine-tuned)	0	0	60.2	-
DT [10]	CLIP ViT B/32	0	0	65.7	-
DT [10]	CLIP ViT L/14	0	0	77.1	-
SSDL [13]	CLIP ViT B/32	0	0	67.0	-
SSDL [13]	CLIP RN50 (fine-tuned)	0	0	71.0	-
DT [29]	CLIP ViT L/14	0	0	-	57.5
DT [29]†	CLIP ViT L/14	0	0	-	72.7
RED-DOT-Baseline	ALBEF Base	0	0	54.6	54.9 (3.4)
RED-DOT-Baseline	BLIP2 ViT L	0	0	68.2	61.1 (2.4)
RED-DOT-Baseline	CLIP ViT B/32	0	0	71.6	63.7 (2.8)
RED-DOT-Baseline	CLIP ViT L/14	0	0	81.7	75.5 (3.7)
CCN [23]	CLIP ViT B/32 (fine-tuned) + ResNet152 + BERT (fine-tuned).	ALL	0	84.7	-
SEN [25]	ResNet50 (Places365), ResNet152 (ImageNet), S-BERT	ALL	0	87.1	-
ECENet [24]	CLIP ViT B/32. BERT, ResNet50	ALL	0	87.7	-
RED-DOT-Baseline	ALBEF Base	1	0	80.8	71.5 (0.3)
RED-DOT-Baseline	BLIP2 ViT L	1	0	85.7	71.5 (3.0)
RED-DOT-Baseline	CLIP ViT B/32	1	0	87.8	70.7 (1.5)
RED-DOT-Baseline	CLIP ViT L/14	1	0	**90.3**	70.6 (3.1)
RED-DOT-DSL	ALBEF Base	1	1	80.1	72.0 (2.3)
RED-DOT-DSL	BLIP2 ViT L	1	1	84.8	72.7 (2.6)
RED-DOT-DSL	CLIP ViT B/32	1	1	84.5	73.9 (0.5)
RED-DOT-SSL	CLIP ViT L/14	1	1	87.9	**76.9** (5.4)

Note: All methods are trained on NewsCLIPpings with the exception of † denoting "R-NESt + CHASMA-D + NC-t2t" [29]. Here, $M, K+$ and $M, K-$ represent the number of relevant and irrelevant evidence, respectively. We report Accuracy on NewsCLIPpings+ and "True vs OOC" Accuracy as well as the standard deviation (in parentheses) on VERITE. The bold denotes the best overall performance.

TABLE IV
POINT-BISERIAL CORRELATION COEFFICIENT (R) AND P-VALUES (P) (IN PARENTHESES) APPLIED ON THE SIMILARITY BETWEEN VARIOUS FEATURES EXTRACTED EITHER WITH CLIP B/32 OR L/14

	NewsCLIPpings+		VERITE	
Similarity	B/32	L/14	B/32	L/14
F_{I^v}, F_{T^v}	−0.38 (0.0)	−0.64 (0.0)	−0.46 (1.4e−35)	−0.59 (6.2e−63)
F_{I^v}, F_{I^e}	−0.51 (0.0)	−0.55 (0.0)	−0.30 (0.1e−15)	−0.33 (2.2e−18)
F_{T^v}, F_{T^e}	−0.28 (0.0)	−0.37 (0.0)	**−0.03 (0.47)**	**−0.03 (0.42)**
F_{T^e}, F_{I^e}	−0.17 (0.0)	−0.28 (0.0)	**−0.03 (0.48)**	**−0.07 (0.08)**

Note: Bold denote relations without statistical significance ($p > 0.05$).

against the state-of-the-art with and without external evidence; 2) the impact of different backbone encoders; and 3) the factors contributing to the improved performance of RED-DOT.

Starting with the NewsCLIPpings dataset, as shown in Table III, we observe that without external evidence, RED-DOT-Baseline (M,K+ =0) outperforms prior methods, namely CLIP (60.2), DT (65.7), and SSDL (67.1), scoring 71.6% and 81.7% while using CLIP ViT B/32 and L/14 respectively. All other factors being equal, this can be primarily attributed to our element-wise modality fusion integrated into the Transformer-based architecture.

Moreover, we observe that all methods that leverage external information significantly outperform methods that do not. With CLIP ViT B/32 as the backbone encoder, RED-DOT-Baseline (M,K+ =1) yields 87.8% on NewsCLIPpings+, competing with and even outperforming the SotA, namely CCN (84.7%), SEN (87.1%) and ECENet (87.7%), while only employing a single piece of evidence per modality and without requiring fine-tuning of the backbone encoder, nor multiple encoders (e.g., BERT and ResNet on top of CLIP) nor additional features (e.g., named entities).

Similarly, with CLIP ViT L/14, RED-DOT-Baseline (M,K+ =1) significantly outperforms the SotA, reaching 90.3% Accuracy, translating into a +3.0% relative improvement over ECENet. This is followed by RED-DOT-SSL, reaching 87.9% on NewsCLIPpings+ and striking the best balance between high performance on both NewsCLIPpings+ and VERITE. These results highlight the advantage of the proposed "evidence reranking" and "modality fusion" modules.

E. Comparative Study: VERITE

In regards to VERITE, leveraging CLIP ViT L/14 as the backbone in *RED-DOT* yields consistently higher performance than the alternatives. More specifically, DT (M,K+ =0) trained on NewsCLIPpings+ with features from CLIP ViT L/14 yields 57.5% accuracy on VERITE while *RED-DOT*-Baseline (M,K+ =0) reaches 75.5% without external evidence surpassing its counterparts that employ ALBEF (54.6), BLIP2 (68.2), or CLIP ViT B/32 (63.7). We can deduce that the selection of the backbone encoder plays a pivotal role. However, the most substantial difference between DT and RED-DOT-Baseline is that the latter leverages our proposed modality fusion module, which demonstrates a noteworthy enhancement in performance. On top of that, by leveraging CLIP ViT L/14 and the proposed RED module, RED-DOT-SSL can further improve performance up to 76.9%, a notable +33.7% relative improvement over *DT*, +1.9% over RED-DOT-Baseline without evidence and +8.9% with evidence. With CLIP B/32, RED-DOT-DSL outperforms *DT* by a notable +28.5%, RED-DOT-Baseline without evidence by +16% and with evidence by +4.5%.

Integrating external evidence into RED-DOT-Baseline enhances performance on VERITE in three out of four encoders (ALBEF, BLIP2, CLIP ViT B/32), but not with CLIP ViT L/14,

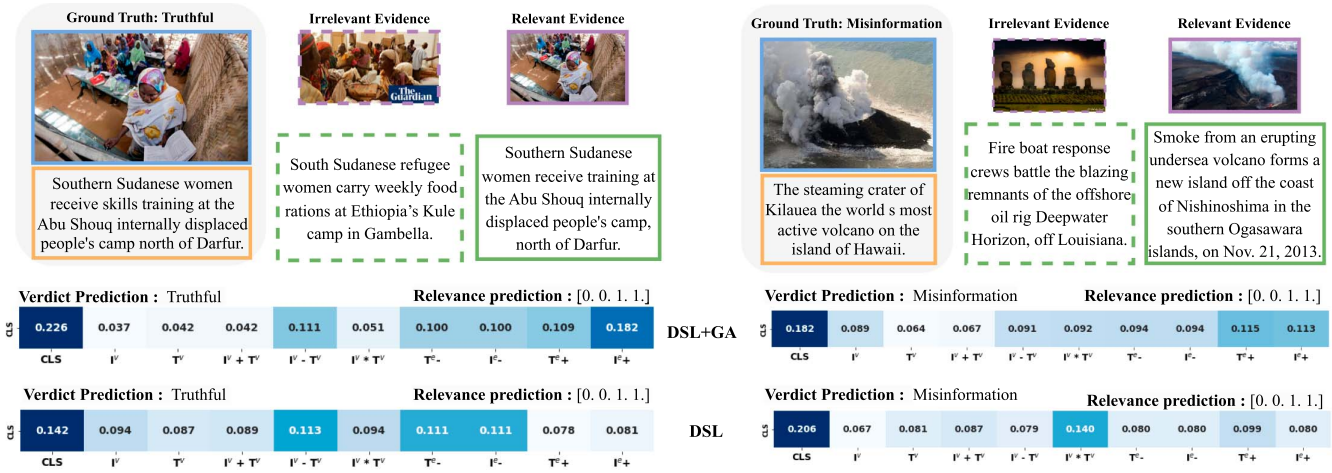

Fig. 4. Inference by RED-DOT variants: DSL and DSL+GA (w/ CLIP ViT B/32) on samples taken from NewsCLIPpings+. We report the attention scores of each method. "Relevance ground truth" is set to [0, 0, 1, 1] for simplicity, regarding [T^e-, I^e-, T^e+, I^e+], respectively.

where performance drops on VERITE despite high performance on NewsCLIPpings+. This discrepancy can be attributed to differing correlation patterns between features extracted by various backbone encoders on NewsCLIPpings+ and VERITE. As seen in Table IV, while image–image and image–text correlations are strong, and statistically significant, on both datasets with CLIP L/14, correlations between textual and visual evidence are weaker on VERITE, thus, training with external evidence enhances performance on NewsCLIPpings+, but not VERITE. Importantly, integrating RED improves performance on VERITE with all backbone encoders, and RED-DOT-SSL using CLIP ViT L/14 achieves the best balance on both datasets. These findings support our hypothesis that multitask learning with "verdict prediction" and "relevant evidence detection" improves detection accuracy. Finally, we retrained RED-DOT-SSL and RED-DOT-Baseline using CLIP ViT L/14 across 10 random seeds, achieving mean accuracies of 76.26% ($\pm$1.0) and 74.92% ($\pm$0.8) on VERITE, respectively, which correspond to a p-value of 0.02 based on the Mann-Whitney U-test, demonstrating statistical significance at a 0.05 p-level.

F. Qualitative Analysis

Fig. 4 shows the inference process on NewsCLIPpings+ test set samples for two RED-DOT variants, DSL and DSL+GA. Both models correctly predict the overall verdict and evidence relevance. However, DSL+GA assigns higher attention scores to relevant items. For example, in the "truthful" pair, the evidence retrieval and reranking module retrieves the same image I^e+ as I^v+, and DSL+GA correctly assigns higher attention scores to I^e+ (0.182), followed by T^e+ (0.109), and lower scores to irrelevant evidence (0.100). In contrast, DSL shows lower attention on I^e+ (0.081) and T^e+ (0.078), with higher scores on irrelevant evidence (0.111). In the "misinformation" pair, DSL+GA assigns higher attention scores to the relevant evidence, while DSL assigns equal scores to irrelevant items, including I^e+. DSL+GA's higher attention on relevant texts and images can potentially enhance interpretability.

VI. Conclusion

In this study we address the challenge of evidence-based MFC and incorporate RED as part of the process, where the model, first has to determine which pieces of evidence are relevant to support or refute the claim under verification and then proceed to assess its veracity. We conduct extensive ablation and comparative experiments to show that the proposed RED-DOT is capable of outperforming its counterparts that are not optimized for RED on VERITE. Moreover, it outperforms the state-of-the-art on NewsCLIPpings+ without requiring numerous evidence, multiple or fine-tuned backbone encoders or additional features.

Despite notable improvements in the field of MFC, it is essential to discuss certain limitations of our research and the field more broadly. Following previous studies [23], [24], [25] we focused on NewsCLIPpings+, which, despite being the only publicly available multimodal dataset with external, nonleaked evidence, it comes with certain constraints. Firstly, the dataset consists of a mere 85 000 samples which may not encompass the full spectrum of diversity concerning historical events, prominent figures, and other crucial contextual elements pertinent to misinformation. Furthermore, NewsCLIPpings+ exclusively focuses on "repurposed images" or "out-of-context misinformation", omitting categories such as "miscaptioned images" [29] and only provides short textual evidence -article titles and captions- that may offer limited information in comparison to complete articles. Therefore, subsequent research endeavors should aim at the collection of more extensive and diverse datasets, that also include full texts as candidate evidence and encompass various forms of multimodal misinformation.

Finally, one should acknowledge that the external information used as "evidence" is retrieved from search engines, which may introduce irrelevant or noisy data, limiting the full potential of MFC methods. Future research should explore improved strategies for gathering, filtering, and evaluating external evidence. Our work takes a first step in this direction by introducing a robust framework for filtering relevant external evidence,

paving the way for more accurate and scalable fact-checking models. Additionally, future studies could examine the use of knowledge databases and structured fact-checking sources to improve retrieval reliability while also considering the challenge of "leaked evidence" [33], [42]. Future research should address these challenges by integrating multilingual capabilities and incorporating additional modalities (e.g., audio and video), to enhance the effectiveness of MFC in combating the cross-linguistic [48] and multiformat [49] spread of misinformation.

REFERENCES

[1] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, "Fake news on social media: the impact on society," *Inf. Syst. Frontiers*, vol. 26, no. 2, pp. 443–458, 2024.

[2] W. L. Bennett and S. Livingston, "The disinformation order: Disruptive communication and the decline of democratic institutions," *Eur. J. Commun.*, vol. 33, no. 2, pp. 122–139, 2018.

[3] J. Roozenbeek et al., "Susceptibility to misinformation about Covid-19 around the world," *Roy. Soc. open Sci.*, vol. 7, no. 10, 2020, Art. no. 201199.

[4] I. J. B. Do Nascimento et al., "Infodemics and health misinformation: a systematic review of reviews," *Bull. World Health Org.*, vol. 100, no. 9, p. 544, 2022.

[5] J. Gamir-Ríos et al., "Multimodal disinformation about otherness on the Internet. the spread of racist, xenophobic and islamophobic fake news in 2020," *Anàlisi*, no. 64, pp. 49–64, 2021.

[6] D. Xu, S. Fan, and M. Kankanhalli, "Combating misinformation in the era of generative ai models," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 9291–9298.

[7] S. Aneja, C. Bregler, and M. Nießner, "COSMOS: Catching out-of-context image misuse using self-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 12, 2023, pp. 14084–14092.

[8] G. Luo, T. Darrell, and A. Rohrbach, "Newsclippings: Automatic generation of out-of-context multimodal media," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6801–6817.

[9] S. Hangloo and B. Arora, "Combating multimodal fake news on social media: methods, datasets, and future perspective," *Multimedia Syst.*, vol. 28, no. 6, pp. 2391–2422, 2022.

[10] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. Petrantonakis, "Synthetic misinformers: Generating and combating multimodal misinformation," in *Proc. 2nd ACM Int. Workshop Multimedia AI Against Disinformation*, 2023, pp. 36–44.

[11] S. Aneja et al., "MMSys' 21 grand challenge on detecting cheapfakes," 2021, *arXiv:2107.05297*.

[12] Y. Zhang, L. Trinh, D. Cao, Z. Cui, and Y. Liu, "Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model," 2023, *arXiv:2304.07633*.

[13] M. Mu, S. Das Bhattacharjee, and J. Yuan, "Self-supervised distilled learning for multi-modal misinformation identification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 2819–2828.

[14] D.-T. Dang-Nguyen et al., "Overview of the grand challenge on detecting cheapfakes at ACM ICMR 2024," in *Proc. Int. Conf. Multimedia Retrieval*, 2024, pp. 1275–1281.

[15] Y. Gu, M. Zhang, I. Castro, S. Wu, and G. Tyson, "Learning domain-invariant features for out-of-context news detection," 2024, *arXiv:2406.07430*.

[16] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 178–206, Feb. 2022.

[17] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Volume 2: Short Papers)*, 2017, pp. 422–426.

[18] I. Augenstein et al., "MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4685–4697.

[19] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: a large-scale dataset for fact extraction and verification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Volume 1 (Long Papers), 2018, pp. 809–819.

[20] R. Aly et al., "Feverous: Fact extraction and verification over unstructured and structured information," 2021, *arXiv:2106.05707*.

[21] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, and L. Huang, "End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2023, pp. 2733–2743.

[22] S. Suryavardan et al., "Factify 2: A multimodal fake news and satire news dataset," 2023, *arXiv:2304.03897*.

[23] S. Abdelnabi, R. Hasan, and M. Fritz, "Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14940–14949.

[24] F. Zhang, J. Liu, Q. Zhang, E. Sun, J. Xie, and Z.-J. Zha, "ECENet: Explainable and context-enhanced network for muti-modal fact verification," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 1231–1240.

[25] X. Yuan, J. Guo, W. Qiu, Z. Huang, and S. Li, "Support or refute: Analyzing the stance of evidence to detect out-of-context mis-and disinformation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2023, pp. 4268–4280.

[26] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2021, pp. 8748–8763.

[27] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9694–9705.

[28] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.

[29] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, "Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias," *Int. J. Multimedia Inf. Retrieval*, vol. 13, no. 1, p. 4, 2024.

[30] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE Access*, vol. 9, pp. 156151–156170, 2021.

[31] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[32] F. Alam et al., "A survey on multimodal disinformation detection," in *Proc. 29th Int. Conf. Comput. Linguistics*. Int. Committee Comput. Linguistics, 2022, pp. 6625–6643.

[33] M. Glockner, Y. Hou, and I. Gurevych, "Missing counter-evidence renders NLP fact-checking unrealistic for misinformation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2022, pp. 5916–5936.

[34] A. Jaiswal, E. Sabir, W. AbdAlmageed, and P. Natarajan, "Multimedia semantic integrity assessment using joint embedding of images and text," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1465–1471.

[35] E. Sabir, W. AbdAlmageed, Y. Wu, and P. Natarajan, "Deep multimodal image-repurposing detection," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1337–1345.

[36] Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 849–857.

[37] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, 2019, pp. 2915–2921.

[38] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *Proc. IEEE fifth Int. Conf. Multimedia Big Data (BigMM)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 39–47.

[39] C. Yu, Y. Ma, L. An, and G. Li, "BCMF: A bidirectional cross-modal fusion model for fake news detection," *Inf. Process. Manage.*, vol. 59, no. 5, 2022, Art. no. 103063.

[40] K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 6149–6157.

[41] S. Suryavardan et al., "Findings of factify 2: multimodal fake news detection," 2023, *arXiv:2307.10475*.

[42] Z. Chrysidis, S.-I. Papadopoulos, S. Papadopoulos, and P. Petrantonakis, "Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking," in *Proc. 3rd ACM Int. Workshop Multimedia AI Against Disinformation*, 2024, pp. 73–81.

[43] C. Boididou et al., "Verifying information with multimedia content on Twitter: a comparative study of automated approaches," *Multimedia tools Appl.*, vol. 77, pp. 15545–15571, 2018.

[44] C. Koutlis, M. Schinas, and S. Papadopoulos, "MemeFier: Dual-stage modality fusion for image meme classification," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2023, pp. 586–591.
[45] G. K. Kumar and K. Nandakumar, "Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features," in *Proc. 2nd Workshop NLP Positive Impact (NLP4PI)*, 2022, pp. 171–183.
[46] P. W. Koh et al., "WILDs: A benchmark of in-the-wild distribution shifts," in *Int. Conf. Mach. Learn. (PMLR)*, 2021, pp. 5637–5664.
[47] C. Longoni, A. Fradkin, L. Cian, and G. Pennycook, "News from generative artificial intelligence is believed less," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, 2022, pp. 97–106.
[48] S. Mohtaj, A. Nizamoglu, P. Sahitaj, V. Schmitt, C. Jakob, and S. Möller, "NewsPolyML: Multi-lingual european news fake assessment dataset," in *Proc. 3rd ACM Int. Workshop Multimedia AI Against Disinformation*, 2024, pp. 82–90.
[49] Q. Xu et al., "M3A: A multimodal misinformation dataset for media authenticity analysis," *Comput. Vis. Image Understanding*, vol. 249, 2024, Art. no. 104205.

Stefanos-Iordanis Papadopoulos received the bachelor's degree in computer science from the Department of Applied Informatics, the University of Macedonia, Thessaloniki, Greece, in 2018 and the master's degree in 2021 in data and web science from the School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2004, where he is currently working toward the Ph.D. degree in multimodal deep learning with the Department of Electrical and Computer Engineering.

Since 2020, he has been a Research Associate with the Information Technologies Institute of the Centre for Research and Technology Hellas, Thessaloniki, Greece. His research interests include multimodal deep learning, focusing on multimedia verification and retrieval, as well as automated multimodal misinformation detection and fact-checking.

Dr. Papadopoulos has been involved in three European Commission funded research projects and co-authored more than 10 peer-reviewed papers in international scientific journals and conferences.

Christos Koutlis received the bachelor's degree in mathematics and the master's degree in statistics and modeling from the Department of Mathematics, Aristotle University of Thessaloniki, in 2012 and 2014, respectively. He received the Ph.D. degree in multivariate time-series analysis from the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, in 2017.

Since 2018, he is a Postdoctoral Researcher with the Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece. His research interests include deep learning based visual and multimodal analysis with focus on multimedia verification and visual bias.

Dr. Koutlis has co-authored over 35 papers published in international scientific journals and conferences (including book chapters) and has been involved in nine European Commission funded research projects.

Symeon (Akis) Papadopoulos received the diploma degree in electrical and computer engineering from Aristotle University of Thessaloniki, in 2004, the Professional Doctorate degree in engineering from the Technical University of Eindhoven, Eindhoven, Netherlands, in 2006, the MBA degree in business from Blekinge Institute of Technology, Karlskrona, Sweden, in 2009, and the Ph.D. degree in computer science from Aristotle University of Thessaloniki, in 2012.

Currently, he is a Principal Researcher with the Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece. He is leading the Media Analysis, Verification and Retrieval Group (MeVer), and is the Co-Founder of Infalia Pvt. Company, a spin-out of CERTH-ITI. His research interests include the intersection of multimedia understanding, social network analysis, information retrieval, big data management, and artificial intelligence.

Dr. Papadopoulos has co-authored more than 50 articles in refereed journals, 15 book chapters and 150 papers in international conferences, three patents, and has edited two books. He has been participating and coordinating a number of relevant EC FP7, H2020, and Horizon Europe projects in the areas of media convergence, social media, and artificial intelligence.

Panagiotis C. Petrantonakis (Senior Member, IEEE) received the diploma degree in electrical and computer engineering and the Ph.D. degree in signal processing and machine learning from Aristotle University of Thessaloniki, Greece, in 2007 and 2011, respectively.

From 2012 to 2016, he was a Postdoctoral Researcher with the Institute of Molecular Biology and Biotechnology of the Foundation for Research and Technology - Hellas (FORTH) and from 2017 to 2022, he was a Postdoctoral Researcher with the Information Technologies Institute of the Centre for Research and Technology - Hellas (CERTH). Currently, he is an Assistant Professor with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki.

Dr. Petrantonakis has published more than 50 papers in peer reviewed journals, book chapters and conferences in the fields of signal processing, machine learning and large scale data analysis.