No Evidence, No Problem: When Less is More for Out of Context Multimodal Misinformation Detection

Anonymous ACL submission

Abstract

The proliferation of multimodal misinformation, particularly Out-of-Context (OOC) image-text mismatches, poses significant challenges for reliable information verification. Existing detection approaches often rely on unimodal signals, limiting their capacity to capture nuanced cross-modal inconsistencies. Although recent multimodal methods have improved performance, many depend on large-scale architectures or external web evidence, which hinders scalability and practical deployment. In this work, we introduce a lightweight and evidence-free framework for OOC misinformation detection that achieves competitive performance with high efficiency. Our approach enhances visual understanding by integrating semantic entity extraction and generated visual captions, which are fused with the accompanying textual caption and input to a prompt-tuned Flan-T5 model. Simultaneously, a fine-tuned CLIP model evaluates image-text alignment. The outputs of both models are combined via a validation-optimized weighted ensemble. Extensive experiments on the NewsCLIPpings dataset demonstrate that our method achieves state-of-the-art accuracy among evidence-free techniques, while offering low computational overhead and strong interpretability, making it well-suited for real-world applications.

1 Introduction

Fake news refers to intentionally disseminated false or misleading information, typically in the form of news reports, designed to influence public opinion, shape emotions, or achieve specific political, economic, or social objectives (Shu et al., 2017; Kouzy et al., 2020). Among various manifestations of fake news, *Out-of-Context* (OOC) misinformation—where images and textual content are deliberately misaligned to deceive audiences—has emerged as a particularly challenging and pervasive problem. Unlike conventional fake news that fabricates textual content, OOC misinformation exploits authentic media elements in misleading contexts, complicating detection efforts relying on traditional fact-checking techniques. Figure 1 illustrates two real-world examples of OOC image-text pairs that have been falsely propagated on social media.

Existing research on misinformation detection can be broadly categorized into three approaches: *text-based*, *image-based*, and *multimodal*. Text-based methods (Ma et al., 2016; Yu et al., 2019; Shu et al., 2019; Dun et al., 2021) typically leverage natural language processing techniques to analyze linguistic patterns, writing style, sentiment, and contextual cues for fake news identification. While effective for purely textual misinformation, these approaches often struggle to detect deception embedded in accompanying visual content.

Image-based techniques (Qi et al., 2019) utilize computer vision methods—such as tampering detection, deepfake analysis, and visual anomaly detection—to identify manipulated or misleading images. However, these methods typically overlook the semantic relationship between images and their textual context, rendering them ineffective at detecting cross-modal inconsistencies.

Neither unimodal strategy suffices for detecting OOC misinformation, where the core deception arises from incongruities between the textual and visual modalities—a tactic frequently employed in social media disinformation campaigns. Authentic images may be paired with fabricated captions or vice versa, making it crucial to jointly evaluate both modalities.

In response, multimodal deep learning approaches have gained traction (Singhal et al., 2019; Giachanou et al., 2020), often leveraging transformer architectures (Dosovitskiy et al., 2020) and convolutional neural networks to model cross-modal relationships. These methods aim to capture semantic inconsistencies between image-caption

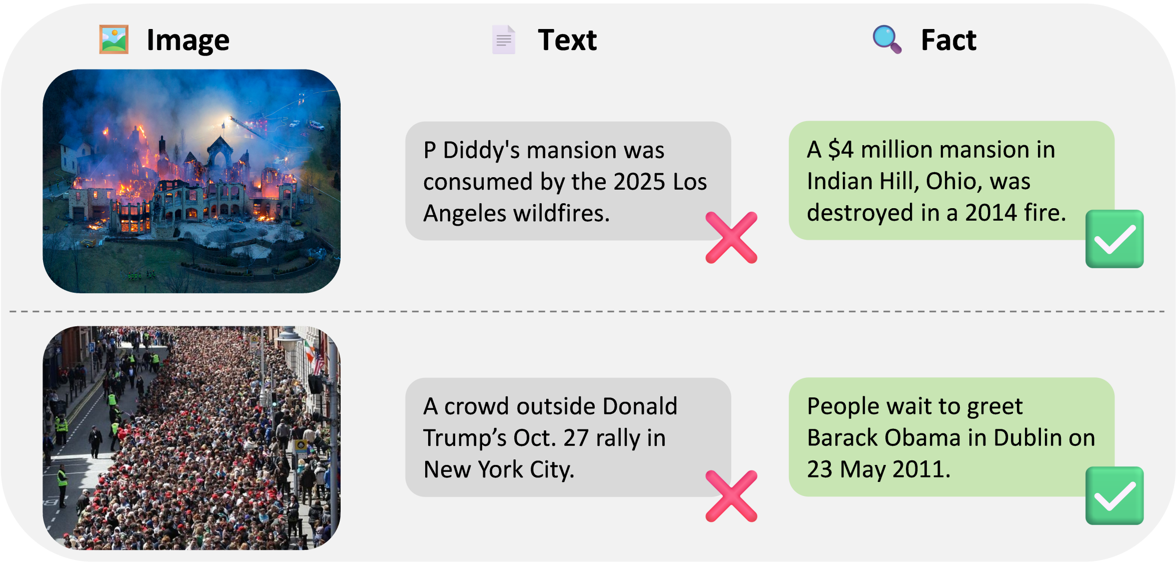
Image
Text
Fact
P Diddy's mansion was consumed by the 2025 Los Angeles wildfires.
A $4 million mansion in Indian Hill, Ohio, was destroyed in a 2014 fire.
A crowd outside Donald Trump's Oct. 27 rally in New York City.
People wait to greet Barack Obama in Dublin on 23 May 2011.

Figure 1: Examples of OOC fake news. Top: A 2014 Ohio fire misrepresented as Diddy's 2025 LA mansion. Bottom: A 2011 Obama visit crowd falsely claimed to be from a Trump rally.

pairs, achieving promising results. Nonetheless, challenges remain with respect to classification accuracy, reliance on external evidence for verification, and computational inefficiency, particularly for deployment at scale.

In this work, we address these limitations by incorporating enriched image-derived semantic cues, including visual captions and entity recognition, to strengthen multimodal feature representations. Our framework fine-tunes and optimally fuses outputs from pretrained text and image encoders, improving detection accuracy without sacrificing computational efficiency. This positions our approach as a practical solution for real-world OOC misinformation detection. Our contributions are as follows:

- We propose an evidence-free multimodal framework that achieves strong performance with high computational efficiency for OOC misinformation detection, combining semantic entity recognition and generated visual captions to enrich visual feature representation.

- We introduce a weighted ensemble strategy optimized on a validation set to effectively fuse predictions from the two modalities.

- Extensive experiments on the NewsCLIPpings dataset demonstrate that our approach achieves state-of-the-art accuracy among evidence-free methods, with reduced computational overhead and increased interpretability.

2 Related Work

2.1 Out-of-Context Misinformation Detection

Out-of-Context (OOC) misinformation involves authentic images or videos repurposed with misleading textual descriptions, posing a subtle yet impactful challenge in fake news detection. Early unimodal methods (Ma et al., 2016; Yu et al., 2019; Shu et al., 2019; Dun et al., 2021; Qi et al., 2019) and initial multimodal approaches (Khattar et al., 2019; Kumari and Ekbal, 2021; Singhal et al., 2019; Giachanou et al., 2020; Hua et al., 2023) have contributed to fake news detection, but often struggle to capture the semantic inconsistencies between visual and textual content that characterize OOC misinformation. Luo et al. (Luo et al., 2021) introduced the NewsCLIPpings dataset to systematically benchmark image-text mismatches, demonstrating the difficulty both humans and machines face in this task.

To improve detection, some approaches utilize external evidence sources. For example, Abdelnabi et al. (Abdelnabi et al., 2022) retrieved web content to verify consistency among images, captions, and related articles, proposing a Cross-Modal Consistency Network (CCN) to jointly evaluate multimodal and metadata alignment. However, reliance on external evidence can limit applicability due to evidence availability, retrieval latency, and potential reliability issues.

Alternative evidence-free methods have also been proposed. SSDL (Mu et al., 2023) leverages self-supervised multimodal pretraining and knowledge distillation to improve cross-modal representation alignment. DPOD (Bhattacharya et al., 2023) adapts CLIP models with domain-aware prompts and out-of-domain examples to enhance generalization across varied misinformation topics. Despite robustness gains, such latent feature-based models often lack interpretability. To enhance explainability, Zhang et al. (Zhang et al., 2023) employed Abstract Meaning Representation to convert text into fact-based queries for visual verification, improving transparency.

More recently, Multimodal Large Language Models (MLLMs) like InstructBLIP have been explored for OOC detection. SNIFFER (Qi et al., 2024) fine-tuned MLLMs on GPT-4-generated instructions to improve reasoning in news verification. LLM-Consensus (Lakara et al., 2024) proposes lightweight, explainable frameworks through dialectic debates among LLM agents, avoiding task-specific fine-tuning. Nevertheless, these approaches are computationally expensive and require continual model updates. Moreover, as highlighted by Yan et al. (Yan et al., 2025), dependence on external evidence—particularly web content increasingly polluted by AI-generated misinformation—introduces critical vulnerabilities that can drastically degrade model reliability.

In summary, while current methods advance OOC detection, challenges persist related to dependence on external evidence, limited generalization, computational inefficiency, and lack of interpretability. These limitations motivate the need for compact, evidence-free approaches that achieve strong accuracy, scalability, and transparency.

2.2 Prompt-based Learning

Prompt-based learning reformulates downstream tasks as natural language inference problems through designed input templates, rather than appending task-specific output heads (Brown et al., 2020; Liu et al., 2022). This paradigm aligns better with large pretrained language models' (PLMs) original training and enhances generalization.

Instruction-tuned models such as Flan-T5 (Chung et al., 2022) have demonstrated strong performance with minimal labeled data, excelling in zero- and few-shot scenarios (Gao et al., 2020). Their responsiveness to natural language prompts enables flexible and interpretable task definitions.

In this work, we leverage the OpenPrompt framework (Ding et al., 2021) to design effective prompt templates for multimodal misinformation detection. By framing the task as masked prediction guided by natural prompts, we exploit Flan-T5's instruction-following capability, improving alignment with human reasoning and enhancing model interpretability.

3 Methodology

3.1 Visual Information Augmentation

To enrich the model's understanding of visual content beyond raw pixel data, we incorporate high-level semantic cues extracted via two complementary strategies.

First, we obtain visual entities (e.g., "person", "car", "building") for each image using the Google Vision API (Google, 2019). Unlike implicit semantics learned end-to-end, these explicit annotations offer interpretability and modularity by summarizing the most salient image elements. This helps the model focus on crucial visual cues when cross-referencing with textual captions. Additionally, the API has demonstrated robustness across diverse domains, ensuring consistency on real-world news images. These annotations are sourced from the preprocessed dataset in (Abdelnabi et al., 2022), facilitating reproducibility and straightforward integration.

Second, we generate visual captions using the GPT-4o mini model from OpenAI (OpenAI, 2024a). Unlike traditional captioning models trained on fixed datasets, GPT-4o mini benefits from multimodal instruction tuning, enabling it to produce context-aware, human-like descriptions. It strikes an effective balance between accuracy, inference speed, and resource efficiency.These captions provide an additional textual abstraction complementing raw images and entity annotations.

While our current implementation focuses solely on GPT-4o mini, we selected it for its ease of integration, competitive performance in preliminary tests, and suitability for scalable deployment. Future work can explore a broader comparison with dedicated visual captioning models.

For a small subset of images (approximately 1.6%) where caption generation fails (e.g., close-up portraits lacking context, low-quality images), we employ fallback strategies such as assigning an empty string or a placeholder. These rare cases minimally affect overall training and evaluation, while preserving pipeline continuity and reducing noise from unreliable visual inputs.

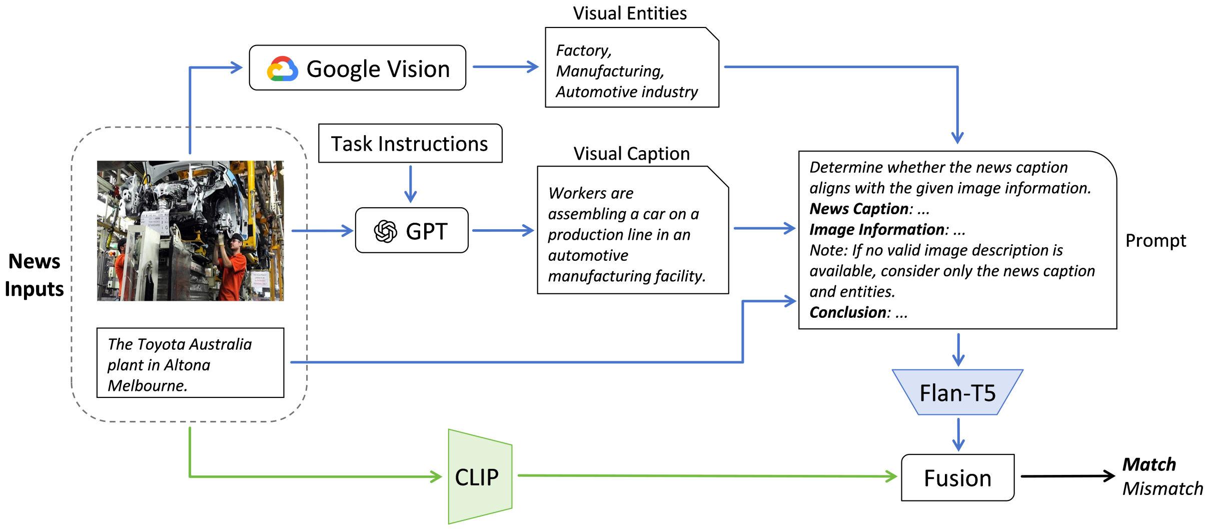
Google Vision
Visual Entities
Factory, Manufacturing, Automotive industry
Task Instructions
GPT
Visual Caption
Workers are assembling a car on a production line in an automotive manufacturing facility.
News Inputs
The Toyota Australia plant in Altona Melbourne.
Determine whether the news caption aligns with the given image information.
News Caption: ...
Image Information: ...
Note: If no valid image description is available, consider only the news caption and entities.
Conclusion: ...
Prompt
Flan-T5
CLIP
Fusion
Match
Mismatch

Figure 2: Overview of the proposed framework architecture.

The enriched visual entities and captions are subsequently used to augment input features during multimodal model fine-tuning, enhancing the detection of semantic inconsistencies between image-text pairs.

3.2 Prompt-tuning Strategy

To effectively capture semantic alignment between visual and textual information, we adopt a prompt-tuning approach based on the Flan-T5 language model (Chung et al., 2022).

Flan-T5 is an instruction-finetuned variant of the T5 model (Raffel et al., 2020), trained on a diverse collection of tasks formulated as text-to-text transformations. It offers an excellent balance of efficiency, versatility, and performance, making it well-suited for resource-constrained scenarios. Its instruction-following capabilities facilitate generalization on downstream tasks with minimal task-specific fine-tuning, aligning well with our goals of lightweight and interpretable misinformation detection.

3.2.1 Prompt Construction

We cast the OOC misinformation detection task as a natural language inference problem. Each input is framed as a question that queries whether the image-related information supports the given news caption. The prompt template is manually designed as:

> *Determine whether the news caption aligns with the given image information.*
> *News Caption: [CAPTION].*
> *Image Information: [IMAGE-INFO].*
> *Note: If no valid image description is available, consider only the news caption and entities.*
> *Conclusion: [MASK]*

Here, `[CAPTION]` is the original news caption, and `[IMAGE-INFO]` combines the detected visual entities and generated image caption. `[MASK]` is the token the model predicts to indicate alignment or mismatch.

The *Note* line acts as a contingency instruction for cases where image captioning fails or visual content is unavailable, directing the model to rely solely on textual inputs. This improves robustness and leverages Flan-T5's instruction tuning to dynamically adapt its reasoning based on input completeness.

This prompt formulation guides the model to perform semantic comparison framed in natural language, fully exploiting Flan-T5's instruction-following strengths.

3.2.2 Verbalizer Design

To map model outputs to the target classes, we define a verbalizer associating specific output tokens with each label. The *match* class is linked to words such as *"match"*, *"correct"*, *"aligned"*, *"consistent"*, and *"true"*. The *mismatch* class corresponds to terms like *"mismatch"*, *"incorrect"*, *"not aligned"*, *"inconsistent"*, and *"false"*. These associations facilitate interpretable token-to-label mapping, enhancing transparency in classification decisions.

3.3 Model Architecture

Our system ingests four input components per sample: the **news caption**, the **raw image**, the list of **visual entities**, and the **visual caption**. These inputs feed two complementary models whose outputs are fused for final classification, as illustrated in Figure 2.

Flan-T5 Module We fine-tune Flan-T5 using the prompt template described above, leveraging the concatenated news caption, visual entities, and visual caption. This lightweight language model learns to classify semantic consistency by producing a probability distribution over the classes *Pristine* and *Falsified*. Flan-T5 excels at structured textual reasoning with relatively low computational cost.

CLIP Module In parallel, we fine-tune CLIP (Radford et al., 2021), a vision-language model trained to embed images and text into a shared latent space. Using the paired news caption and raw image, CLIP is trained to predict their alignment, also outputting class probabilities. CLIP is adept at capturing cross-modal correspondences between visual and textual data.

Fusion We combine Flan-T5 and CLIP predictions through a weighted ensemble, where the fusion weight is optimized on a validation set. This ensemble leverages Flan-T5's strengths in linguistic reasoning and CLIP's visual-semantic alignment capabilities, leading to improved overall detection accuracy while maintaining efficiency and interpretability.

4 Experiments

4.1 Dataset

The NewsCLIPpings dataset (Luo et al., 2021) is a large-scale benchmark specifically designed for multimodal fake news detection. Each sample in the dataset consists of a textual description and an accompanying image, and the dataset is divided into two categories: pristine samples, where the image and text are contextually aligned, and falsified samples, where the image has been replaced with a semantically similar one from a different event to create an OOC scenario.

NewsCLIPpings is constructed from the VisualNews (Liu et al., 2020) corpus, which sources articles from major outlets such as the BBC, The Guardian, USA Today, and The Washington Post. It has become one of the most comprehensive benchmarks for evaluating models on OOC misinformation detection.

Following previous studies, this project uses the Merge/Balanced subset of the dataset, which contains an equal number of real and fake samples. It is split into 71,072 samples for training, 7,024 for validation, and 7,264 for testing, maintaining a 10:1:1 ratio.

4.2 Experimental Setup

All model training and inference were conducted on a single NVIDIA A6000 GPU with 48 GB of memory. We fine-tuned our prompt-based language model using the OpenPrompt (Ding et al., 2021) framework, along with PyTorch and scikit-learn for model training and evaluation.

During fine-tuning, we adopted a grouped parameter optimization strategy. Parameters such as bias and LayerNorm.weight were excluded from weight decay and assigned a weight decay factor of zero. All remaining parameters were regularized using a weight decay factor of 0.01. Optimization was carried out using the AdamW (Loshchilov and Hutter, 2019) optimizer, with the learning rate set to 1e-4. The loss function used was cross-entropy (Mannor et al., 2005) loss. We used a batch size of 16 and trained the model for a total of 10 epochs.

To explore the balance between efficiency and performance, we conducted experiments with both Flan-T5-Small and Flan-T5-Large (Chung et al., 2022). CLIP (Radford et al., 2021) was fine-tuned using the ViT-B/32 (Dosovitskiy et al., 2020) variant with paired image and caption data.

5 Results

5.1 Performance Comparison

We present a comparative analysis of our model's performance against a series of state-of-the-art (SOTA) baselines on the NewsCLIPpings dataset. To ensure fairness, we restrict our comparison to models that do not utilize external evidence such as web search information, and maintain relatively small parameter sizes to ensure efficiency and feasibility in real-world scenarios.

The baselines include two models trained from scratch — SAFE (Zhou et al., 2020) and

Model	**Accuracy (%)**
SAFE	52.8
EANN	58.1
VisualBERT	58.6
VINVL	65.4
SSDL	65.6
CLIP	66.0
SDG	68.0
Neu-Sym detector	68.2
GPT-4o	70.7
DPOD	74.4
DT-Transformer	77.1
Ours	**84.4**

Table 1: Performance comparison between our model and baselines.

Model	**VisCap**	**Fusion**	**All**	**Fake**	**Real**	**F1 Score**
Flan-T5-Small	✗	✗	75.3	75.2	75.3	0.75
Flan-T5-Small	✗	✓	79.0	78.7	79.2	0.79
Flan-T5-Small	✓	✗	80.7	78.5	82.9	0.80
Flan-T5-Small	✓	✓	81.8	80.3	**83.3**	0.81
Flan-T5-Large	✗	✗	76.8	85.3	68.3	0.79
Flan-T5-Large	✗	✓	80.9	83.3	78.4	0.81
Flan-T5-Large	✓	✗	83.9	87.5	80.3	0.84
Flan-T5-Large	✓	✓	**84.4**	**87.8**	80.9	**0.85**

Table 2: Ablation study of visual caption and fusion strategies. The table reports both Accuracy and F1 Score for each configuration.

EANN (Wang et al., 2018) — as well as a variety of models leveraging pre-trained architectures, such as VisualBERT (Li et al., 2019), VINVL (Huang et al., 2022), SSDL (Mu et al., 2023), CLIP (Radford et al., 2021), SDG (Shalabi et al., 2023), Neu-Sym detector (Zhang et al., 2023), GPT-4o (OpenAI, 2024b), DPOD (Bhattacharya et al., 2023), and DT-Transformer (Papadopoulos et al., 2023). The performance metrics for these models are collected from existing literature and published benchmark results.

As shown in the Table 1, our proposed method achieves 84.4% accuracy, outperforming all listed baselines by a significant margin. This demonstrates the effectiveness of combining lightweight prompt-tuned language models and image-language models through a weighted fusion strategy.

Compared to models like CLIP or even more recent ones like DPOD and DT-Transformer, our model maintains a competitive advantage without relying on large-scale parameters or external evidence. The performance gain can be attributed to the enriched visual understanding from both entity-level and caption-level information, and the synergy between the prompt-based text model and the vision-language alignment model.

These results confirm that our method sets a new standard for efficient and accurate OOC misinformation detection in the absence of external verification resources.

5.2 Ablation Study

To investigate the contribution of different components in our model, we conduct an ablation study on both the Flan-T5-Small and Flan-T5-Large backbones. Specifically, we examine the impact of two key factors: (1) the inclusion of visual captions in image representation, and (2) the integration of the prompt-tuned language model with CLIP. The results are summarized in Table 2.

From the results, we observe the following:

- Adding visual captions significantly improves both models. For instance, Flan-T5-Small's accuracy rises from 75.3% to 80.7% without CLIP, thanks to the rich semantic context captions provide—coherent sentences that go beyond sparse entity labels. When combined with CLIP, the gain remains at 1.1%, suggesting partial redundancy but still a complementary effect.

- Integrating CLIP consistently boosts performance by enhancing cross-modal alignment. Flan-T5-Small improves from 75.3% to 79.0% with CLIP alone. For Flan-T5-Large, CLIP adds 4.1%, while visual captions offer a larger gain of 7.1%, indicating the larger model captures some of CLIP's functionality, reducing its marginal impact.

- The highest performance is achieved when combining both cues. Flan-T5-Large with visual captions and CLIP reaches 84.4% accuracy and an F1 score of 0.85, demonstrating the benefit of fusing semantic richness with cross-modal reasoning, even without relying on external evidence.

- Flan-T5-Large consistently outperforms its smaller counterpart across all settings. However, with nearly 10 times more parameters, it comes at a significant computational cost. This highlights the practical value of Flan-T5-Small, which achieves strong results when

Caption: Virginia Tech s Thor includes artificial elastic muscles. It will not be ready until 2014 so will be replaced by a less advanced substitute at this stage.
Ground Truth: Real
Prediction: Fake

Caption: Gambling tables at Venetian Macau Queensland wants some Asian highrollers too.
Ground Truth: Fake
Prediction: Real

Caption: Apple experienced problems of its own following the release of iOS 8.
Ground Truth: Real
Prediction (FlanT5): Fake

Figure 3: Three representative error cases selected from the test set.

paired with visual captions and CLIP, offering a more efficient option for deployment.

These results validate our design choices and confirm that each component contributes positively to the final performance. In particular, the combination of visual captioning and cross-modal fusion via CLIP is essential for achieving SOTA results within the class of models that operate without accessing external evidence sources.

5.3 Error Analysis

To better understand our model's limitations, we conducted a qualitative analysis of misclassified test samples. Figure 3 presents three representative cases highlighting common failure modes: temporal ambiguity, contextual mismatch, and modality imbalance.

In the first case, the caption correctly describes Virginia Tech's humanoid robot and its unavailability until 2014. Despite relevant visual content, the model misclassified it as fake due to a lack of temporal cues in the image, revealing difficulty in reasoning about abstract, time-sensitive concepts.

The second case involves a fake caption referencing the Venetian Macau's gambling tables. The image, however, depicts a generic shopping-like scene with entities such as "Display window" and "Shopping". The model incorrectly predicted it as real, showing its struggle with fine-grained location grounding and named-entity resolution.

In the third case, the caption discusses iOS 8 issues, paired with an image of someone holding an iPhone 6. Flan-T5 flagged it as fake due to a perceived mismatch between software and hardware, while CLIP focused on topical relevance. The final ensemble prediction was correct, showcasing the models' complementary strengths.

These errors suggest future improvements could come from enhanced temporal reasoning, better visual grounding, and retrieval-based verification.

5.4 Discussion

In previous sections, we demonstrated that our model achieves the highest accuracy among all methods that do not rely on external evidence. However, we acknowledge that real, trustworthy, and high-quality evidence can significantly improve performance. Compared to evidence-based models (Abdelnabi et al., 2022; Yuan et al., 2023; Qi et al., 2024; Lakara et al., 2024), our model exhibits approximately a 5% drop in accuracy. In this section, we investigate how our model complements evidence-based reasoning and analyze the benefits and limitations of incorporating such external information.

We conducted a supplementary experiment using 200 test samples (balanced real/fake). For each image, we used Google Vision API (Google, 2019) to retrieve webpages, following the protocol in (Abdelnabi et al., 2022). This yielded 136 valid search results, from which 112 samples produced usable textual evidence (e.g., titles, image captions).

We evaluated three settings: Flan-T5-Large alone, GPT-4o mini with evidence, and a weighted fusion of both. As shown in Table 3, on the full 200-sample set, the fusion model reached 88.0% accuracy—just 2% higher than Flan-T5-Large—due to GPT-4o mini's poor performance without evidence. However, on the 112 samples with valid evidence, fusion achieved 94.6%, outperforming Flan-T5-Large by 7.1% and GPT-4o mini by 1.7%. These

Model	Setting	Accuracy (%)
Flan-T5-Large	All 200 samples	86.0
GPT-4o mini	All 200 samples	85.0
Fusion	All 200 samples	88.0
Flan-T5-Large	112 samples with evidence	87.5
GPT-4o mini	112 samples with evidence	92.9
Fusion	112 samples with evidence	**94.6**

Table 3: Accuracy comparison of different models under varying evidence availability.

results underscore the value of combining multimodal reasoning with retrieved evidence when available.

Despite the promising results, several limitations emerged during this process. First, only 56% of the samples yielded valid evidence. In many cases, no relevant webpages were returned, or the retrieved pages lacked accessible or meaningful metadata. This significantly limits the scalability and consistency of evidence-based approaches, particularly in time-sensitive or obscure scenarios where relevant web content may be scarce or non-indexed.

Second, for a subset of webpages, the evidence retrieval process was noticeably slow. Some pages contained excessive multimedia content, such as high-resolution images, videos, or dynamic loading structures, which increased parsing time and sometimes led to failure in extracting relevant text. These issues not only make it difficult to generate the entire evidence dataset, but also pose challenges for real-time deployment.

Third, there exists a potential risk in the quality and reliability of retrieved evidence itself. Some top-ranked sources may come from user-generated content or posts from platforms like X, Facebook, or non-reputable news websites. When the evidence comes from such sources, it may itself be misleading or entirely false, compounding the misinformation problem rather than mitigating it. This introduces a new form of vulnerability where the model may incorrectly validate fake content based on unreliable evidence.

While we do not position our model as universally lightweight, it is substantially more efficient than recent high-performing models such as SNIFFER (Qi et al., 2024), which rely on multimodal large language models and web evidence retrieval. Compared to SNIFFER, which uses models exceeding 13B parameters, our system (under 1B parameters) is over 90% smaller. Despite this, it achieves 84.4% accuracy (vs. 88.4%), with only a 4% drop, and requires no evidence pipeline—making it significantly faster and more deployable in practice.

This trade-off reflects a deliberate balance between performance and efficiency, making our model more suitable for scalable or latency-sensitive applications. Rather than claiming absolute lightweight design, we emphasize its practicality relative to heavier architectures.

6 Conclusion

In this paper, we proposed a multimodal framework for detecting out-of-context misinformation by integrating visual captioning, CLIP-based alignment, and a prompt-tuned Flan-T5 model. Experiments demonstrate strong performance without external evidence, outperforming prior methods in this setting. Further improvements were observed by combining our model with GPT-based evidence reasoning, highlighting its adaptability to real-world scenarios where evidence is partially available.

Future work will focus on enhancing the use of external evidence to further improve detection accuracy. Although incorporating web-based evidence has shown promising gains, better strategies are needed for handling such evidence efficiently and reliably. This includes expanding image-based web retrieval coverage, improving the speed and precision of content extraction through focused scraping, and developing automated methods to assess the credibility of retrieved sources. Additionally, exploring fine-grained image grounding and temporal reasoning may help capture more nuanced or time-sensitive mismatches between images and text.

Limitations

While our approach demonstrates strong performance without relying on external evidence, it still encounters notable challenges. Compared to models that leverage high-quality external evidence, our model exhibits lower accuracy. It also struggles with understanding abstract temporal references, accurately grounding fine-grained location claims, and detecting subtle inconsistencies across modalities. These limitations are particularly evident in complex real-world scenarios where visual and textual elements diverge in nuanced or context-dependent ways.

In our further experiments with external web evidence, we found that retrieving relevant supporting content is not always feasible. Many samples could not be matched to appropriate webpages, and even when matches were found, the extraction process was often hindered by webpage structure and noise. Moreover, relying on external sources introduces the risk of incorporating misleading or low-credibility content, especially from social media or unverified platforms. These issues collectively affect the consistency, reliability, and scalability of evidence-based enhancements, highlighting the need for more robust and trustworthy retrieval and filtering mechanisms.

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Amartya Bhattacharya, Debarshi Brahma, Suraj Nagaje Mahadev, Anmol Asati, Vikas Verma, and Soma Biswas. 2023. Can out-of-domain data help to learn domain-specific prompts for multimodal misinformation detection?

Tom Brown, Benjamin F Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Thomas Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey C.S. Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *ArXiv (Cornell University)*, 4.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, and 12 others. 2022. Scaling instruction-finetuned language models. *arXiv:2210.11416 [cs]*.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv (Cornell University)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:81–89.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal multi-image fake news detection. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*.

Google. 2019. Cloud computing services | google cloud.

Jiaheng Hua, Xiaodong Cui, Xianghua Li, Keke Tang, and Peican Zhu. 2023. Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing*, 136:110125.

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

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. *The World Wide Web Conference*.

Ramez Kouzy, Abi Jaoude Joseph, Afif Kraitem, Molly , Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie Akl, and Khalil Baddour. 2020. Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12.

Rina Kumari and Asif Ekbal. 2021. Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications*, 184:115412.

Kumud Lakara, Juil Sock, Christian Rupprecht, Philip Torr, John Collomosse, and Christian Schroeder. 2024. Mad-sherlock: Multi-agent debates for out-of-context misinformation detection. *arXiv (Cornell University)*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv (Cornell University)*.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visual news: Benchmark and challenges in news image captioning. *arXiv (Cornell University)*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv (Cornell University)*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Shie Mannor, Dori Peleg, and Reuven Rubinstein. 2005. The cross entropy method for classification. *Proceedings of the 22nd international conference on Machine learning - ICML '05*.

Michael Mu, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2023. Self-supervised distilled learning for multi-modal misinformation identification. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 33:2818–2827.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Hello gpt-4o.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. 2023. Synthetic misinformers: Generating and combating multimodal misinformation. *Zenodo (CERN European Organization for Nuclear Research)*.

Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. *2019 IEEE International Conference on Data Mining (ICDM)*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv.org*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Fatma Shalabi, Huy H Nguyen, Hichem Felouat, Ching-Chun Chang, and Isao Echizen. 2023. Image-text out-of-context detection using synthetic multimodal misinformation. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 605–612.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19:22–36.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*.

S Singhal, Shah R R, T Chakraborty, P Kumaraguru, and S Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. *IEEE Xplore*, page 39–47.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Zehong Yan, Peng Qi, Wynne Hsu, and Mong Li Lee. 2025. Mitigating genai-powered evidence pollution for out-of-context multimodal misinformation detection. *arXiv (Cornell University)*.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2019. Attention-based convolutional approach for misinformation identification from massive and noisy microblog posts. *Computers & Security*, 83:106–121.

Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 2023. Support or refute: Analyzing the stance of evidence to detect out-of-context mis- and disinformation. *arXiv (Cornell University)*.

Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. 2023. Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model. *arXiv (Cornell University)*.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection. *arXiv (Cornell University)*.