International Journal of Multimedia Information Retrieval (2024) 13:4
https://doi.org/10.1007/s13735-023-00312-6

REGULAR PAPER

VERITE: a Robust benchmark for multimodal misinformation detection accounting for unimodal bias

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

Received: 21 July 2023 / Revised: 18 October 2023 / Accepted: 21 November 2023 / Published online: 8 January 2024

Abstract
Multimedia content has become ubiquitous on social media platforms, leading to the rise of multimodal misinformation (MM) and the urgent need for effective strategies to detect and prevent its spread. In recent years, the challenge of multimodal misinformation detection (MMD) has garnered significant attention by researchers and has mainly involved the creation of annotated, weakly annotated, or synthetically generated training datasets, along with the development of various deep learning MMD models. However, the problem of unimodal bias has been overlooked, where specific patterns and biases in MMD benchmarks can result in biased or unimodal models outperforming their multimodal counterparts on an inherently multimodal task, making it difficult to assess progress. In this study, we systematically investigate and identify the presence of unimodal bias in widely used MMD benchmarks, namely VMU-Twitter and COSMOS. To address this issue, we introduce the "VERification of Image-TExt pairs" (VERITE) benchmark for MMD which incorporates real-world data, excludes "asymmetric multimodal misinformation" and utilizes "modality balancing". We conduct an extensive comparative study with a transformer-based architecture that shows the ability of VERITE to effectively address unimodal bias, rendering it a robust evaluation framework for MMD. Furthermore, we introduce a new method—termed Crossmodal HArd Synthetic MisAlignment (CHASMA)—for generating realistic synthetic training data that preserve crossmodal relations between legitimate images and false human-written captions. By leveraging CHASMA in the training process, we observe consistent and notable improvements in predictive performance on VERITE; with a 9.2% increase in accuracy. We release our code at: https://github.com/stevejpapad/image-text-verification

Keywords Multimodal learning · Deep learning · Misinformation detection · Unimodal bias · Benchmark

1 Introduction

The proliferation of misinformation poses a significant societal challenge with potential negative impacts on democratic processes [7], social cohesion [13], public health [43], political and religious persecution [14] among others. The widespread usage of digital media platforms in recent years has only exacerbated the problem [39]. In the context of social media platforms, multimedia content has been shown to often be more attention-grabbing and widely disseminated than plain text [28], while the presence of an image can significantly enhance the persuasiveness of a false statement [37]. Against this backdrop, while the work of fact-checkers becomes increasingly important it also becomes increasingly more difficult, considering the scale of content produced and shared daily on social media. In response, researchers have been investigating a range of AI-based methods for detecting misinformation, e.g. detecting inaccurate claims with the use of natural language processing [32], detecting synthetic images, such as DeepFakes, with the use of deep learning [42] or multimodal misinformation with the use of multimodal deep learning [17].

✉ Stefanos-Iordanis Papadopoulos
stefpapad@iti.gr

Christos Koutlis
ckoutlis@iti.gr

Symeon Papadopoulos
papadop@iti.gr

Panagiotis C. Petrantonakis
ppetrant@ece.auth.gr

[1] Information Technology Institute, Centre for Research & Technology, Hellas, Thessaloniki, Greece

[2] Department of Electrical & Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece

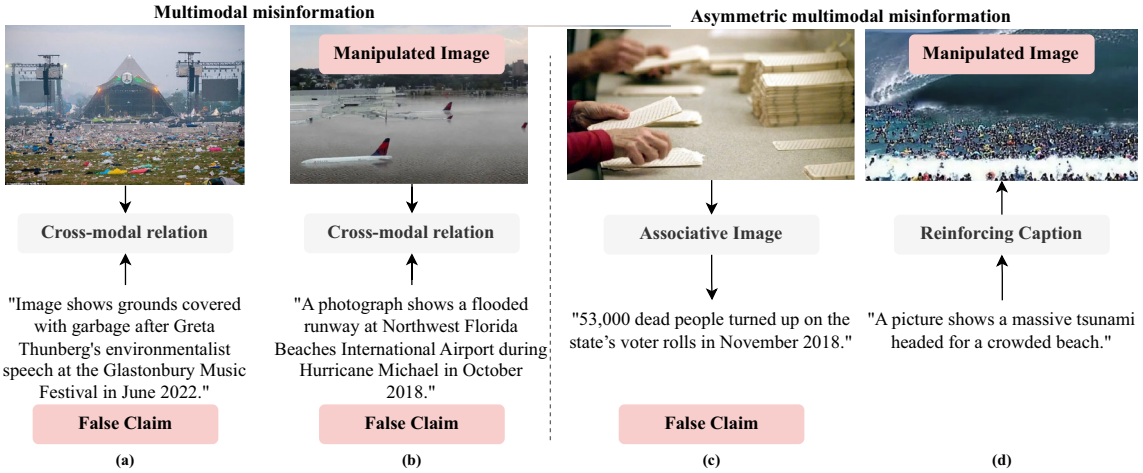

Fig. 1 Examples of multimodal misinformation (taken from VERITE) and asymmetric multimodal misinformation (taken from COSMOS benchmark)

Multimodal misinformation (MM) typically refers to false or misleading information that is spread using multiple modes of communication, such as text, images, audio and video [3]. Here, we focus on image-caption pairs that collaboratively contribute to the dissemination of misinformation. For instance, in Fig. 1a, an image depicts the grounds of a musical festival covered in garbage, accompanied by the false claim that it was taken in June 2022 "after Greta Thunberg's environmentalist speech", while the image was actually taken in 2015.[1]

Previous studies on automated multimodal misinformation detection (MMD) have predominantly explored three approaches in terms of training datasets: annotated [9, 55], weakly annotated [20, 35, 38] and synthetically generated datasets [19, 31, 44]. These distinct routes facilitated the development and evaluation of multimodal models designed to detect and combat misinformation effectively [22, 45, 50, 52]. However, previous studies have overlooked the investigation of unimodal bias. Training datasets exhibiting certain patterns and biases (asymmetries and imbalances) towards one modality can lead to biased models or unimodal methods capable of outperforming their multimodal counterparts in a purportedly multimodal task. If these patterns persist within the evaluation benchmarks, they can obscure the impact of unimodal bias, hindering our ability to effectively assess progress in the field of MMD. In our investigation, we uncover that the widely used VMU-Twitter dataset [9] exhibits an image-side unimodal bias, while the COSMOS evaluation benchmark [4] exhibits a text-side unimodal bias, raising questions about their reliability as evaluation benchmarks for MMD.

Against this backdrop, the primary aim of this study is to create a robust evaluation framework that accounts for unimodal bias. To this end, we have create the "VERification of Image-TExt pairs" (VERITE) evaluation benchmark which accounts for unimodal bias by (1) consisting of real-world data, (2) excluding "asymmetric multimodal misinformation" (Asymmetric-MM) and (3) employing "modality balancing". We introduce the term Asymmetric-MM–and contrast it with MM—to highlight cases where one dominant modality is responsible for propagating misinformation, while other modalities have little or no influence. An example of Asymmetric-MM can be seen in Fig. 1c where a claim pertains to "deceased people turning up to vote", and an image merely thematically related to the claim is added primarily for cosmetic enhancement. Focusing on the dominant modality, a robust text-only (or, in other scenarios, image-only) detector would suffice for detecting misinformation, rendering the other modality inconsequential in the detection process. We hypothesize that this asymmetry can exacerbate unimodal bias. Furthermore, we introduce the concept of "modality balancing" which ensures that all images and captions are presented twice during evaluation, once in their truthful pair and once in their misleading pair, thus compelling a model to consider both modalities and their relation when discerning between truth and misinformation. We conduct a comprehensive comparative analysis where we train a transformer-based architecture using different datasets, including VMU-Twitter, Fakeddit, and various synthetically generated datasets. Our empirical results demonstrate that VERITE effectively mitigates and prevents the occurrence of unimodal bias.

Our second contribution is the introduction of "Cross-modal HArd Synthetic MisAlignment" (*CHASMA*), a new method for generating synthetic training datasets that aims to maintain crossmodal relation between legitimate images and misleading human-written texts to create plausible mis-

[1] https://www.snopes.com/fact-check/glastonbury-greta-thunberg.

leading pairs. More specifically, *CHASMA* utilizes a large pre-trained crossmodal alignment model (CLIP [41]) to pair legitimate images (from VisualNews [30]) with contextually relevant but misleading texts (from Fakeddit). *CHASMA* maintains the sophisticated linguistic patterns (e.g. exaggeration, irony, emotions) that are often found in human-written texts, unlike methods that rely on Named Entity Inconsistencies (NEI) for generating MM [44]. The inclusion of *CHASMA* in the training process consistently enhances the predictive performance on the VERITE benchmark, particularly evident in aggregated datasets, resulting in a notable 9.2% increase in accuracy.

The main contributions of our work can be summarized as follows:

- Systematically investigate the issue of unimodal bias within widely used evaluation MMD benchmarks (VMU-Twitter and COSMOS).
- Create the VERITE benchmark, which effectively mitigates the problem of unimodal bias and provides a more robust and reliable evaluation framework for MMD.
- Develop CHASMA, a novel approach for creating synthetic training data for MMDthat consistently leads to improved detection accuracy on the VERITE benchmark.

2 Related work

The automated detection of misinformation is a challenging task that has garnered increasing attention from researchers in recent years. A range of methods is being explored to identify misinformation in text [32] and images [42]. Consequently, multiple datasets have been created for fake news detection [48] and manipulated images [18]. These challenges involve unimodal settings. However, there is a need for MMD models that can handle cases where the combination of an image and its caption lead to misinformation. Given the complexity of this task, large training datasets are required to train robust MMD models. In this section, we explore the available research on existing datasets, including both annotated and synthetically generated as well as available evaluation benchmarks for MMD.

2.1 Annotated multimodal misinformation datasets

The "image verification corpus", often referred as the "Twitter" dataset, ("VMU-Twitter" from now on) was used in the MediaEval2016 Verifying Multimedia Use (VMU) challenge [9] and comprises 16,440 tweets regarding 410 images for training and 1,090 tweets regarding 104 images for evaluation. Since the images are accompanied by tweets, the dataset has been widely used for MMD [22, 45, 50, 52]. In addition, the Fauxtography dataset comprises manually fact-checked image-caption pairs sourced from Snopes[2] and Reuters,[3] with a total of 1,233 pairs, of which 592 are classified as truthful and 641 as misleading [55]. However, their very limited size raises doubts about the effectiveness and generalizability of deep neural networks trained on these datasets.

To address the challenges of collecting and annotating large-scale datasets, researchers have also explored weakly annotated datasets. The MuMiN dataset, for instance, consists of 21 million tweets on Twitter, linked to 13,000 fact-checked claims, with a total of 6,573 images [38]. While this dataset provides rich social information such as user information, articles, and hashtags, its limited number of images may also be insufficient for MMD. The NewsBag is another large-scale multimodal dataset that was created by scraping the Wall Street Journal and Real News for truthful pairs and The Onion[4] and The Poke[5] for misleading pairs [20]. However, the latter sites publish humorous and satirical articles which may not reflect real-world misinformation [24].

Fakeddit is a large weakly labelled dataset consisting of 1,063,106 instances collected from various subreddits[6] and grouped into two, three, or six classes based on their content [35]. The instances are classified as either Truthful or Misleading and then separated into six classes, including true, satire, misleading content, manipulated content, false connection, or impostor content. Of the total instances, 680,798 have both an image and a caption, with 413,197 of them being Misleading and 267,601 being Truthful. Despite being weakly labelled, Fakeddit provides a large-scale resource for training machine learning models to detect misleading multimodal content.

2.2 Synthetic multimodal misinformation datasets

Due to the need for large-scale datasets, the labour-intensive nature of manual annotation and the potential for weak labelling to introduce noise, researchers have also been exploring the use of synthetically generated training data for MMD. These methods can be categorized into two groups based on the type of misinformation they generate, namely OOC pairs or NEI.

OOC-based datasets can be created through random sampling techniques, such as in the case of the MAIM [19] and COSMOS [4] datasets. However, these methods tend to produce easily detectable non-realistic pairs, making them unsuitable for training effective misinformation detection models [40]. An alternative approach is to use feature-based

[2] https://www.snopes.com/fact-check/category/photos.

[3] https://www.reuters.com/fact-check.

[4] https://www.theonion.com.

[5] https://www.thepoke.co.uk.

[6] https://www.reddit.com.

sampling to retrieve more realistic pairs that more realistically resemble multimodal misinformation. The NewsCLIPings dataset [31] was created using scene-learning, person matching and CLIP in order to retrieve images from within the VisualNews dataset in order to create OOC samples. Similarly, the Twitter-COMMs dataset was created via CLIP-based sampling on Twitter data related to climate, COVID, and military vehicles [8].

On the other hand, NEI-based methods rely on substituting named entities in the caption—such as people, locations, and dates—with other entities of the same type, resulting in misleading inconsistencies between the image and caption. Since random retrieval and replacement of entities may be easily detectable [40], several methods have been proposed to retrieve relevant entities based on cluster-based retrieval for MEIR [44], rule-based retrieval for TamperedNews [34], and CLIP-based retrieval for CLIP-NESt [40]. Finally, aggregating synthetically generated datasets—combining both OOC and NEI—has been shown to further improve performance [40].

2.3 Unimodal bias and evaluation benchmarks

Unimodal bias has mainly been observed and investigated in the domain of visual question answering (VQA), wherein biased models rely on surface-level statistical patterns within one modality (usually the textual modality), while disregarding the information present in the other modality (usually the visual modality) [15]. Evaluation benchmarks have been devised to enhance fairness and robustness of evaluating of VQA models [2] and various methods have been proposed for counteracting unimodal bias during training [10]. However, comparable efforts in addressing unimodal bias have not been explored within the context of MMD.

Currently, there is no widely accepted benchmark for evaluating MMD models. Most studies assess their approaches on a split of their weakly annotated [35, 38] or their synthetically generated datasets [19, 31, 34, 44], which may not provide a realistic estimate of how these methods will perform when confronted with real-world misinformation. The COSMOS benchmark is one of the few works that collect an evaluation set consisting of real-world multimodal misinformation and make it publicly available [4]. It consists of 1,700 pairs and is balanced between truthful and misleading pairs—collected from credible news sources and Snopes.com respectively—and has been used in two challenges for "CheapFakes detection" [5, 6]. Nevertheless, in [40], it was found that text-only methods, especially NEI-based ones, can outperform their multimodal counterpart on COSMOS, raising questions about its reliability as an MMD benchmark. Another widely used dataset for MMD is the VMU-Twitter dataset [9], despite consisting mainly of manipulated and digitally created images. In this paper, we systematically investigate the factors behind unimodal bias in MMD and create a new evaluation benchmark that accounts for it.

3 Methodological framework

3.1 Problem definition

In this study, we focus on the challenge of multimodal misinformation detection (MMD) and specifically on image-caption pairs that collaboratively contribute to the propagation of misinformation. Typically, MMD can be defined as follows: Given a dataset $(x_i, y_i)_{i=1}^{N}$, where $x_i = (I_i, C_i)$ represents an image-caption pair and $y_i \in \{0, 1\}$ denotes the ground truth label indicating the presence or absence of misinformation, the objective is to learn a mapping function $f : x \rightarrow y$ that accurately predicts the presence of misinformation in a given image-caption pair. However, instead of addressing MMD as a binary classification problem ([4, 19, 31, 34, 40, 44, 55]), we introduce a new taxonomy that includes three classes:

1. Truthful (True): An image-caption pair (I_i^t, C_i^t) is considered True when the origin, content, and context of an image are accurately described in the accompanying caption.
2. Out-Of-Context (OOC) image-text pairs: It involves a deceptive combination of a truthful caption C_i^t and an out-of-context image I_i^x or a legitimate image I_i^t with an out-of-context caption C_i^x; with "^{x}" denoting the different context but otherwise truthful information.
3. MisCaptioned images (MC): It involves an image I_i^t being paired with a misleading caption C_i^f that misrepresents the origin, content, and/or meaning of the image; with "^{f}" denoting falsehood or manipulation.

We consider the structural differences between OOC and MC to warrant separate classification since MC cases predominantly involve the introduction of falsehoods within the textual modality that are linked to the image, whereas OOC scenarios involve the juxtaposition of otherwise truthful text with a legitimate yet decontextualized image, resulting in the propagation of misinformation.

Furthermore, we investigate the problem of unimodal bias in the context of MMD, the phenomenon of unimodal models or models biased towards one modality outperforming their unbiased multimodal counterparts on an inherently multimodal task. Unimodal bias can emerge during the training process as a consequence of certain patterns and biases, wherein models tend to emphasize superficial statistical correlations within a single modality. If these patterns persist

within the evaluation benchmarks, they have the potential to obscure the presence of unimodal biases within the results. We hypothesize that one such problematic pattern is "asymmetric multimodal misinformation" (Asymmetric-MM)—which we contrast against MM—where false claims are accompanied by a loosely connected image (associative imagery) or manipulated images are accompanied by captions that simply reinforce the misleading content of the image (reinforcing captions). Examples are provided in Fig. 1c and d. Both scenarios create an asymmetry between the two modalities, rendering one modality as the dominant source of misinformation, while the second modality has little or no influence. It is important to note that instances of MC images (including NEI) may exhibit a certain degree of "asymmetry" in that misinformation is primarily propagated through the textual modality. Nevertheless, we do not consider them to be Asymmetric-MM because the text in MC pairs remains connected to and misrepresents some aspect of the image, such as depicted entities or events.

Previous studies did not make a distinction between MM and Asymmetric-MM while collecting or annotating their datasets. Given 200 random samples from COSMOS and following the classification taxonomy of Snopes[7], we found that 48% of COSMOS pairs are "false claims" (41% associative imagery and 7% reinforcing captions), while 52% were classified as "miscaptioned", which we consider to be MM because it implies a relationship between the two modalities. After de-duplicating the images of the COSMOS benchmark, the rates were 41% miscaptioned, 35% associative imagery, 4% reinforcing captions and 20% duplicates. On Fakeddit—given 300 random samples—roughly 45% of pairs were considered Asymmetric-MM, with 41% being manipulated images and 4% with associative imagery. Moreover, we consider that roughly 14% of Fakeddit's samples are MM since the remaining 40% were mostly funny memes, visual jokes, pareidolia imagery and other content that is not generally considered to be misinformation.[8]

3.2 Creating the VERITE evaluation benchmark

Due to the lack of a robust evaluation benchmark for MMD that accounts for unimodal bias, we introduce the "VERification of Image-TExt pairs" (VERITE) benchmark. VERITE comprises three classes: True, OOC, and MC pairs. The data collection process is illustrated in Fig. 2 and involves the following steps:

1. Define inclusion criteria
   - Consider fact-checked articles from Snopes and Reuters that are classified as "MisCaptioned" (MC).
   - Exclude articles classified as "false claim", "legend", "satire', "scam", "misattributed" and other categories that do not adhere to our definition of MM.
   - Exclude articles regarding video footage or animated content and keep image-related cases, unless a screenshot of the video is provided that clearly captures the content and claim of the caption.
   - Include manipulated images (digital art, AI-generated imagery, etc.) only if they are not created with intention to misinform, and their initial origin, content, context, or meaning has been misrepresented within the claim.
2. Select images and captions
   - Review the article and collect the misleading claim C_i^f.
   - Collect the image I_i^t that is related to C_i^f.
   - Extract the truthful claim C_i^t for I_i^t from the article.
   - Examine if claim C_i^f is linked to I_i^t and misrepresents some aspect of it (e.g. origin, content, context, depicted entities etc). If not, exclude for being Asymmetric-MM.
3. Refine captions and images
   - Remove "giveaway" words such as "supposedly", "allegedly", "however" or phrases like "this is not the case", that negate the false claim. Such words and phrases, if learned during the training process, could be used as "shortcuts" by MMD models.
   - Rephrase C_i^f to mimic the syntactic and grammatical structure of C_i^t in order to avoid potential linguistic biases.
   - Rephrase both C_i^t and C_i^f to follow the format: "An image shows.." or "Photograph showing.." in order to create a direct link between the two modalities.
   - Examine both C_i^t and C_i^f for spelling and grammatical errors using "Google Docs spelling and grammar check".
   - Verify that the images are of reasonable quality and do not have any watermarks. If needed, use reverse image search to find the exact same image in better quality.
4. OOC Image retrieval
   - Extract relevant keywords, or their synonyms, in C_i^t to create a query Q.
   - Use Google image search to retrieve one OOC image I_i^x based on Q.

[7] "Fact Checks Rating" in https://www.snopes.com/sitemap.

[8] The assessment of the COSMOS benchmark follows the taxonomy of Snopes, which is based on the judgement of professional fact-checkers, while the assessment of Fakeddit was conducted by the authors and thus should be interpreted as a rough estimate and not definitive.

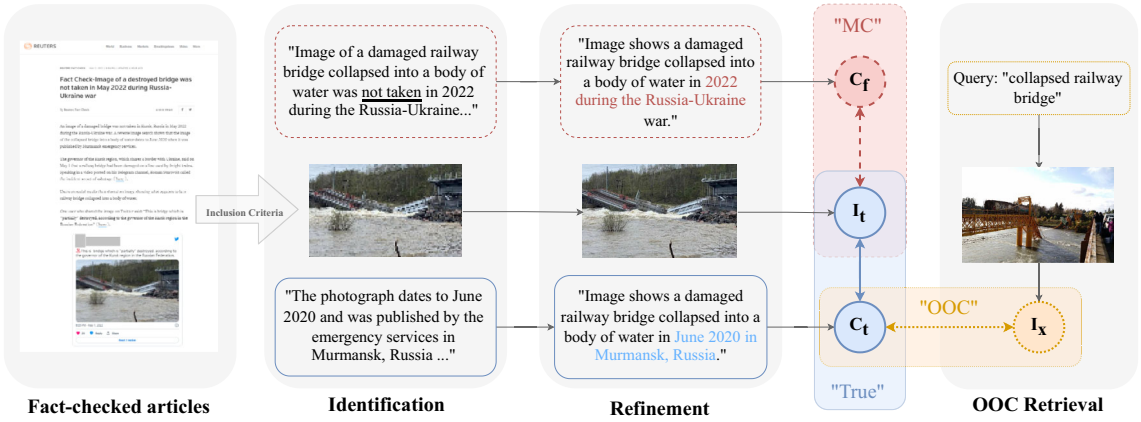

Fig. 2 Data collection, filtering and refinement process for creating VERITE

- Ensure that C_i^t and I_i^x share a discernible and meaningful connection (identical or similar origin, content, context or depicted entities) and their alignment is deceptive.

To illustrate the aforementioned process in practice, let us consider the example shown in Fig. 2. Starting with a fact-checked article,[9] we collect I_i^t showing a damaged railway that has collapsed into a body of water and C_i^f falsely claiming that the event occurred in "2022 during the Russia-Ukraine war". We also collect the truthful C_i^t which is provided by professional fact-checkers. C_i^t clarifies that the event took place in "June 2020 in Murmanask, Russian" and thus is unrelated to the 2022 Russia–Ukraine war. Afterwards, we extract keywords from C_i^t and use Q ="collapsed railway bridge" as the query and retrieve I_i^x from Google Images. Similar to I_i^t, I_i^x also depicts a collapsed railway bridge but it was captured in Chile, not Russia, thus misaligning the "location" entity.

We collected 260 articles from Snopes and 78 from Reuters that met our criteria, which translates to 338 (I_i^t, C_i^t), 338 (I_i^t, C_i^f) and 324 (I_i^x, C_i^t) pairs for True, MC, and OOC, respectively. The collected Snopes articles date as far back as January 2001 up to January 2023, whereas Reuters—only allowing searches up to two years in the past—date from January 2021 to January 2023. The collected data cover a wide and diverse array of topics and cases including world news (29.04%), politics (27.94%), culture and arts (8.82%), entertainment (7.72%), sports (3.67%), the environment (3.66%), religion (2.94%), travel (2.57%), business (2.20%), science and technology (2.19%), health and wellness (1.46%) and others.[10]

We introduce the term "modality balancing" to denote that I_i^t and C_i^t are included twice in the dataset: once with the truthful label and once within the misleading label, as seen in Fig. 2. More specifically, each image is present once in its truthful pair and once in the MC pair, while each caption is present once in its truthful pair and once in the OOC pair. This approach ensures that the model will have to focus on both modalities to consistently discern between factual and misleading I-C pairs.

3.3 Crossmodal hard synthetic misalignment

Previous studies on synthetic training data for MMD have primarily relied on OOC pairs or NEI. These methods create formulaic manipulations, either by re-sampling existing pairs or substituting named entities, and therefore lack the imaginative or expressive characteristic of human produced misinformation such as emotions or irony. Conversely, large weakly annotated datasets may contain noisy labels and high rates of Asymmetric-MM.

To address these issues, we propose a new method for generating MM termed Crossmodal HArd Synthetic MisAlignment (CHASMA). Given a truthful (I_i^t, C_i^t) pair and their $V_{I_i^t}$, $T_{C_i^t}$ visual and textual embeddings extracted from CLIP, retrieve the most plausible misleading caption C_j^f from a collection of misleading captions $\mathcal{C}_\mathcal{F}$ with $T_{\mathcal{C}_\mathcal{F}}$ textual embeddings, in order to produce a miscaptioned (I_i^t, C_j^f)

[9] https://www.reuters.com/article/factcheck-destroyed-bridge-idUSL2N2WU1CM.

[10] To extract and estimate the frequency distribution of news categories, we used https://huggingface.co/Yueh-Huan/news-category-classification-distilbert.

pair with:

$$\underset{C_j^f \in \mathcal{C}_\mathcal{F}}{\text{argmax}} \begin{cases} sim(T_{C_i^t}, T_{C_j^f}), & p \leq 0.5 \\ sim(V_{I_i^t}, T_{C_j^f}), & p > 0.5 \end{cases} \quad (1)$$

where $p \in [0, 1]$ is a uniformly sampled number that determines calculating the cosine similarity (sim) between text-to-text or image-to-text pairs.

We apply crossmodal hard synthetic misalignment between VisualNews [30] dataset—consisting of 1,259,732 (I_i^t, C_i^t) pairs—and the Fakeddit dataset (I_j^f, C_j^f) [35]. Out of the 400K misleading captions in $\mathcal{C}_\mathcal{F}$ in the Fakeddit dataset, the misalignment process only retains 145,891. The resulting generated dataset, termed *CHASMA*, is balanced between 1.2M (I_i^t, C_i^t) truthful and 1.2M (I_i^t, C_j^f) miscaptioned pairs. Since C_j^f from Fakeddit may have been aligned with more than one image from VisualNews, we also create *CHASMA-D* by removing duplicate instances of C_j^f. We balance the classes of *CHASMA-D* through random down-sampling. The resulting dataset consists of 145,891 (I_i^t, C_i^t) and an equal number of (I_i^t, C_j^f).

We randomly sample 100 instances from the generated data and determined that approximately 73% of generated (I_i^t, C_j^f) can be considered MM, while 12% are Asymmetric-MM. Moreover, 6% of the pairs in the dataset are accidentally correct pairs, for instance, an image of firefighters near a fire being paired with the caption "Firemen battling a blaze". Finally, 9% of pairs are unclear, containing click-bait captions as "You'll never guess how far new home prices have dropped" which are paired with a weakly relevant image and cannot necessarily be considered misinformation. Naturally, the proposed method is not perfect, with approximately 27% of its samples not aligning with our definition of MM. Nevertheless, it provides a significant improvement over the original Fakeddit dataset where roughly 45% are Asymmetric-MM and only 14% are MM.

As seen in the examples of Fig. 3 (bottom), misleading captions C_j^f from Fakeddit can contain humour, irony and be more imaginative than named entity substitutions. However, their connections with the images I_j^f are often Asymmetric-MM or can be easy to detect (e.g. an illustrated image being humorously paired with a real demonstration). Conversely, *CHASMA* maintains the "desired" aspects of C_j^f (e.g. sarcasm, emotions, etc.) but pairs them with more relevant imagery, thus creating "hard" samples and by extension more robust training data. For example, consider the case shown in Fig. 3, where an illustrated image is humorously paired with a caption about a demonstration and is subsequently "misaligned" with an image of a real protest, thus creating a more realistic misleading pair.

In contrast to NEI-based methods, our generated samples consists of human-written misinformation rather than simple named entity manipulations. Finally, unlike NewsCLIPings, CHASMA utilizes CLIP-based retrieval to generate MC rather than OOC pairs and employs both intra-modal and crossmodal similarity to create synthetic samples.

3.4 Detection model

In our experiments, we encode all image-caption pairs (I, C) using the pre-trained CLIP ViT-L/14 [41] both as the image encoder $E_I(\cdot)$ and the textual encoder $E_C(\cdot)$ that produce

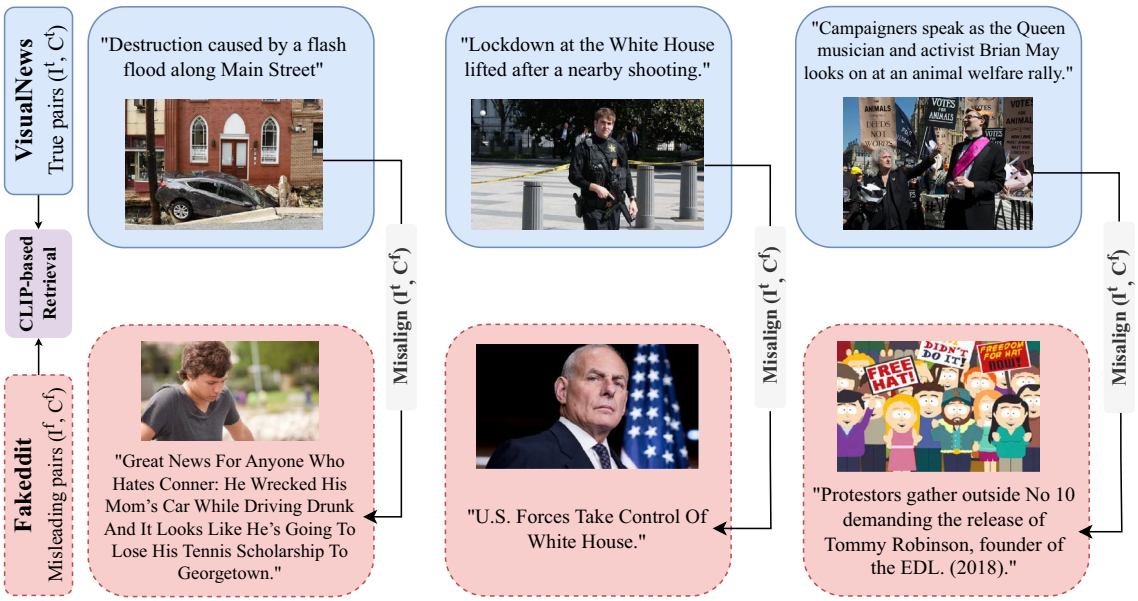

Fig. 3 Training samples from CHASMA when applied across the VisualNews and Fakeddit datasets

the corresponding vector representations $V_I \in R^{m\times 1}$ and $T_C \in R^{m\times 1}$, respectively, where $m = 768$ the encoder's embedding dimension. CLIP is an open and widely used model for multimodal feature extraction in numerous multimedia analysis and retrieval tasks [16, 27, 29] including multimedia verification and has yielded promising results [31, 40, 47, 53, 54].

We concatenate the extracted features across the first or "token" axis as $[V_I, T_C] \in R^{m\times 2}$. ("batch dimension" omitted for clarity). As the "detector" $D(\cdot)$, we use the transformer encoder [49] but exclude positional encoding and use average pooling instead of a CLS token. $D(\cdot)$ comprises L layers of h attention heads and a feed-forward network of f dimension and outputs y:

$$y = \mathbf{W}1 \cdot \mathrm{GELU}(\mathbf{W}0 \cdot \mathrm{LN}(\mathrm{D}([V_I, T_C]))) \quad (2)$$

where LN stands for layer normalization, $\mathbf{W}_0 \in \mathbb{R}^{m\times 2}$ is a GELU activated fully connected layer and $\mathbf{W}_1 \in \mathbb{R}^{n\times \frac{m}{2}}$ is the final classification layer with $n = 1$ for binary and $n = 3$ for multiclass tasks (learnable bias terms are considered but omitted here for clarity). The network is optimized based on the categorical cross-entropy or the binary cross-entropy loss function for multiclass or binary tasks, respectively.

For unimodal experiments, we only pass V_I or T_C through $D(\cdot)$ and define $\mathbf{W}_0 \in \mathbb{R}^{m\times 1}$. In these cases, $D(\cdot)$ only receives a single input token. Therefore, its attention scores are uniformly assigned a value of 1, resulting in the absence of distinct attention weights. We denote this "Transformer" detector as $D^-(\cdot)$; $D(\cdot)$ minus multi-head self-attention, since the latter has no contributing role.

Moreover, in order to investigate the role that multi-head self-attention plays in unimodal bias, we conduct additional experiments using the variant $D^-(\cdot)$ where the two modalities are concatenated along the second or "dimensional" axis, resulting $[V_I; T_C] \in \mathbb{R}^{2m\times 1}$.

4 Experimental setup

4.1 Training datasets and competing methods

First, we train $D(\cdot)$ on the **VMU-Twitter** (MediaEval 2016[11]) dataset and compare it against numerous MMD models, namely: event adversarial neural network (*EANN*) using VGG-19 and TextCNN [50], multimodal variational autoencoder (*MVAE*) using VGG-19 and Bi-LSTMs [22], *SpotFake* using VGG-19 and BERT [45], bidirectional crossmodal fusion (*BCMF*) network using DeiT and BERT [52] and a transformer-based architecture employing faster-RCNN and BERT to capture intra-modal relations and a multiplicative multimodal method to capture inter-modal relations (*Intra+Inter*) [46].

Table 1 Number of samples per class in each training and testing dataset. "*" denotes datasets whose "false" pairs exhibit more similarities to, but may not entirely align with, our definition of miscaptioned (MC) images. Validation sets are used but omitted here

	True	OOC	MC
Training Dataset			
VisualNews	1,007,744	–	–
RSt	1,007,744	1,007,744	–
NC-t2t	258,036	258,036	–
CSt	1,007,744	1,007,744	–
MEIR	82,156	–	57,940
CLIP-NESt	1,007,744	–	847,693
R-NESt	1,007,744	–	924,586
Fakeddit*	267,601	–	413,197
CHASMA	1,007,744	–	1,007,744
CHASMA-D	145,891	–	145,891
VMU-Twitter*	7292	–	9148
Testing Dataset			
VMU-Twitter*	467	–	623
COSMOS*	850	–	850
VERITE	338	324	338

Afterwards, we compare $D(\cdot)$ when trained on the original **Fakeddit** [35], our **CHASMA** and **CHASMA-D** datasets as well as numerous synthetically generated datasets, including OOC: NewsCLIPings text-text (**NC-t2t**) [31], random sampling by topic (**RSt**) [40] as well as NEI: MEIR [44], random named entity swapping by topic (**R-NESt**) and CLIP-based named entity swapping by topic (**CLIP-NESt**) [40]. The number of samples per class for each dataset is given in Table 1.

Furthermore, we experiment with dataset aggregation, the combination of various generated datasets. Aggregated datasets are denoted with a plus sign, for instance *R-NESt + NC-t2t*. For the multiclass task, we combine one OOC dataset and at least one MC dataset to represent the OOC and MC classes, respectively. To evaluate the contribution of CHASMA (or CHASMA-D) in MMD we perform an ablation experiment where they are either integrated or excluded from aggregated datasets. Note that, during training, we apply random down-sampling to address any class imbalance.

Figure 4 presents a high-level overview of our employed pipeline. We incorporate truthful image-caption pairs from the VisualNews dataset and employ an OOC-based (e.g. NewsCLIPings) and a MC-based generation method (e.g. CHASMA) to create false OOC and MC pairs, respectively.

[11] https://github.com/MKLab-ITI/image-verification-corpus.

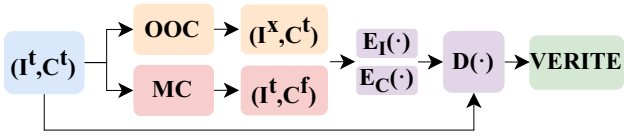

Fig. 4 High-level overview of the employed pipeline

Subsequently, we utilize CLIP to extract the visual and textual features from the image-caption pairs and then train the multiclass transformer detector $D(\cdot)$ before ultimately assessing its performance on the VERITE benchmark.

4.2 Evaluation protocol

Considering the distribution shift between training (generated) and test sets (real-world), utilizing an "out-of-distribution" validation set could potentially result in slightly better test accuracy [23]. However, due to the relatively small sizes of both COSMOS and VERITE datasets, we decided to avoid this approach. Instead, after training, we retrieve the best performing hyper-parameter combination based on the "in-distribution" validation set (generated) and evaluate it on the final test (real-world) sets: COSMOS and VERITE. For evaluation, we report the accuracy score (image-only, text-only or multimodal) for binary classification on COSMOS and multiclass accuracy on VERITE. Moreover, we experiment with a binary version of VERITE (VERITE-B) where both "OOC" and "MC" pairs are combined into a single class denoting misinformation. Here, we report the accuracy for each pair of classes, namely "True vs OOC" and "True vs MC". The number of samples per class for each evaluation dataset can be seen in Table 1.

Prior works using the VMU-Twitter dataset do not specify the validation set used for hyperparameter tuning [22, 45, 52]. By inspecting their code[12], [13] we can deduce that the test set was used for this purpose, which is problematic. We follow this protocol only for comparability and also train $D(\cdot)$ using a corrected protocol, where the development set is randomly split into training (90%) and validation (10%).

To evaluate the presence and magnitude of unimodal bias, we employ two metrics: the percentage increase in accuracy ($\Delta\%$) between a unimodal model and its multimodal counterpart, and Cohen's d (d) effect size. Negative $\Delta\%$ and positive d values serve as indicators for the presence of unimodal bias.

4.3 Implementation details

$D(\cdot)$ is trained for a maximum of 30 epochs (early stopping at 10 epochs) by the Adam optimizer with a learning rate of $lr = 5e-5$. For tuning the hyperparameters of $D(\cdot)$ consider the following values: $L \in \{1, 4\}$ transformer layers of $f \in \{128, 1024\}$ dimension of the feed-forward network model, $h \in \{2, 8\}$ attention heads. The dropout rate is constant at 0.1 and the batch size at 512. This grid-search results in a total of 8 experiments per modality (image-only, text-only, multimodal), thus 24 per dataset. For experiments on the VMU-Twitter dataset, we reduce the batch size to 16 and define $lr \in \{5e-5, 1e-5\}$, since it is a much smaller dataset. We set a constant random seed (0) for Torch, Python random and NumPy to ensure the reproducibility of our experiments. We conducted the experiments on a computer equipped with an AMD Ryzen 3960X 24-Core CPU, 128GB of RAM, and a single GeForce RTX 3060 GPU.

[12] https://github.com/dhruvkhattar/MVAE.

[13] https://github.com/shiivangii/SpotFake.

5 Experimental results

Image-side unimodal bias on VMU-Twitter: We begin by comparing the performance of $D(\cdot)$ with various models trained and evaluated on the VMU-Twitter dataset. In Table 2, we observe that among multimodal models, $D^-(I; C)$ achieves the third-highest result (80.5%), after Intra+Inter (83.1%) and BCMF (81.5%). However, it is noteworthy that the image-only model $D^-(I)$ achieves the highest overall accuracy (83.7%). This finding indicates the presence of image-side unimodal bias within models trained and evaluated on the VMU-Twitter. Table 7 also demonstrates that $D(I, C)$ displays a greater percentage decrease (-4.78%) compared to $D^-(I; C)$ (-3.92%), thus VMU-Twitter does not seem to allow the full utilization of multi-head self-attention.

Figure 5 demonstrates that the multimodal model $D(I, C)$ produces the same outputs regardless of whether the image is paired with its corresponding caption or two randomly selected captions. $D(I, C)$ predicts that all pairs are "true"

Table 2 Performance of transformer $D(I, C)$ and $D^-(\cdot)$ for caption-only (C), image-only (I) or multimodal inputs ($I; C$) when trained and evaluated on the VMU-Twitter dataset. Bold denotes the highest binary accuracy

Model	$E_I(\cdot)$	$E_C(\cdot)$	Accuracy
EANN [50]	VGG-19	TextCNN	71.5
MVAE [22]	VGG-19	BiLSTM	74.5
SpotFake [45]	VGG-19	BERT	77.8
BCMF [52]	DeiT	BERT	81.5
Intra+Inter [46]	Faster-RCNN	BERT	83.1
$D^-(C)$	–	CLIP	74.7
$D^-(I)$	CLIP	–	**83.7**
$D^-(I; C)$	CLIP	CLIP	80.5
$D(I, C)$	CLIP	CLIP	79.7

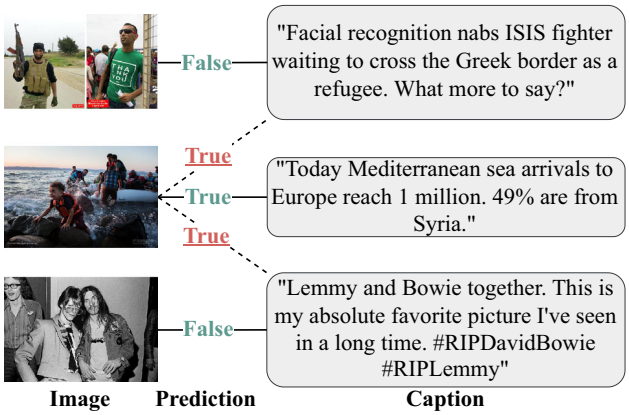

Fig. 5 Inference by $D(I, C)$ on three samples from VMU-Twitter. Moreover, we examine the model's image-side unimodal bias by inputting the middle image along with each of the three captions. $D(I, C)$ predicts "true" with all three captions, which means that the model does not take the caption into consideration. Red underlines denote mistaken predictions

regardless of the accompanying text. This example visually highlights the presence of image-side unimodal bias within the model's inference process.

The occurrence of image-side unimodal bias can be attributed to two primary factors. Firstly, VMU-Twitter was originally designed as an image verification corpus, comprising a substantial number of manipulated or edited images. Consequently, the significance of the accompanying text diminishes, as the primary source of misinformation lies within the image itself, what we term Asymmetric-MM. Secondly, VMU-Twitter exhibits an imbalance between the number of texts and images used for training and testing. With only 410 images available for training and 104 images for testing, compared to approximately 17k and 1k tweets, respectively, each image appears multiple times in the dataset, albeit with different texts. This discrepancy can lead to the model disregarding the textual modality, further reinforcing the image-side bias. Considering these factors, it appears that VMU-Twitter may not be an optimal choice for training and evaluating models for the task of MMD and might be better suited for its original purpose, namely, image verification.

As discussed in Sect. 4.2, it is worth highlighting that the evaluation protocol employed in [22, 45, 52] is problematic, using the test set during the validation and/or hyper-parameter tuning process. Under the corrected evaluation protocol, $D^-(I)$ achieves 81.0% accuracy, $D^-(I; C)$ achieves 77.3% (−4.56%), and $D(I, C)$ achieves 76.66% (−5.35%). The aforementioned conclusions regarding image-side bias remain consistent even under the corrected evaluation protocol.

Finally, note that a direct comparison between the models in Table 2 is not possible as they employ different image and text encoders. Consequently, we refrain from asserting that we have attained "state-of-the-art" performance on the VMU-Twitter. Instead, the results showcase that $D(\cdot)$ can provide competitive and reasonably strong performance—while being a relatively simple architecture—and will be leveraged in all preceding experiments.

Text-side unimodal bias on COSMOS: We proceed by training $D(\cdot)$ on various datasets for binary classification and evaluating on the COSMOS benchmark, as illustrated in Table 3. We observe that the text-only $D^-(C)$ trained on *CHASMA-D* achieves 72.6% accuracy on COSMOS, the highest accuracy score on COSMOS. However, this translates into the text-only model outperforming its multimodal counterparts, $D^-(I; C)$ and $D(I, C)$ by −7.85% and −14.88%, respectively. As seen in Table 7, on average, $D^-(C)$ outperforms $D^-(I; C)$ by 2.34% and $D(I, C)$ by 3.47% with a d of 0.25 and 0.4, respectively, highlighting that COSMOS does not seem to allow the full utilization of multi-head self-attention. We also observe in Table 3 that both $D(I, C)$ and $D^-(I; C)$ suffer from text-side uni-

Table 3 Results on the COSMOS benchmark. We report the performance of Transformer $D(I, C)$ and $D^-(\cdot)$ for caption-only (C), image-only (I) or multimodal inputs (C; I). Bold denotes the highest binary accuracy

Training	$D^-(I)$	$D^-(C)$	$D^-(I; C)$	$D(I, C)$
RSt	50.0	50.0	51.1	51.5
NC-t2t	45.1	50.0	50.8	52.6
CSt	45.5	47.4	52.6	52.2
MEIR	50.7	51.9	52.6	53.2
CLIP-NESt	50.0	55.4	53.6	53.4
R-NESt	50.0	59.5	54.1	55.2
Fakeddit	57.4	61.7	57.9	52.5
CHASMA	64.5	67.6	60.3	58.9
CHASMA-D	64.4	**72.6**	66.9	61.8

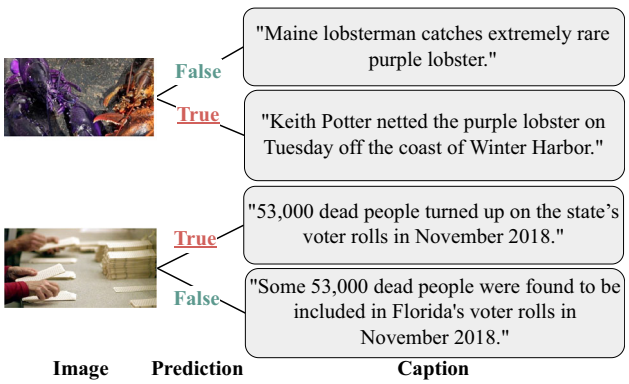

Fig. 6 Inference on two misleading samples from COSMOS with near-duplicate texts by $D(I, C)$ trained on CHASMA-D. Red underlines denote mistaken predictions

Table 4 Binary classification results on the test set of each dataset

Training Dataset	$D^-(C)$	$D^-(I)$	$D^-(I;C)$	$D(I,C)$
RSt	50.0	50.0	96.5	96.6
CSt	57.4	57.2	75.1	75.9
NC-t2t	50.0	53.7	84.0	84.3
MEIR	72.2	65.5	76.1	73.9
R-NESt	82.6	52.1	91.0	91.2
CLIP-NESt	65.5	54.0	70.4	70.3
Fakeddit	90.9	90.1	95.1	94.5
CHASMA	90.1	50.0	93.0	91.3
CHASMA-D	86.9	60.4	94.1	94.3

modal bias on COSMOS, only when trained with NEI-based datasets (CLIP-NESt and R-NESt) or datasets relying on human-written misinformation (Fakeddit and CHASMA). Text manipulations and human-written texts may display certain linguistic patterns that inadvertently the models learn to attend to while reducing attention towards the visual modality.

Figure 6 provides a visual representation of the behaviour of the multimodal model $D(I, C)$ when trained on CHASMA-D and evaluated on COSMOS. It showcases that the model can generate different outputs when applied to near-duplicate image-caption pairs, where the textual content exhibits only very minor differences that do not significantly alter the fact that it represents misinformation. Considering these results, we can conclude that COSMOS is not an ideal choice when it comes evaluating models for the task of multimodal misinformation detection. The dataset's characteristics and composition allow for the presence and reinforcement of text-sided unimodal bias, thereby yielding misleading or falsely optimistic outcomes.

Unimodal bias is not (entirely) algorithmic: Table 4 presents the performance of $D(\cdot)$ when trained on various datasets and evaluated on their respective test sets. When trained on OOC-based datasets (RSt, NC-t2t, and CSt) $D(\cdot)$ performs poorly in both image—and text-only settings—with an average of 53.6% and 52.5%, respectively, while achieving high multimodal accuracy. Expectedly, as both the image and the caption in OOC samples are factually accurate, but only their relation is corrupted, it is not possible to determine the existence of misinformation by solely examining one modality.

In contrast, $D^-(C)$ trained on NEI datasets (MEIR, R-NESt, CLIP-NESt) and CHASMA perform in closer proximity to the multimodal one, with $D^-(C)$ scoring 81.4%, compared to 86.6% by $D^-(I;C)$ and 85.9% by $D(I,C)$. At the same time, the image-only setting yields significantly lower performance for NEI methods and *CHASMA*; the only exception being *Fakeddit*, which comprises a higher percentage of manipulated images. Once again, these results suggest that methods relying on text manipulation or human-written misinformation may introduce linguistic patterns and biases that render the image less important.

However, unlike the COSMOS benchmark, no unimodal method surpasses its multimodal counterparts on the test sets. This is also demonstrated in Table 7, where neither $\Delta\%$ nor d indicate the presence of any unimodal bias. We can deduce that unimodal bias is partially algorithmic—an MMD model may rely on certain superficial unimodal patterns during training—but more importantly, these biases are significantly exacerbated by certain characteristics of VMU-Twitterand COSMOS—one of which is the high prevalence of Asymmetric-MM instances—thus raising concerns about their reliability as evaluation benchmarks.

VERITE alleviates unimodal bias: The analysis of Table 7 reveals that both $\Delta\%$ and Cohen's d effect sizes indicate the absence of any unimodal bias on the VERITE benchmark. Notably, $D(I,C)$ displays an average 27.94% increase in accuracy when compared to text-only $D^-(C)$ and 43.27% when compared to image-only $D^-(I)$. These results emphasize that a model biased towards one modality can not achieve satisfactory performance on VERITE. Furthermore, it is worth noting that $D(I,C)$ consistently outperforms $D^-(I;C)$, demonstrating that VERITE effectively allows for the power of multi-head self-attention to be leveraged, unlike COSMOS and VMU-Twitter.

Additionally, we train $D(\cdot)$ for binary classification and evaluate its performance on *VERITE-B*; the binary version of *VERITE*. The primary aim of these experiments is to investigate the implications of removing "modality balancing" from *VERITE* in relation to unimodal bias. This entails that each image no longer appears twice in *VERITE*, once in the "True" class and once in the "Miscaptioned" class, and each caption no longer appears twice, once in the "True" class and once in the "out-of-context" class; since they are separated into two separate evaluations. In Table 6, we observe that $D^-(I;C)$ trained on R-NESt or CHASMA-D exhibits minor instances of unimodal bias in the "True vs MC" evaluation. The scale of this bias becomes more pronounced when multi-head self-attention is employed in $D(I,C)$. Additionally, when trained with *Fakeddit*, $D(I,C)$ showcases unimodal bias within the "True vs OOC" metric. These findings bear similarities to the patterns identified within the COSMOS benchmark, albeit at a smaller scale, presumably due to the lack of Asymmetric-MM in VERITE. Based on these results, we can infer that "modality balancing" plays a crucial role in mitigating the manifestation of unimodal bias within *VERITE*. Hence, we advise against employing *VERITE-B* as an evaluation benchmark for multimodal misinformation detection, especially of MC pairs. Instead, we recommend utilizing the original *VERITE* benchmark, as it has demonstrated its robustness as a comprehensive evaluation framework.

Table 5 Multiclass classification results on the *VERITE* dataset with different training MC data. For OOC data, NC-t2t is used in all experiments

MC Data	$D(I, C)$				$D^-(I; C)$	$D^-(C)$	$D^-(I)$
	Accuracy	True	MC	OOC	Accuracy	Accuracy	Accuracy
CLIP-NESt	47.7	64.5	**35.2**	43.2	40.1	33.1	33.8
R-NESt	47.7	79.0	19.2	44.8	43.9	33.6	34.6
CHASMA	48.7	79.0	16.3	50.9	47.9	37.3	34.4
CHASMA-D	49.0	81.7	13.0	**52.5**	47.7	40.6	37.5
R-NESt+CHASMA	49.6	76.9	23.1	48.8	51.1	39.5	34.8
R-NESt+CHASMA-D	50.0	80.2	24.3	45.4	46.9	41.7	33.2
CLIP-NESt+CHASMA	50.8	**83.7**	21.0	47.5	49.6	41.8	33.4
CLIP-NESt+CHASMA-D	**52.1**	70.7	33.4	52.2	49.3	43.7	34.8

Table 6 Results on *VERITE-B* by $D(\cdot)$ trained on different datasets for binary classification. The objective of these experiments is to investigate the impact on unimodal bias when eliminating "modality balancing" from *VERITE*. Evaluation metrics used include "True vs OOC" and "True vs MC" accuracy. In parentheses, we report the percentage improvement ($\Delta\%$) of each multimodal model compared to the text-only model. Bold denotes the best performance per evaluation metric

	True vs OOC			True vs MC		
Training Dataset	$D^-(C)$	$D^-(I; C)$	$D(I, C)$	$D^-(C)$	$D^-(I; C)$	$D(I, C)$
Fakeddit	50.4	51.5 (2.2)	48.3 (−4.2)	58.7	55.9 (−4.8)	53.6 (−8.7)
CHASMA-D	50.4	52.6 (4.4)	52.0 (3.2)	**64.8**	64.5 (−0.5)	58.4 (−9.9)
R-NESt	50.0	66.2 (32.4)	67.2 (34.4)	59.2	59.6 (0.68)	58.6 (−1.0)
NC-t2t	46.5	72.4 (55.7)	72.0 (54.8)	50.0	54.4 (8.8)	54.6 (9.2)
R-NESt + CHASMA-D + NC-t2t	50.6	72.4 (43.1)	**72.7** (42.8)	58.4	63.9 (9.4)	61.2 (1.3)

Table 7 Examination of unimodal bias on different evaluation datasets. We report the average percentage increase in terms of accuracy ($\Delta\%$) and the average effect size measured by Cohen's d (d). Negative $\Delta\%$ and positive d values indicate the presence and magnitude of unimodal bias (denoted with bold)

Multimodal	$D(I, C)$				$D^-(I; C)$			
vs Unimodal	$D^-(C)$		vs $D^-(I)$		$D^-(C)$		vs$D^-(I)$	
Dataset	$\Delta\%$	d	$\Delta\%$	d	$\Delta\%$	d	$\Delta\%$	d
Test sets	25.33	−2.33	49.39	−1.02	25.67	−2.36	49.92	−1.05
COSMOS	**−3.47**	**0.41**	4.09	−0.28	**−2.34**	**0.25**	5.32	−0.39
VMU-Twitter	6.69	–	**−4.78**	–	7.76	–	**−3.82**	–
VERITE	27.94	−3.56	43.27	−10.41	21.38	−2.19	36.28	−4.68

On the performance of CHASMA: Table 5 provides a detailed overview of the results obtained on the VERITE evaluation benchmark. In our training process for multi-class misinformation detection, we employ $D(\cdot)$ using one out-of-context (OOC) dataset in combination with at least one NEI-based or CHASMA dataset, or both. We observe that $D(I, C)$ trained on *CHASMA + NC-t2t*(48.7%) or *CHASMA-D + NC-t2t*(49%) outperform both *CLIP-NESt + NC-t2t*(47.7%) and *R-NESt + NC-t2t*(47.7%). Furthermore, when $D(I, C)$ is trained on aggregated datasets that include *CHASMA*. It consistently outperforms those that do not. Notably, *CLIP-NESt + Misalign + NC-t2t* achieves the highest overall multiclass accuracy of 52.1%, representing a 9.22% improvement over *CLIP-NESt + NC-t2t*. Similar patterns are also reproduced while using $D^-(I; C)$. These findings highlight the effectiveness of the proposed methodology. By producing "harder" training samples and reducing the rate of Asymmetric-MM, CHASMA can significantly improve predictive performance on real-world data. Finally, it is worth noting that while $D(\cdot)$ trained on CHASMA displayed a high rate of text-side unimodal bias on COSMOS, this phenomenon is not present in the VERITE evaluation benchmark.

6 Conclusions

In this study, we address the task of multimodal misinformation detection (MMD) where an image and its accompanying caption collaborate to spread misleading or false information. Our primary focus lies in addressing the issue of unimodal bias, arising in datasets that exhibit distinct patterns

and biases towards one modality, which allows unimodal methods to outperform their multimodal counterparts in an inherently multimodal task. Our systematic investigation found that datasets widely used for MMD, namely VMU-Twitter and COSMOS, can enable image-side and text-side unimodal bias, respectively, raising questions about their reliability as benchmarks for MMD.

To address the aforementioned concerns, we introduce the VERITE evaluation benchmark, designed to provide a comprehensive and robust framework for multimodal misinformation detection. VERITE encompasses a diverse array of real-world data, excludes "asymmetric multimodal misinformation" (Asymmetric-MM)—where one modality plays a dominant role in propagating misinformation while others have little or no influence—and implements "modality balancing"; where each image and caption appear twice in the dataset, once within their truthful and once within a misleading pair. We conduct an extensive comparative study with a transformer-based architecture which demonstrates that VERITE effectively mitigates and prevents the manifestation of unimodal bias, offering an improved evaluation framework for MMD.

In addition, we introduce *CHASMA*, a novel method for generating synthetic training data that retain crossmodal relations between image-caption pairs. *CHASMA* employs a large pre-trained crossmodal alignment model to generate hard examples that retain crossmodal relations between legitimate images and misleading human-written captions. Empirical results show that using *CHASMA* in the training process consistently improves detection accuracy and has achieved the highest performance on VERITE.

The proposed approach achieved 52.1% accuracy for multiclass MMD. Nevertheless, we are optimistic that CHASMA and VERITE can serve as a foundation for future research, leading to further advancements in this area. For instance, future works could experiment with improved multimodal encoders [25, 26], news- or event-aware encoders [27], advanced modality fusion techniques [21, 51, 52], utilize external evidence [1] or explore new methods for generating training data [11]. As future research unfolds, VERITE could be expanded to include additional types of MM (e.g. AI-generated content) or additional modalities (e.g. videos), or be repurposed for other relevant tasks (e.g. fact-checked article retrieval [36]). Moreover, since MMD is only one part of multimedia verification [33], "claim detection" and "check-worthiness" [12] could be employed to distinguish between Asymmetric-MM and MM and determine whether to use a unimodal detector (e.g. false claim or manipulated image) or a multimodal misinformation detector in each scenario. Finally, while our focus has been on alleviating unimodal bias at the evaluation level, it may be worth exploring methods for reducing unimodal bias from an algorithmic perspective [10]. In all these endeavours, VERITE can serve as a robust evaluation benchmark.

Acknowledgements The publication of the article in OA mode was financially supported by HEAL-Link

Author Contributions S-IP involved in conceptualization, methodology, data curation, investigation, writing—original draft, and visualization.CK involved in conceptualization, writing—review and editing, and supervision. SP involved in conceptualization, writing—review and editing, supervision, resources, project administration, and funding acquisition. PC.P involved in conceptualization, review and editing, and supervision.

Funding Open access funding provided by HEAL-Link Greece. This work is partially funded by the Horizon Europe "vera.ai - VERification Assisted by AI" project under grant agreement no. 101070093.

Data availability All datasets are publicly available and have been properly referenced in the text.

Code availability Code is publicly available at https://github.com/stevejpapad/image-text-verification.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical approval Not applicable

Consent to participate Not applicable

Consent for publication All authors have provided their consent for publication.

References

1. Abdelnabi S, Hasan R, Fritz M (2022) Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14940–14949
2. Agrawal A, Batra D, Parikh D, et al (2018) Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4971–4980
3. Alam F, Cresci S, Chakraborty T et al (2022) A survey on multimodal disinformation detection. In: Proceedings of the 29th

international conference on computational linguistics, international committee on computational linguistics, pp 6625–6643
4. Aneja S, Bregler C, Niebner M (2023) Cosmos: catching out-of-context image misuse using self-supervised learning. In: Proceedings of the AAAI conference on artificial intelligence, pp 14084–14092
5. Aneja S, Midoglu C, Dang-Nguyen DT, et al (2021b) Mmsys' 21 grand challenge on detecting cheapfakes. arXiv preprint arXiv:2107.05297
6. Aneja S, Midoglu C, Dang-Nguyen DT, et al (2022) Acm multimedia grand challenge on detecting cheapfakes. arXiv preprint arXiv:2207.14534
7. Bennett WL, Livingston S (2018) The disinformation order: disruptive communication and the decline of democratic institutions. Eur J Commun 33(2):122–139. https://doi.org/10.1177/0267323118760317
8. Biamby G, Luo G, Darrell T et al (2022) Twitter-comms: detecting climate, covid, and military multimodal misinformation. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1530–1549
9. Boididou C, Middleton SE, Jin Z et al (2018) Verifying information with multimedia content on twitter: a comparative study of automated approaches. Multimed Tools Appl 77:15545–15571. https://doi.org/10.1007/s11042-017-5132-9
10. Cadene R, Dancette C, Cord M, et al (2019) Rubi: Reducing unimodal biases for visual question answering. Advances in neural information processing systems 32
11. Cardenuto JP, Yang J, Padilha R, et al (2023) The age of synthetic realities: Challenges and opportunities. arXiv preprint arXiv:2306.11503 https://doi.org/10.48550/arXiv.2306.11503
12. Cheema GS, Hakimov S, Sittar A et al (2022) Mm-claims: A dataset for multimodal claim detection in social media. In: Findings of the association for computational linguistics: NAACL 2022, pp 962–979
13. Duffy A, Tandoc E, Ling R (2020) Too good to be true, too good not to share: the social utility of fake news. Inf Commun Soc 23(13):1965–1979. https://doi.org/10.1080/1369118X.2019.1623904
14. Gamir-Ríos J, Tarullo R, Ibáñez-Cuquerella M, et al (2021) Multimodal disinformation about otherness on the internet. the spread of racist, xenophobic and islamophobic fake news in 2020. Anàlisi pp 49–64. https://doi.org/10.5565/rev/analisi.3398
15. Goyal Y, Khot T, Summers-Stay D, et al (2017) Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6904–6913
16. Guzhov A, Raue F, Hees J et al (2022) Audioclip: Extending clip to image, text and audio. ICASSP 2022–2022 IEEE international conference on acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 976–980
17. Hangloo S, Arora B (2022) Combating multimodal fake news on social media: methods, datasets, and future perspective. Multimedia Syst 28(6):2391–2422. https://doi.org/10.1007/s00530-022-00966-y
18. Heller S, Rossetto L, Schuldt H (2018) The ps-battles dataset-an image collection for image manipulation detection. arXiv preprint arXiv:1804.04866 https://doi.org/10.48550/arXiv.1804.04866
19. Jaiswal A, Sabir E, AbdAlmageed W, et al (2017) Multimedia semantic integrity assessment using joint embedding of images and text. In: Proceedings of the 25th ACM international conference on Multimedia, pp 1465–1471, https://doi.org/10.1145/3123266.3123385
20. Jindal S, Sood R, Singh R, et al (2020) Newsbag: A multimodal benchmark dataset for fake news detection. In: CEUR Workshop Proc., pp 138–145
21. Jing J, Wu H, Sun J et al (2023) Multimodal fake news detection via progressive fusion networks. Inf Process Manag 60(1):103120. https://doi.org/10.1016/j.ipm.2022.103120
22. Khattar D, Goud JS, Gupta M, et al (2019) Mvae: Multimodal variational autoencoder for fake news detection. In: The world wide web conference, pp 2915–2921, https://doi.org/10.1145/3308558.3313552
23. Koh PW, Sagawa S, Marklund H, et al (2021) Wilds: A benchmark of in-the-wild distribution shifts. In: International conference on machine learning, PMLR, pp 5637–5664
24. Levi O, Hosseini P, Diab M, et al (2019) Identifying nuances in fake news vs. satire: using semantic and linguistic cues. arXiv preprint arXiv:1910.01160 https://doi.org/10.48550/arXiv.1910.01160
25. Li J, Selvaraju R, Gotmare A et al (2021) Align before fuse: vision and language representation learning with momentum distillation. Adv Neural Inf Process Syst 34:9694–9705
26. Li J, Li D, Savarese S, et al (2023) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 https://doi.org/10.48550/arXiv.2301.12597
27. Li M, Xu R, Wang S, et al (2022) Clip-event: Connecting text and images with event structures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16420–16429
28. Li Y, Xie Y (2020) Is a picture worth a thousand words? an empirical study of image content and social media engagement. J Mark Res 57(1):1–19. https://doi.org/10.1177/00222437198811
29. Lin Z, Geng S, Zhang R, et al (2022) Frozen clip models are efficient video learners. In: European conference on computer vision, Springer, pp 388–404
30. Liu F, Wang Y, Wang T et al (2021) Visual news: Benchmark and challenges in news image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 6761–6771
31. Luo G, Darrell T, Rohrbach A (2021) Newsclippings: Automatic generation of out-of-context multimodal media. arXiv preprint arXiv:2104.05893 https://doi.org/10.48550/arXiv.2104.05893
32. Mridha MF, Keya AJ, Hamid MA et al (2021) A comprehensive review on fake news detection with deep learning. IEEE Access 9:156151–156170. https://doi.org/10.1109/ACCESS.2021.3129329
33. Mubashara A, Michael S, Zhijiang G, et al (2023) Multimodal automated fact-checking: A survey. arXiv preprint arXiv:2305.13507
34. Müller-Budack E, Theiner J, Diering S, et al (2020) 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, https://doi.org/10.1145/3372278.3390670
35. Nakamura K, Levy S, Wang WY (2020) 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
36. Nakov P, Da San Martino G, Elsayed T, et al (2021) The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In: advances in information retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43, Springer, pp 639–649, https://doi.org/10.1007/978-3-030-72240-1_75
37. Newman EJ, Garry M, Bernstein DM et al (2012) Nonprobative photographs (or words) inflate truthiness. Psychon Bull Rev 19:969–974. https://doi.org/10.3758/s13423-012-0292-0
38. Nielsen DS, McConville R (2022) Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In: Proceedings of the 45th international ACM SIGIR con-

ference on research and development in information retrieval, pp 3141–3153, https://doi.org/10.1145/3477495.3531744
39. Olan F, Jayawickrama U, Arakpogun EO, et al (2022) Fake news on social media: the impact on society. Information Systems Frontiers pp 1–16. https://doi.org/10.1007/s10796-022-10242-z
40. Papadopoulos SI, Koutlis C, Papadopoulos S, et al (2023) Synthetic misinformers: Generating and combating multimodal misinformation. In: Proceedings of the 2nd ACM international workshop on multimedia AI against Disinformation, pp 36–44, https://doi.org/10.1145/3592572.3592842
41. Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR, pp 8748–8763
42. Rana MS, Nobi MN, Murali B et al (2022) Deepfake detection: a systematic literature review. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3154404
43. Roozenbeek J, Schneider CR, Dryhurst S et al (2020) Susceptibility to misinformation about covid-19 around the world. Royal Society Open Sci 7(10):201199. https://doi.org/10.1098/rsos.201199
44. Sabir E, AbdAlmageed W, Wu Y, et al (2018) Deep multimodal image-repurposing detection. In: Proceedings of the 26th ACM international conference on Multimedia, pp 1337–1345, https://doi.org/10.1145/3240508.3240707
45. Singhal S, Shah RR, Chakraborty T, et al (2019) Spotfake: A multimodal framework for fake news detection. In: 2019 IEEE fifth international conference on multimedia big data (BigMM), IEEE, pp 39–47, https://doi.org/10.1109/BigMM.2019.00-44
46. Singhal S, Pandey T, Mrig S et al (2022) Leveraging intra and inter modality relationship for multimodal fake news detection. Companion Proc Web Conf 2022:726–734
47. Tahmasebi S, Hakimov S, Ewerth R et al (2023) Improving generalization for multi-modal fake news detection. In: Proceedings of the 2023 ACM international conference on multimedia retrieval, pp 581–585
48. Thorne J, Vlachos A, Christodoulopoulos C et al (2018) Fever: a large-scale dataset for fact extraction and verification. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (Long Papers), pp 809–819
49. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advances in neural information processing systems 30
50. Wang Y, Ma F, Jin Z, et al (2018) Eann: Event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, pp 849–857, https://doi.org/10.1145/3219819.3219903
51. Wu Y, Zhan P, Zhang Y et al (2021) Multimodal fusion with co-attention networks for fake news detection. Findings of the association for computational linguistics: ACL-IJCNLP 2021:2560–2569
52. Yu C, Ma Y, An L et al (2022) Bcmf: a bidirectional cross-modal fusion model for fake news detection. Inf Process Manag 59(5):103063. https://doi.org/10.1016/j.ipm.2022.103063
53. Zhang Y, Tao Z, Wang X, et al (2023) Ino at factify 2: Structure coherence based multi-modal fact verification. arXiv preprint arXiv:2303.01510
54. Zhou Y, Yang Y, Ying Q, et al (2023) Multimodal fake news detection via clip-guided learning. In: 2023 IEEE International conference on multimedia and expo (ICME), IEEE, pp 2825–2830
55. Zlatkova D, Nakov P, Koychev I (2019) Fact-checking meets fauxtography: verifying claims about images. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 2099–2108