CRC: Knowledge-Enhanced Caption Reconstruction and Comparison for Detecting Out-of-Context Misinformation

Jie Li[1], Yuqiu Deng[1], Simin Yang[1], Jinrui Wang[1], Yunbo Rao[2†]

Beijing University of Posts and Telecommunications[1]

University of Electronic Science and Technology of China[2]

Abstract

The proliferation of multimodal misinformation, particularly out-of-context (OOC) misinformation, poses a pressing societal challenge due to its deceptive nature and ease of fabrication, where authentic modalities are deliberately mismatched. Existing approaches to OOC misinformation detection often fail to extract the rich semantic information available in the news image. To address this, we propose Knowledge-enhanced **C**aption **R**econstructor and **C**omparison (CRC), a novel framework integrating web knowledge for misinformation detection. Specifically, we first retrieve textual knowledge from the web using news images and visual knowledge along with its corresponding captions from the web using news captions. Next, Vision-Language Models (VLMs) generate synthetic captions from news images using prompts augmented with textual knowledge. For inconsistency detection, news images are compared against visual knowledge, while synthetic captions are compared against both original news captions and retrieved captions. Finally, consistency embeddings are combined through dual-modal attention module for final predictions. Experiments on large-scale datasets demonstrate effectiveness of our method.

Keywords: Out-of-context misinformation detection; Knowledge-enhanced; Natural language processing; Vision language Model; Data augmentation

1. Introduction

The rise of multimodal content, combining text and images, has become a cornerstone of modern news dissemination, particularly on social media platforms. However, this progress has also facilitated the widespread dissemination of misinformation [1], particularly through the actions of malicious actors employing sophisticated manipulation techniques. Among the various forms of misinformation, out-of-context (OOC) misinformation is notably prevalent due to its simplicity and effectiveness. This type of misinformation involves the deliberate pairing of authentic images or videos with misleading captions, creating a deceptive alignment that misrepresents the original context. For instance, during the COVID-19 pandemic, numerous instances of OOC misinformation circulated on social media platforms, such as images from unrelated events falsely presented as evidence of severe pandemic-related outcomes, thereby undermining public trust and exacerbating social unrest. Addressing this issue is critical, as OOC misinformation

[†]Corresponding author: Yunbo Rao (Email: uestc2008@126.com)

poses significant challenges to information integrity and public discourse, necessitating further research into its detection and mitigation.

To address the challenges of OOC misinformation, significant efforts have been made to develop automated detection methods. Manual fact-checking by specialized agencies, while valuable, struggles to keep pace with the rapid proliferation of misinformation on social media platforms. Consequently, deep learning-based approaches have emerged to enhance detection efficiency. Recent studies [2, 7, 21, 22, 24, 10] have focused on learning joint latent embeddings to assess the consistency between images and captions to identify OOC misinformation. Additionally, CCD [4] sought to mitigate biases introduced by single-modality analysis, adopting a causal perspective to improve prediction accuracy. These advancements underscore the need for continued exploration of robust detection methods to counter the evolving nature of misinformation. However, these methods neglect external factual verification critical for real-world claims (e.g., validating if an image truly matches a stated location). Recent advancements integrate external knowledge through web-retrieved evidence [26] and multimodal large language models [34], with frameworks like LLM-Consensus [18] further leveraging debate mechanisms to synthesize verification insights. These developments highlight the growing emphasis on combining intrinsic multimodal analysis with external knowledge retrieval for comprehensive OOC detection.

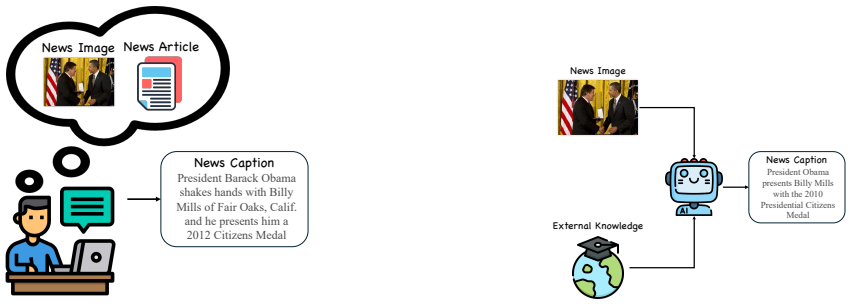
However, these approaches suffer from critical limitations: they ignore the deep semantic consistency between cross-modal content and fail to adequately leverage the rich semantic information embedded within news images. Instead of exploring the inherent semantic information, existing methods [26] rely primarily on surface-level feature matching, missing the opportunity to unlock deeper visual semantics that could provide more robust misinformation detection.

Inspired by the editorial process, where editors reveal deep semantic information within images through careful alignment with original articles as support, we propose Knowledge-enhanced **C**aption **R**econstructor and **C**omparison (CRC) to explicitly surface visual semantics by reconstructing semantically aligned image captions through VLMs to bridge the gap between cross-modal and single-modal semantics comparison. Specifically, our approach: (1) retrieve textual knowledge from the web using news images and visual knowledge along with its corresponding captions from the web using news captions; (2) filters the most relevant knowledge by semantic similarities; (3) perform instruction-tuning Vision-Language Models (VLMs) to generate context-aligned captions using news images augmented with filtered textual knowledge in VLMs prompts; (4) detects inconsistencies through separate visual/textual classifiers comparing news images against retrieved images and synthetic captions against news/retrieved captions, then fuses these consistency embeddings via dual-modal attention module for final predictions. By reconstructing captions that capture deep image semantics, our method effectively identifies OOC misinformation while surfacing visual semantics to provide interpretable results through the generated captions. Comprehensive experiments on NewsCLIPpings dataset [25] demonstrate the effectiveness of our approach.

Our contributions are summarized as follows:

- We propose CRC, a novel framework for detecting out-of-context (OOC) misinformation, pioneering the emulation of professional editorial workflows by leveraging Vision Language Models (VLMs) for caption reconstruction with external knowledge support.

- We develop a robust multi-modal, multi-tasks verdict prediction mechanism within CRC for OOC misinformation detection.



(a) Illustration of an editorial process

(b) Caption Reconstruction process in CRC

Figure 1: Illustration for editorial process and caption reconstruction process in CRC.

- Comprehensive experiments conducted on a large-scale dataset validate the efficacy of our framework, demonstrating its potential to enhance detection accuracy.

2. Related Work

2.1. Fake News Detection

Fake news detection represents a well-established challenge in natural language processing [5]. Early approaches primarily relied on textual information to identify false statements in news articles [20]. Stance-based models leveraged or emulated user opinions to infer news credibility [9]. CompareNet [6] introduced a textual fake news detection model based on graph neural networks, leveraging entity-level knowledge graph features to detect misinformation by extracting entities from the text.

In recent years, the proliferation of fake news incorporating multimodal information, such as images, videos, and audio, has shifted research focus toward multimodal fake news detection [23]. These approaches require both fact-checking and cross-modal consistency analysis capabilities. For instance, FKA-Owl[12] proposed a large vision-language model framework for detecting maliciously manipulated images and captions. Additionally, CCD[4] introduced a general causal framework to mitigate biases from single-modality analysis, addressing confounding factors in multimodal settings. These advancements provide a foundation for specialized subtasks, such as out-of-context misinformation detection.

2.2. Out-of-context Misinformation Detection

Out-of-context (OOC) misinformation detection, a subtask of multimodal fake news detection, focuses on identifying news pairs with authentic but mismatched images and captions. The NewsCLIPpings dataset [25], specifically curated for OOC misinformation detection, represents a highly challenging benchmark due to its emphasis on subtle inconsistencies. Several approaches have been proposed to address this challenge. For example, CCN [26] developed a cross-modal consistency checking network that utilizes evidence retrieved from image and text queries to assess alignment. SEN [35] employed a Stance Extraction Network to analyze the stance of retrieved evidence, evaluating consistency between images and captions.

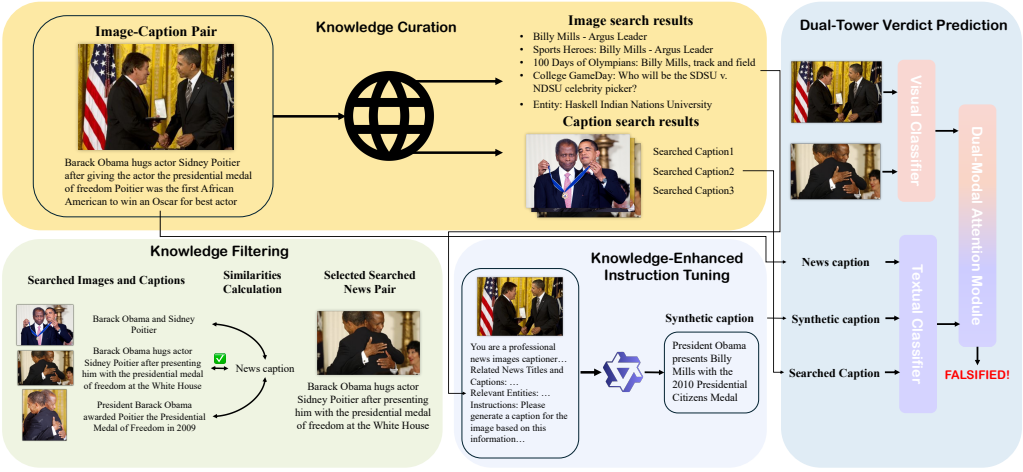


Figure 2: The overall framework of CRC. The image and caption in the news pair are used as query for textual knowledge and visual knowledge and its associated caption respectively. Then, we select the most relevant caption among captions retrived by the news caption as augmentation for both modalities. We further adapt the VLMs to news captioning domain via instruction tuning with textual knowledge as the prompt input. Finally, we perform verdict prediction with a dual-tower verdict prediction module to fully utilize the information in both modalities.

Zhang et al. [16] developed an interpretable misinformation detection framework by integrating neural-symbolic reasoning with large-scale multimodal models, enabling transparent analysis of false information. SNIFFER [34] introduced an OOC-misinformation-specific multimodal large language model (MLLM) that detects inconsistencies while also generating explanatory outputs. Furthermore, LLM-Consensus [18] developed a debating framework incorporating external knowledge to enhance OOC detection. Additionally, EXCLAIM[11] introduced an agent-based framework to harness the reasoning capabilities of large language models for improved consistency analysis.

Different from these approaches, our framework, CRC, incorporates the editorial process, as depicted in Figure 1(a), to facilitate OOC misinformation detection by reconstructing captions and integrating external knowledge.

3. Problem Formulation

Given an image-caption news pair consisting of an image and a corresponding caption that is associated with both the image and the news article, our objective is to determine whether the news pair is falsified or pristine.

As complementary information, we utilize additional news pairs associated with the caption as visual knowledge and captions from image search results as textual knowledge to enhance the semantic richness of each modality for consistency analysis.

Formally, we define the dataset as $D = \{(I_i, C_i, W_i^I, W_i^C, T_i, y_i)\}_{i=1}^N$, where:

- I_i is the news image sourced from a news article,
- C_i is the news caption associated with the news image,

- W_i^I is the images retrieved from web search results using C_i as a query (referred to as visual knowledge),
- W_i^C is the caption retrieved from web search results associated with W_i^I (referred to as retrieved captions),
- T_i is the textual knowledge retrieved from web-based image search using I_i as input (referred to as textual knowledge),
- $y_i \in \{0, 1\}$ is a binary label indicating whether the pair (I_i, C_i) is falsified ($y_i = 1$) or pristine ($y_i = 0$).

The task is formulated as a mapping: $(I_i, C_i, W_i^I, W_i^C, T_i) \rightarrow y_i$.

A pristine pair indicates that the caption C_i accurately describes the news image I_i in the context of the original news article. A falsified pair involves a news caption C_i that misrepresents a news image I_i, creating a false narrative. It is worth noting that both the caption and the image are pristine, with falsification arising solely from their inconsistency. The prediction process leverages $I_i, C_i, W_i^I, W_i^C, T_i$ to identify subtle inconsistencies between the image and caption, requiring robust analysis of their multimodal interactions. This formulation provides a foundation for developing methods to address the OOC misinformation detection challenge, as explored in subsequent sections.

4. The Proposed Approach

We propose a novel framework, CRC, for detecting out-of-context (OOC) misinformation by effectively integrating multimodal information from images, captions, and web-retrieved results. As illustrated in 2 the proposed framework consists of four key components: Knowledge Curation, Knowledge Filtering, Knowledge-Enhanced Instruction Tuning, and Dual-Tower Verdict Prediction.

4.1. Knowledge Curation

Out-of-context (OOC) misinformation is often challenging to detect, even for human evaluators, due to the high visual and semantic similarity between paired images and captions. To address this, we incorporate external knowledge to enrich the semantic content of each modality, providing additional context for cross-modal consistency analysis.

Following prior work [26], we obtain visual knowledge W^I and its associated retrieved captions W^C by performing web search using the original caption C as a query. This enhances the image modality by retrieving contextually relevant images and their descriptions. Conversely, we enrich the textual modality by conducting image-based web search using the news image I to retrieve relevant textual knowledge T.

We leverage publicly available data provided by Abdelnabi et al. [26] for the initial web search retrievals. To address cases of missing or incomplete textual knowledge, we additionally employ the open-source project PicImageSearch[1] to supplement the image-based caption retrieval. This enhancement ensures a more complete knowledge base for downstream reasoning and verdict prediction.

This cross-modal knowledge curation provides a more comprehensive understanding of the news pair and serves as a foundation for subsequent modules.

[1] `https://github.com/kitUIN/PicImageSearch`

4.2. Knowledge Filtering

The curated visual knowledge and retrieved caption may contain irrelevant information that can degrade the model's performance. To address this, we propose a similarity-based filtering method to remove such irrelevant knowledge.

Specifically, for all web-searched visual knowledge W_i^I and retrieved captions W_i^C queried by the caption C_i, we first encode the query C_i and all retrieved captions W_i^C using a sentence-transformer [33]. We then compute the cosine similarity between the encoded query and each retrieved caption:

$$j^* = \arg\max_j \cos\left(E^{C_i}, E^{W_{i,j}^C}\right)$$

where E^{C_i} and $E^{W_{i,j}^C}$ denote the embeddings of the query and the j-th retrieved caption, respectively. Finally, we select the visual knowledge and the retrieved caption pair W_{i,j^*}^I and W_{i,j^*}^C with the highest similarity score as filtered training data.

By applying this knowledge filtering, we ensure contextual consistency in the augmented data, which improves both the accuracy and robustness of the detection model.

4.3. Knowledge-Enhanced Instruction Tuning

We employ Vision Language Models (VLMs) to generate synthetic captions as an augmentation for the dataset. However, pretrained VLMs such as the Intern-VL series [13, 14, 15] or Qwen-VL series [29, 30, 31] often fail to accurately reconstruct captions for images due to their limited access to background knowledge. For instance, VLMs may struggle to recognize ordinary individuals appearing in news images. To address this limitation, we provide image-related external knowledge, including potential entities and related news titles, to assist VLMs in producing more accurate captions.

To better align the synthetic caption dataset with the domain of the original captions, we perform instruction tuning on the VLMs.

For instruction tuning, we carefully design a prompt template, presented in Appendix A, which takes the image and retrieved external knowledge as input and the matched caption as output. Specifically, we first select all news pairs with label $y_i = 0$ from the training set. Then, using the image I_i and image search results T_i (retrieved by querying with I_i) as input, we populate the prompt template. The VLM generates a synthetic caption based on this prompt. Through this instruction tuning process, VLMs learn to emulate expert editors by reconstructing image captions accurately, minimizing hallucinations and irrelevant information.

We select Qwen2.5-VL-7B [31] as the base model and employ the Low-Rank Adaptation (LoRA) [19] parameter-efficient fine-tuning (PEFT) [27] technique to fine-tune the vision-language models (VLMs). This approach effectively reduces training costs while preserving the model's generalization ability and its capability to integrate original information.

Following instruction tuning, we generate synthetic captions C_i' for each news pair in the training set using I_i and T_i as inputs for the VLMs. These synthetic captions serve as references to the original captions.

4.4. Dual-Tower Verdict Prediction

To comprehensively capture and analyze the multimodal information, we propose a dual-tower verdict prediction approach. The prediction model is composed of three key components: textual classification, visual classification, and the final verdict prediction module. This architecture enables the model to separately process textual and visual consistency prediction before combining them to make a robust decision regarding the authenticity of the news pair.

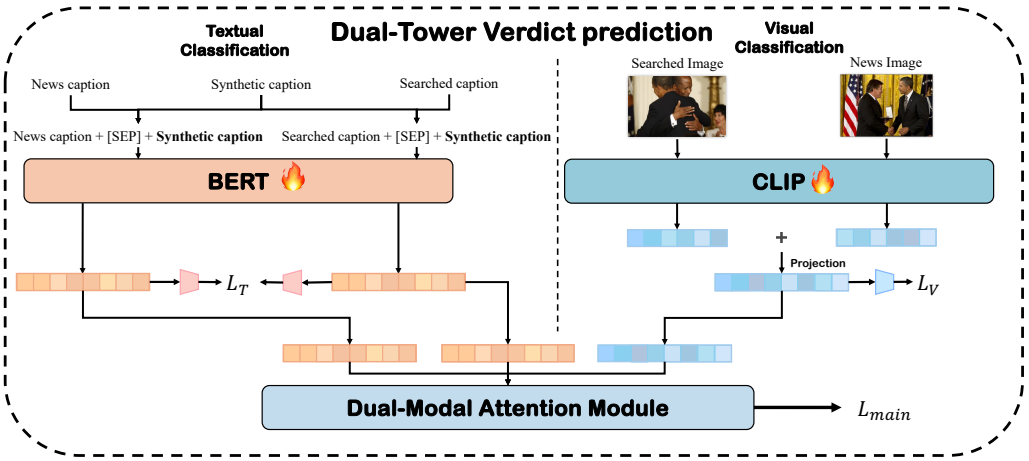


Figure 3: The framework of Dual-Tower Verdict Prediction module. It is split into two parts: textual classification and visual classificaition, and finally aggregated by dual-modal attention module to generate the final prediction.

4.4.1. Textual Classification

The textual classification component aims to evaluate the semantic alignment between the original caption C_i and the synthetically generated caption C'_i. For this purpose, we adopt BERT [3] as the foundational encoder, fine-tuned via the LoRA [19] parameter-efficient fine-tuning (PEFT) technique. We opt for BERT over Sentence-BERT [33] because of the significant entity overlap commonly present in both captions; while Sentence-BERT excels at capturing sentence-level semantic similarity, BERT is better suited for nuanced consistency checks where entity-level distinctions are crucial, particularly in the context of news captions.

Our method concatenates the original caption C_i and the synthetic caption C'_i using the special separator token [SEP], then encodes this combined input through the BERT encoder to obtain a joint embedding representation $E_i^{Concat} \in \mathbb{R}^{d_t}$ then use a Linear layer to project the classification embedding to $H_i^C \in \mathbb{R}^d$:

$$H_i^C = \text{Linear}(\text{BERT}(Concat(C_i, C'_i))$$

Subsequently, a fully connected layer maps H_i^C to a scalar classification score:

$$s_i^C = W^{BERT} H_i^C + b^{BERT}$$

where W^{BERT} and b^{BERT} denote the weights and bias parameters of the classification layer.

In addition to the original caption, we incorporate web-retrieved captions W_i^C queried by C_i , which can provide valuable context to enhance the semantic consistency evaluation. Similar to the above procedure, we concatenate W_i^C and the synthetic caption C'_i, encode the resulting sequence into embedding $E_i'^{Concat} \in \mathbb{R}^{d_t}$ then project $E_i'^{Concat}$ to $H_i^W \in \mathbb{R}^d$, finally calculate the classification score s_i^W:

$$H_i^W = \text{Linear}(\text{BERT}(Concat(W_i^C, C'_i))$$

$$s_i^W = W^{BERT} H_i^W + b^{BERT}$$

4.4.2. Visual Classification

The visual classification module aims to assess the semantic consistency between the original image I_i and the retrieved visual knowledge W_i^I, which is obtained via search based on the caption C_i. This design enables the model to leverage external visual context to better verify the authenticity of the image-caption pair.

Specifically, we use CLIP [17] to separately encode the original image I_i and the retrieved visual knowledge W_i^I, producing embeddings $I_i^{CLIP} \in \mathbb{R}^d$ and $W_i^{I,CLIP} \in \mathbb{R}^d$, respectively. These embeddings are then concatenated to form a joint representation:

$$CLS_i^{CLIP} = [I_i^{CLIP}; W_i^{I,CLIP}] \in \mathbb{R}^{2d}$$

To ensure compatibility with other model components, a linear projection layer maps this concatenated vector to a feature embedding:

$$H_i^{CLIP} = \text{Linear}(CLS_i^{CLIP})$$

where $H_i^{CLIP} \in \mathbb{R}^d$; W^{CLIP} and b^{CLIP} denote the weights and bias of the linear transformation.

To effectively capture the semantic relationship between the original image and the retrieved visual knowledge, we adopt CLIP [17] as the base encoder due to its demonstrated capability in learning joint image-text embeddings. CLIP's pretraining on large-scale image-caption pairs enables it to produce rich, semantically aligned visual features that are well-suited for cross-modal consistency verification. By leveraging CLIP's robust encoding, our model benefits from enhanced discriminative power in detecting subtle discrepancies between the original image and contextually related retrieved images, which is critical in out-of-context misinformation detection.

Finally, a fully connected layer computes the visual classification score indicating the semantic consistency between the original image and the retrieved visual knowledge:

$$s_i^{CLIP} = W^{CLIP} H_i^{CLIP} + b^{CLIP}$$

This score helps determine whether the external visual context supports or contradicts the original image content, contributing to the overall misinformation detection.

4.4.3. Verdict Prediction

The final verdict prediction module integrates the classification embeddings and scores obtained from the news captions classification, retrieved captions classification, and visual classification models. This module employs a transformer encoder layer, referred to as the **Dual-Modal Attention Module**, which takes as input the three classification embeddings H_i^C, H_i^W, and H_i^{CLIP}, with a learnable classification token prepended to the sequence.

Formally, the input sequence is:

$$Input = [[CLS]; H_i^C; H_i^W; H_i^{CLIP}]$$

The transformer encoder processes this sequence to produce fused embeddings:

$$H_i^V = \text{Dual-Modal Attention Module}(Input)$$

where $H_{i,0}^V$ corresponds to the output embedding of the classification token. Finally, a fully connected linear layer maps this embedding to the final verdict prediction score:

$$s_i^C = W^{VP} H_i^V + b^{VP}$$

The Dual-Modal Attention Module effectively captures the interdependencies among the multimodal classification embeddings, enabling a comprehensive and robust decision on the consistency and authenticity of the image-caption pair.

4.5. Model Training and Optimization

Our model is trained using a multi-task loss function that integrates the classification objectives of individual modules. Specifically, we independently apply cross-entropy loss to the textual classification, visual classification, and the final dual-modal attention module, defined as:

$$\mathcal{L} = \mathcal{L}_{text} + \mathcal{L}_{visual} + \mathcal{L}_{verdict}$$

where $\mathcal{L}$ represents the total loss, $\mathcal{L}_{text}$, $\mathcal{L}_{visual}$, and $\mathcal{L}_{verdict}$ denote the cross-entropy losses for the textual classification, visual classification, and verdict prediction, respectively.

This multi-task supervision encourages each module to learn complementary representations while jointly optimizing the final misinformation detection performance.

5. Experiments and Analysis

In this section, we present a comprehensive evaluation of our proposed method. We begin by introducing the evaluation datasets and experimental setup, followed by a detailed comparison with the baselines. We then provide in-depth analyses through ablation studies, performance assessment under various data constraints, and illustrative case studies to demonstrate the effectiveness of our approach.

5.1. Experiment Setup

5.1.1. Dataset

We conducted our experiments on the NewsCLIPpings dataset [25], widely recognized as the most challenging testbed for evaluating out-of-context misinformation detection systems. This comprehensive dataset encompasses four distinct scenarios where authentic news captions and images are deliberately mismatched to create deceptive content. The scenarios are designed as follows:

- Semantics/CLIP Text-Image: Identifies captions with highest cosine similarity to images via CLIP [17] embeddings and pairs them as misleading matches.
- Semantics/CLIP Text-Text: Calculates similarity between caption pairs, then selects highly similar but unrelated image-caption pairs. Both semantic scenarios create content that can deceive human experts.
- Person/SBERT-WK Text-Text: Creates deceptive news by mismatching images and captions with the same person entity but lowest SBERT-WK [32] similarity scores, fabricating false event associations.
- Scene/ResNet-Place: Pairs images without people based on highest ResNet-Place [28] scene embedding similarity, creating misleading associations through visually similar backgrounds.

We evaluate our proposed method using the merged balanced split that combines equal proportions of samples from each scenario. The resulting dataset comprises 71,072 training samples, 7,024 validation samples, and 7,264 testing samples, ensuring comprehensive representation across all misinformation types.

For the knowledge retrieval component, we utilize search results denoted as W_i^I, W_i^C, and T_i from the curated dataset compiled by CCN [26]. Upon analysis, we identified gaps in the image search results coverage. To address this limitation, we conducted supplementary image searches using the open-source PicImageSearch library[2] with Bing Image Search[3] as our primary search engine, thereby enhancing the comprehensiveness of our evidence repository.

5.1.2. Experimental Setup

Our implementation architecture comprises several key components with carefully selected foundation models. For the instruction tuning module, we employ Qwen2.5-VL-7B [31] as our base model, leveraging its multimodal capabilities while applying the parameter-efficient LoRA [19] technique for fine-tuning to maintain computational efficiency. For the classification components, we utilize the BERT-base architecture for textual classification tasks and the CLIP-base model for visual modality classification.

The framework is implemented using PyTorch and the Transformers library. All training procedures were executed on a computing server equipped with dual NVIDIA-4080-Super GPUs. We optimize the models using Adam with a learning rate of 8e-5 and process batches of 80 samples to balance computational constraints with training stability.

5.2. Experimental Results

5.2.1. Performance Comparison

We evaluated our proposed method CRC against several established baseline approaches in the field of multimodal misinformation detection:

1) EANN [8] utilizes adversarial training to analyze factual content and extract event-level features through a dual-branch architecture. It combines modality-specific representations using hierarchical fusion techniques for effective multimodal detection. 2) VisualBERT [37] represents a foundational multimodal pre-training framework that concatenates region-based visual features with textual embeddings as input to a unified Transformer encoder. This approach employs masked language modeling and image-text matching objectives to develop cross-modal alignments within a shared embedding space. 3) CLIP [17] serves as a robust pretrained model specifically designed for image-caption learning. We also incorporate CLIP as the foundation model for our visual modality classification component. 4) LORGAN [39] detects phrase-level inconsistencies through latent logic variables and logical rule aggregation for explainable predictions. We use the model with CLIP as the backbone as the baseline. 4) DT-Transformer [36] builds upon CLIP's multimodal architecture by implementing a hierarchical attention mechanism and introducing auxiliary refinement modules to enhance cross-modal interaction. This method utilizes variational multi-source generated datasets as training data to improve modality alignment. 5) GPT-4o[38]: Powerful MLLM developed by OpenAI, we evaluate its ability in a zero-shot manner. 6) CCN [26] introduces a cross-modal consistency checking network that leverages evidence retrieved through both image and text searches. The model generates verdict predictions by evaluating the consistency between input samples and the retrieved evidence.

[2]https://github.com/kitUIN/PicImageSearch
[3]https://www.bing.com

Method	All	Fake	Real
EANN	58.1	61.8	56.2
VisualBERT	58.6	38.9	78.4
CLIP	66.0	64.3	67.7
CLIP-LORGAN	70.5	-	-
DT-Transformer	77.1	78.6	75.6
GPT-4o	73.8	75.5	73.4
CCN	83.9	82.8	84.9
CRC *(Ours)*	**85.2**	**85.5**	**85.1**

Table 1: Comparative study on CRC and baselines.(%)

Table 1 presents a comprehensive comparison between CRC and established baseline approaches. As evident from the results, CRC consistently outperforms all competing methods, achieving an absolute accuracy improvement of 1.2% over the second-best performer. Notably, CRC and CCN are the only methodologies that leverage web search results as additional knowledge sources, and correspondingly, they are the only two approaches that exceed the 80% accuracy threshold. This clear performance distinction underscores the critical importance of incorporating external world knowledge into the detection framework for effectively identifying out-of-context misinformation.

5.2.2. Ablation Studies

To assess the contribution of individual components within our CRC framework, we conducted a series of systematic ablation experiments. Our evaluation focused on isolating the impact of textual classification (original vs. synthetic captions), enhanced textual classification (retrieved captions vs. synthetic captions), visual classification (original news images vs. web-retrieved images), and our multi-task learning approach. The results are presented in Table 2.

Table 2: Ablation study of CRC's components. The checkmark indicates that the corresponding component is used. Textual: Classification between original and synthetic captions; Retrieved Captions: Classification between retrieved and synthetic captions; Visual: Classification between searched images and original images; Multi-tasks: Apply loss on text and visual detection respectively; Tuning: Apply Instruction Tuning on VLMs.

Textual	Retrieved Captions	Visual	Multi-tasks	Tuning	All	Fake	Real
✓	✗	✗	✓	✓	84.9	**88.0**	81.8
✗	✓	✗	✓	✓	79.3	82.2	76.4
✗	✓	✓	✓	✓	79.4	87.7	71.1
✓	✓	✗	✓	✓	85.0	86.8	83.2
✓	✗	✓	✓	✓	84.2	86.2	82.2
✓	✓	✓	✗	✓	84.3	85.7	82.9
✓	✓	✓	✓	✗	83.6	85.1	81.8
✓	✓	✓	✓	✓	**85.2**	85.5	**85.1**

The analysis of Table 2 reveals several key insights. The textual classification comparing original and synthetic captions serves as a robust foundation (Row 1), achieving 84.9% accuracy. Integrating web-retrieved textual knowledge (Row 4) further improves performance to 85.0%, demonstrating the value of external context. The visual classification module acts as an effective complementary component; adding it to the comprehensive textual analysis (Row 8 vs. Row

4) elevates the overall accuracy to its peak of 85.2% and significantly boosts ”Real” sample detection from 83.2% to 85.1%. This underscores that a synergistic combination of multimodal signals is most effective.

Furthermore, two components are critical for optimal performance. First, instruction tuning proves essential, contributing a significant 1.6% increase in accuracy when comparing the full model with and without it (Row 8 vs. Row 7). Second, the multi-task learning strategy provides a clear advantage, with the full model (Row 8, 85.2%) outperforming an identical setup without it (Row 6, 84.3%). These findings confirm that while multi-faceted textual analysis is central, the integration of visual modality, instruction tuning, and a multi-task objective are all key to CRC's state-of-the-art performance.

To further quantify the role of external knowledge, we evaluated a variant of CRC without any web-retrieved textual or visual knowledge during inference. As shown in Table 3, removing this knowledge leads to a 3.6% drop in accuracy. This highlights that external knowledge provides crucial context, such as names and locations that the VLM may not inherently know, thereby improving the quality of synthetic captions and the final prediction.

Table 3: Evaluation of the impact of external knowledge on CRC's performance.

Method	All	Fake	Real
CRC (w/o Knowledge)	81.6	84.0	79.2
CRC (w Knowledge, Top-k)	71.7	82.4	61.0
CRC (w Knowledge)	85.2	85.5	85.1

5.2.3. Performance under Data Constraints

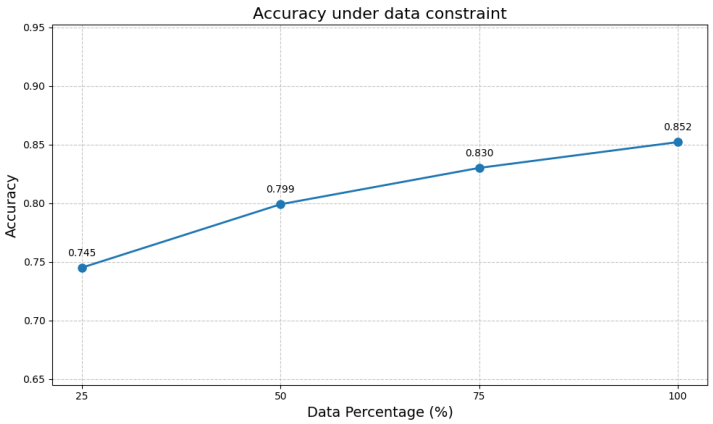


Figure 4: Performance of CRC under different training data constraints.

To assess the impact of training data volume, we evaluated CRC's performance using varying percentages of the training set. As shown in Figure 4, model accuracy scales positively with data volume. Performance steadily increases from 74.5% (with 25% of the data) to 79.9% (50%), 83.0% (75%), and reaches its peak of 85.2% on the full dataset. This trend underscores the model's ability to effectively leverage larger datasets, while also showing that it can achieve reasonable performance even with limited data.

5.2.4. *Quantitative Evaluation of Captions*

To demonstrate the effectiveness of our key components on the quality of the generated captions, we conducted a quantitative evaluation using ROUGE-L [41] and BLEU-4 [40]. We compared the synthetic captions generated by our full model against those from ablated versions. The results in Table 4 show that the full CRC framework, with both instruction tuning and external knowledge, produces captions that are semantically closer to the ground truth.

Table 4: Quantitative evaluation of synthetic caption quality using ROUGE-L and BLEU-4 scores. Higher is better.

Method	ROUGE-L	BLEU-4
CRC (w/o Tuning)	0.19	0.03
CRC (w/o Knowledge)	0.20	0.04
CRC (Full Model)	**0.26**	**0.10**

5.2.5. *Case Study*

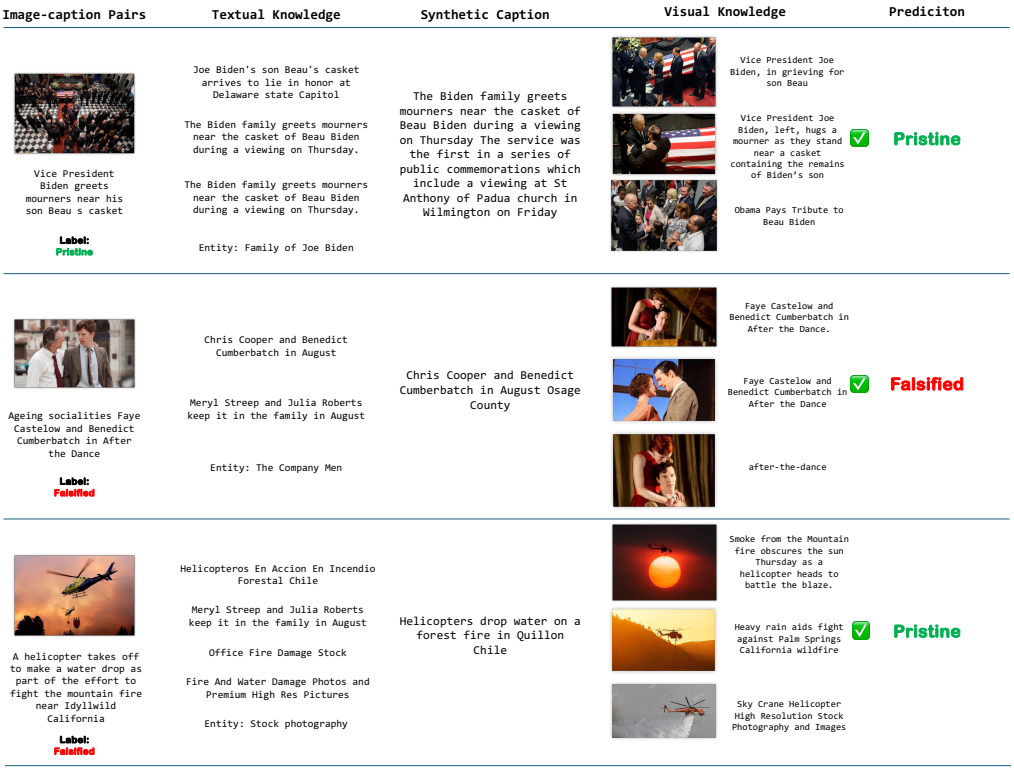


Figure 5: Prediction samples of CRC.

Figure 5 presents qualitative examples from the inference process. CRC demonstrates a strong ability to generate accurate synthetic captions, providing a robust reference against the original caption. As shown in the first two examples, with the assistance of retrieved textual

knowledge, the VLM acquires crucial context (e.g., identities, events) and produces high-quality captions comparable to those written by human editors. This, in turn, aids the verdict prediction module. However, the third example highlights a potential limitation: in generic scenes where the original and synthetic captions are highly similar, the verdict module may struggle. Despite the VLM's accurate captioning, the module can overlook subtle semantic discrepancies, potentially leading to a prediction error.

6. Limitation

Our study has two primary limitations. First, concerning **efficiency and computational cost**, our CRC framework involves a multi-step process including web retrieval, VLM inference, and multiple classification tasks. This design, while effective for accuracy, is computationally intensive and may pose challenges for real-time, large-scale deployment. A trade-off exists between the depth of analysis and processing speed.

Second, regarding **generalization and dataset diversity**, our experiments were conducted exclusively on the NewsCLIPpings dataset. While this is a challenging benchmark for journalistic out-of-context misinformation, the characteristics of misinformation can vary significantly across other domains, such as social media, advertisements, or scientific articles. Therefore, the framework's performance and universality on these other types of content remain to be validated.

7. Conclusion

In this paper, we propose CRC, an innovative framework that emulates the editorial process to detect out-of-context misinformation. Our method curates and filters pivotal knowledge from the web, then leverages instruction-tuned Vision Language Models (VLMs) to reconstruct a semantically faithful synthetic caption for the news image. Subsequently, a dual-tower architecture performs a multi-faceted comparison: a textual module compares the synthetic caption against both the original and retrieved captions, while a visual module assesses consistency with retrieved images. Finally, these signals are fused for a final verdict. Extensive experiments on a large-scale dataset validate the effectiveness of our approach, setting a new state-of-the-art. Future work will focus on improving computational efficiency and validating the framework's generalization across diverse domains.

Author Contributions

Jie Li proposed the core methodology framework and critically revised the final manuscript; Yuqiu Deng designed and performed all experiments, analyzed data, and drafted the manuscript; Simin Yang and Jinrui Wang refined the text structure. All authors approved the final version.

References

[1] Li, X., Qiao, J., Yin, S., Wu, L., Gao, C., Wang, Z. & Li, X. A Survey of Multimodal Fake News Detection: A Cross-Modal Interaction Perspective. *IEEE Transactions On Emerging Topics In Computational Intelligence*. (2025)

[2] Jaiswal, A., Wu, Y., AbdAlmageed, W., Masi, I. & Natarajan, P. AIRD: Adversarial Learning Framework for Image Repurposing Detection.. *Computer Vision And Pattern Recognition (CVPR)*. pp. 11330-11339 (2019)

[3] Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings Of The 2019 Conference Of The North*. pp. 4171-4186 (2019)

[4] Chen, Z., Hu, L., Li, W., Shao, Y. & Nie, L. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection.. *Annual Meeting Of The Association For Computational Linguistics (ACL)*. pp. 627-638 (2023)
[5] Chen, C. & Shu, K. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*. **abs/2311.05656** (2024)
[6] Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N. & Zhou, M. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge. *Proceedings Of The 59th Annual Meeting Of The Association For Computational Linguistics And The 11th International Joint Conference On Natural Language Processing (Volume 1: Long Papers)*. (2021)
[7] Aneja, S., Bregler, C. & Nießner, M. COSMOS: Catching Out-of-Context Misinformation with Self-Supervised Learning. *ArXiv.org*. **abs/2101.06278** (2021)
[8] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L. & Gao, J. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection.. *ACM SIGKDD Conference On Knowledge Discovery And Data Mining (KDD)*. pp. 849-857 (2018)
[9] Yao, B., Shah, A., Sun, L., Cho, J. & Huang, L. End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models. *Proceedings Of The 46th International ACM SIGIR Conference On Research And Development In Information Retrieval*. pp. 2733-2743 (2023)
[10] Zhu, Y., Li, Y., Wang, J., Gao, M. & Wei, J. FaKnow: A Unified Library for Fake News Detection. *Data Intelligence*. (2025)
[11] Wu, Y., Zhang, Z., Wang, F., Luo, Y., Xiong, H. & Tang, N. EXCLAIM: An Explainable Cross-Modal Agentic System for Misinformation Detection with Hierarchical Retrieval. (2025)
[12] Liu, X., Li, P., Huang, H., Li, Z., Cui, X., Liang, J., Qin, L., Deng, W. & He, Z. FKA-Owl: Advancing Multimodal Fake News Detection through Knowledge-Augmented LVLMs.. *ACM International Conference On Multimedia (MM)*. pp. 10154-10163 (2024)
[13] Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., Ma, J., Wang, J., Dong, X., Yan, H., Guo, H., He, C., Shi, B., Jin, Z., Xu, C., Wang, B., Wei, X., Li, W., Zhang, W., Zhang, B., Cai, P., Wen, L., Yan, X., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J. & Wang, W. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*. **67** (2024)
[14] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y. & Dai, J. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. (2024)
[15] Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., Gao, Z., Cui, E., Cao, Y., Liu, Y., Wang, H., Xu, W., Li, H., Wang, J., Lv, H., Chen, D., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J. & Wang, W. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. (2025)
[16] Zhang, Y., Trinh, L., Cao, D., Cui, Z. & Liu, Y. Interpretable Detection of Out-of-Context Misinformation with Neural-Symbolic-Enhanced Large Multimodal Model. *ArXiv*. (2023)
[17] Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. *International Conference On Machine Learning*. pp. 8748-8763 (2021)
[18] Lakara, K., Channing, G., Sock, J., Rupprecht, C., Torr, P., Collomosse, J. & Witt, C. LLM-Consensus: Multi-Agent Debate for Visual Misinformation Detection. (2024)
[19] Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference On Learning Representations*. (2022)
[20] Su, Q., Wan, M., Liu, X. & Huang, C. Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective. *Natural Language Processing Research*. **1**, 1 (2020)
[21] Jaiswal, A., Sabir, E., Abd-Almageed, W. & Natarajan, P. Multimedia Semantic Integrity Assessment Using Joint Embedding Of Images And Text.. *ACM International Conference On Multimedia (MM)*. pp. 1465-1471 (2017)
[22] Müller-Budack, E., Theiner, J., Diering, S., Idahl, M. & Ewerth, R. Multimodal Analytics for Real-world News using Measures of Cross-modal Entity Consistency.. *International Conference On Multimedia Retrieval (ICMR)*. pp. 16-25 (2020)
[23] Comito, C., Caroprese, L. & Zumpano, E. Multimodal fake news detection on social media: a survey of deep learning techniques. *Social Network Analysis And Mining*. **13** (2023)
[24] Hu, L., Zhao, Z., Qi, W., Song, X. & Nie, L. Multimodal matching-aware co-attention networks with mutual knowledge distillation for fake news detection.. *Information Sciences*. **664** pp. 120310 (2024)
[25] Luo, G., Darrell, T. & Rohrbach, A. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. (2021)
[26] Abdelnabi, S., Hasan, R. & Fritz, M. Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context

Images via Online Resources.. *Computer Vision And Pattern Recognition (CVPR)*. pp. 14920-14929 (2022)
[27] Han, Z., Gao, C., Liu, J., Zhang, J. & Zhang, S. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *Transactions On Machine Learning Research*. **abs/2403.14608** (2024)
[28] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. & Torralba, A. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **40**, 1452-1464 (2018)
[29] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C. & Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. (2023)
[30] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J. & Lin, J. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *ArXiv.org*. **abs/2409.12191** (2024)
[31] Yang, Q., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z., Quan, S. & Wang, Z. Qwen2.5 Technical Report. *ArXiv.org*. **abs/2412.15115** (2024)
[32] Wang, B. & Kuo, C. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models.. *IEEE/ACM Transactions On Audio, Speech And Language Processing*. **28** pp. 2146-2157 (2020)
[33] Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.. *Conference On Empirical Methods In Natural Language Processing (EMNLP)*. pp. 3980-3990 (2019)
[34] Qi, P., Yan, Z., Hsu, W. & Lee, M. Sniffer: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection.. *Computer Vision And Pattern Recognition (CVPR)*. pp. 13052-13062 (2024)
[35] Yuan, X., Guo, J., Qiu, W., Huang, Z. & Li, S. Support or Refute: Analyzing the Stance of Evidence to Detect Out-of-Context Mis- and Disinformation. *Proceedings Of The 2023 Conference On Empirical Methods In Natural Language Processing*. pp. 4268-4280 (2023)
[36] Papadopoulos, S., Koutlis, C., Papadopoulos, S. & Petrantonakis, P. Synthetic Misinformers: Generating and Combating Multimodal Misinformation.. *International Conference On Multimedia Retrieval (ICMR)*. pp. 36-44 (2023)
[37] Li, L., Yatskar, M., Yin, D., Hsieh, C. & Chang, K. VisualBERT: A Simple and Performant Baseline for Vision and Language. *ArXiv.org*. **abs/1908.03557** (2019)
[38] OpenAI. GPT-4o System Card. *ArXiv.org*. **abs/2410.21276** (2024)
[39] Ma, H., Zhang, J., Liu, Q., Wu, S. & Wang, L. INTERPRETABLE MULTIMODAL OUT-OF-CONTEXT DETECTION WITH SOFT LOGIC REGULARIZATION. *ArXiv*. (2024)
[40] Papineni, K., Roukos, S., Ward, T. & Zhu, W. Bleu: a Method for Automatic Evaluation of Machine Translation. *ACL*. (2002)
[41] CY, Lin. Rouge: A package for automatic evaluation of summaries *ACL*. 2004

Author Biography

Appendix A. The Prompt Template of VLMs

```
# Task: Image Captioning

You are a professional news images captioner. Your task is
to caption the image based on the given complementary
material.
The image is from a news article. Below is the
complementary material with related news titles and
possible entities:

## Related News Titles and Captions

{related_news_titles_and_captions}

## Matched Content

{matched_content}

## Relevant Entities

{relevant_entities}

## Best Guess

{best_guess_label}

# Instructions

Please generate a caption for the image based on this
information. Your caption should be concise, factual, and
journalistically appropriate.
```

Figure A.6: The prompt template of VLMs.