# Natural Language-centered Inference Network for Multi-modal Fake News Detection

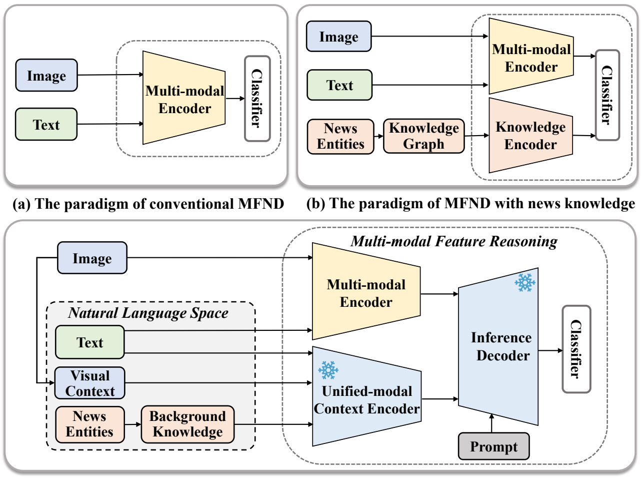
**Qiang Zhang**, **Jiawei Liu**[*], **Fanrui Zhang**, **Jingyi Xie** and **Zheng-Jun Zha**
University of Science and Technology of China, China
{zq_126, zfr888, hsfzxjy}@mail.ustc.edu.cn, {jwliu6, zhazj}@ustc.edu.cn

## Abstract

The proliferation of fake news with image and text in the internet has triggered widespread concern. Existing research has made important contributions in cross-modal information interaction and fusion, but fails to fundamentally address the modality gap among news image, text, and news-related external knowledge representations. In this paper, we propose a novel Natural Language-centered Inference Network (NLIN) for multi-modal fake news detection by aligning multi-modal news content with the natural language space and introducing an encoder-decoder architecture to fully comprehend the news in-context. Specifically, we first unify multi-modal news content into textual modality by converting news images and news-related external knowledge into plain textual content. Then, we design a multi-modal feature reasoning module, which consists of a multi-modal encoder, a unified-modal context encoder and an inference decoder with prompt phrase. This framework not only fully extracts the latent representation of cross-modal news content, but also utilizes the prompt phrase to stimulate the powerful in-context learning ability of the pre-trained large language model to reason about the truthfulness of the news content. In addition, to support the research in the field of multi-modal fake news detection, we produce a challenging large scale, multi-platform, multi-domain multi-modal Chinese Fake News Detection (CFND) dataset. Extensive experiments show that our CFND dataset is challenging and the proposed NLIN outperforms state-of-the-art methods.

## 1 Introduction

Fake news is spreading online at an ever-increasing rate, posing a serious challenge to the credibility of news media platforms [Abdelnabi *et al.*, 2022; Liu *et al.*, 2023]. At the same time, the widespread dissemination of fake news may cause mass panic and social instability, *e.g.*, some unscrupulous individuals have exploited fake news to mislead health care practices, undermine the credibility of governments and interfere with presidential elections [Narayan *et al.*, 2022; Zhang *et al.*, 2023a]. In addition, with the swift development of multimedia technology, fake news makers are increasingly turning more and more to multi-modal content, such as attractive images, to capture and mislead the general public, thus making fabricated stories more credible and disseminating them more speedily. Therefore, in order to reduce the harmful effects of fake news dissemination, automatic Multi-modal Fake News Detection (MFND) has received more and more research attention.



(a) The paradigm of conventional MFND (b) The paradigm of MFND with news knowledge

(c) The proposed paradigm

Figure 1: Comparison with previous paradigms. (a) Traditional MFND follows the paradigm of fusing news image and text features by designing multi-modal fusion encoder, then performing fake news detection. (b) The MFND with news knowledge paradigm employs news entities represented in the knowledge graph as news background knowledge and fuses them into the cross-modal space. (c) The proposed paradigm aims to unify news content with background knowledge into the natural language space and reason about the authenticity of news through an encoder-decoder architecture equipped with prompt phrases.

To address this problem, a series of multi-modal fake news detectors have been proposed to explore cross-modal information interaction and fusion for identifying anomalies in fake news. Most of the previous work follows the paradigm of fusing news image and text features by multi-

[*] Corresponding author

modal encoder and subsequently performing fake news detection (as shown in Figure 1 (a)). These approaches simply utilize concatenation operations [Singhal *et al.*, 2019a; Wang *et al.*, 2021], attention mechanism [Jin *et al.*, 2017; Zhou *et al.*, 2022] or auxiliary tasks [Chen *et al.*, 2022a; Khattar *et al.*, 2019] to capture the underlying semantic correlations between image and text features. However, these methods fail to obtain persuasive and interpretable results as they lack the link to external facts. Therefore, in order to verify the authenticity of news at the knowledge level, a few of approaches excavate background knowledge information about the news and serve as a source of objective evidence by extracting news entities and linking them to the knowledge graph, which is structured as shown in Figure 1 (b). For example, some of them consider news entities and their contexts as external knowledge and utilize attention mechanisms to assess the weights of news entity representations [Zhang *et al.*, 2023b; Tseng *et al.*, 2022] or discover inconsistent semantic information at the knowledge level [Sun *et al.*, 2021]. However, these aforementioned methods neglect or do not fundamentally address the intractable problem of modality gap among image, text, and knowledge representations of news, resulting in the model's inability to adequately extract cross-modal correlation features of news content. Therefore, in order to overcome this dilemma, a potentially effective solution is to unify multi-modal news content into textual modality and utilize the powerful in-context learning ability and rich implicit knowledge of large language model (LLM) to extract the news cross-modal correlation features and reason about the authenticity of news content.

Although fake news crosses geographical and linguistic boundaries, most of the work and datasets available in the field are concentrated in the English domain [Boididou *et al.*, 2018; Zubiaga *et al.*, 2017], with limited content in other language domains. Taking the Chinese domain as an example, the existing multi-modal fake news detection datasets are only Weibo [Jin *et al.*, 2017] and Weibo-21 [Nan *et al.*, 2021], and their data sources are often limited to one platform on the web, which makes the representation of fake news often more homogeneous. At the same time, their data size is restricted, and the news content is more outdated, which is not conducive to detecting the current news content, and bring certain limitations for the development of Chinese fake news detection. Therefore, a large-scale, multi-platform, multi-domain Chinese fake news detection dataset is urgently needed.

In order to support the research in the field of multi-modal fake news detection, we produce a challenging multi-modal Chinese Fake News Detection (CFND) dataset, which is collected from multiple platforms, contains news from multiple domains. And CFND consists of 26,665 news samples, significantly larger than previous Chinese datasets. Additionally, we propose a novel Natural Language-centered Inference Network (NLIN) for multi-modal fake news detection, whose general structure is shown in Figure 1 (c). This detection paradigm is mainly divided into three phases: (1) News pre-processing phase, we extract the visual context using three approaches, namely, image caption, dense labeling and OCR, so that we could achieve image-to-text conversion with minimal information loss. In terms of news-related external knowledge, we first extract visual entity set from the visual context and textual entity set from the news text. Then the news entities are linked to the Wikidata [Vrandečić and Krötzsch, 2014] database, so as to obtain the contextual description of each entity, and construct the news background knowledge. (2) In the encoding phase, we employ the LLM encoder finetuned with LoRA [Hu *et al.*, 2021] to embed the news content which has unified to the natural language space, for obtaining the news visual context feature and textual context feature. Meanwhile, in order to prevent the loss of the original news information, we extract the news multi-modal features with the CLIP model [Radford *et al.*, 2021] and map them to the textual embedding space with a multi-modal mapping network. (3) In the decoding phase, we introduce an inference decoder with prompt phrase. This module utilizes the prompt phrase to enable the decoder from the LLM to reason about the authenticity of the news content based on the previously obtained news visual context feature, textual context feature and multi-modal feature, while transforming the generative problem into a classification problem, *i.e.*, the location-specific output from the decoder serving as the input features for the final classification. Extensive experiments demonstrate that our CFND dataset is challenging and the proposed NLIN outperforms state-of-the-art methods.

The main contributions of this paper are as following:

- We unify all the news-related information (image, text and news-related external knowledge) into the natural language space, thus addressing the effects of modality gap fundamentally.
- We introduce an encoder-decoder architecture equipped with prompt phrases for fully comprehending the news context and inferring its authenticity.
- We produce a challenging large scale, multi-platform, multi-domain multi-modal Chinese Fake News Detection (CFND) dataset.

## 2 Related Works

### 2.1 Text-based Fake News Detection

Traditional fake news detection is mainly based on analysing the semantic features of news textual content to determine its authenticity [Ma *et al.*, 2015]. Early work focused on designing complex handcrafted features, such as lexical and syntactic features [Pérez-Rosas *et al.*, 2017] based on news text, user comments or user interactions on social network [Wu *et al.*, 2015]. Recently, deep learning models have achieved promising results in detecting fake news. For example, Tseng *et al.* [Tseng *et al.*, 2022] introduced sub-event segmentation algorithms to aggregate user comments, as well as modelling the news content and user comments at various degrees of semantic granularity. Han *et al.* [Han *et al.*, 2021] extracted news entities and relationships to form a single knowledge graph, and transform the problem of detecting fake news into subgraph classification task.

### 2.2 Multi-modal Fake News Detection

Recently, multi-modal fake news detection has received considerable attention because news content forms tend to coexist with multi-modal information such as images and text.

Khattar *et al.* [Khattar *et al.*, 2019] performed fake news detection by using a multi-modal variable autoencoder to learn cross-modal correlations and shared representations of multi-modal content. Chen *et al.* [Chen *et al.*, 2022a] applied the contrast learning approach to transform unimodal features into the same feature space and aggregates unimodal and multi-modal features to varying degrees by quantifying the ambiguity between text and image. Wu *et al.* [Wu *et al.*, 2021] applied the frequency domain information of the image as the complement and fuse it with image and text features for fake news detection. Sun *et al.* [Sun *et al.*, 2021] detected fake news by capturing inconsistent semantic information of news content at cross-modal and knowledge levels. However, these above models neglect or do not fundamentally address the intractable problem of modality gap among image, text and knowledge representations of news, which in turn limits the performance of fake news detection systems.

### 2.3 Multi-modal Fake News Detection Dataset

The field of multi-modal fake news detection has constructed many datasets for research purposes. In the English domain, the Pheme [Zubiaga *et al.*, 2017] dataset consists of tweets from the Twitter platform based on five breaking news stories. The Politifact and GossipCop datasets were collected from the political and entertainment domains of the FakeNewsNet [Shu *et al.*, 2020] database respectively. The Twitter [Boididou *et al.*, 2018] dataset was created for the MediaEval Validating Multimedia Usage task release. In the Chinese domain, Weibo [Jin *et al.*, 2017] and Weibo-21 [Nan *et al.*, 2021] are two widely used Chinese multi-modal fake news detection datasets. They are both collected on the Sina Weibo platform, while the Weibo-21 dataset contains multiple categories of news. However, these datasets tend to be confined to a single platform with small data volume, which brings about some limitations.

## 3 CFND Dataset Construction

### 3.1 Data Collection

For the fake news data, we utilize six active Chinese fact-checking websites as data sources, whose news content is typically derived from public announcements, news articles and social media platforms such as Sina Weibo and TikTok. And the authenticity of news content on the websites is evaluated by expert fact-checking personnel. For the real news data, we apply four official news websites as data sources. During the data collection process, we selected news content from different categories in proportion to ensure coverage across various domains. Additionally, for each news content, we select both the news title and its corresponding image content as the text and image of a sample in the CFND dataset.

### 3.2 Data Pre-processing

Due to the fact that when crawling the news content of the website, the results often differ from the standard dataset format. Therefore, we manually check all the crawled data. The checks are described below:

- Since some crawled data may be irrelevant information, such as advertisements or irrelevant images, we perform noise data identification and removal. Additionally, we check the quality of the images and remove images with too low resolution or blurred images to ensure that the model inputs are of high quality.
- In some cases, the same news data may exist on multiple platforms, so we apply the operation of de-duplication to remove these similar data and reduce the redundancy in the dataset.
- In Chinese fact-checking websites, news headlines often tend to use rhetorical questions to express non-factual claims. Therefore, in order to alleviate the doubtfulness of headlines, we transform rhetorical questions into declarative phrases.
- To determine news domains, we reference and synthesize categorization methods from various news websites, and select five domains, while hiring 5 experts to manually label the news domains. Initially, each expert annotated independently, followed by a cross-checking process. If at least 4 experts agreed, the final domain labeling was determined.

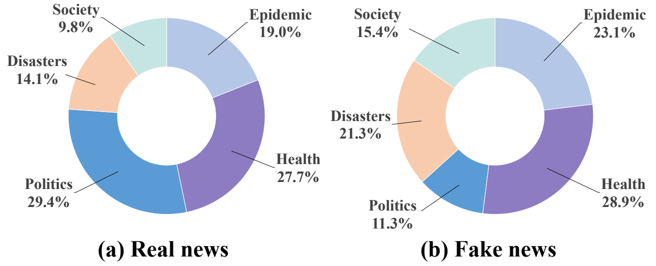


Figure 2: The statistics of CFND dataset.

After pre-processing the dataset, we obtain a total of 10,271 fake news items and 16,394 real news items, and make sure that each news item has a corresponding image content. Furthermore, we divide the whole dataset into training set, validation set and test set with the ratio of 60%, 20% and 20% respectively, and maintaining the proportion of real and fake news in each set.

### 3.3 Data Analysis

We analyze the news content in the CFND dataset, as shown in Figure 2. We can observe that its news content involves five fields, including epidemic, health, politics, disasters and society. In the fake news data, the content about health and epidemic information is the most, because the content of these two types of information is closely related to people's daily life, and rumor mongers usually spread rumors in this field to gain public attention. In the real news data, the content about politics is the most, followed by the health and epidemic field. Meanwhile, we compare and analyze the CFND dataset with the existing multi-modal fake news detection datasets, and the results are shown in Table 1. It can be observed that compared to previous datasets, CFND has a larger data volume and simultaneously encompasses news content from multiple platforms and domains.

| Dataset | Real | Fake | Image | Multi-domain | Language |
|---|---|---|---|---|---|
| Pheme [Zubiaga *et al.*, 2017] | 3,830 | 1,972 | 5,802 | ✗ | English |
| Twitter [Boididou *et al.*, 2018] | 6,225 | 9,404 | 411 | ✗ | English |
| Politifact [Shu *et al.*, 2020] | 624 | 432 | 783 | ✗ | English |
| GossipCop [Shu *et al.*, 2020] | 16,817 | 5,323 | 18,417 | ✗ | English |
| Weibo [Jin *et al.*, 2017] | 4,779 | 4,779 | 9,558 | ✗ | Chinese |
| Weibo-21 [Nan *et al.*, 2021] | 4,640 | 4,488 | 9,128 | ✓ | Chinese |
| **CFND** | **16,394** | **10,271** | **26,665** | ✓ | **Chinese** |

Table 1: Comparison with existing multi-modal fake news detection datasets.

# 4 Method

## 4.1 Overview

Multi-modal fake news detection is commonly described as a binary classification problem, aiming to determine the veracity of a given post with both news text $\boldsymbol{T}$ and image $\boldsymbol{I}$. To address this problem, we propose a novel Natural Language-centered Inference Network (NLIN) for multi-modal fake news detection, whose architecture is shown in Figure 3. Specifically, we first convert the news image and background knowledge into plain textual content. Then, we employ a unified-modal context encoder to extract features from the news content, which are unified into the natural language space, thus obtaining both visual and textual context features of the news. Meanwhile, to prevent the loss of original news information, we extend the news multi-modal feature through the CLIP model and map it to the text embedding space. Finally, we concatenate the news visual context feature, textual context feature as well as the multi-modal feature and input them together into the inference decoder, while using prompt phrase to transform the generation problem into a classification problem.

## 4.2 News Pre-processing

In this section, we present an outline of the ways in which news images and background knowledge are translated into the natural language space respectively.

**Visual Description.** In order to maximize the extraction of semantic information embedded in the image, we perform a total of the following three transformations on the image:

- Obtain the global semantic information of the image with the usage of BLIP [Li *et al.*, 2022] model for generating the image caption $\boldsymbol{C}$.
- Obtain the local semantic information of the image with the usage of VinVL [Zhang *et al.*, 2021] model to acquire the dense labelling $\boldsymbol{L}$.
- Utilize EasyOCR [EasyOCR, ] model to detect the embedded text $\boldsymbol{O}$ in the image.

At this point, we have obtained the visual context information of the image, $\boldsymbol{V}_C = (\boldsymbol{C}, \boldsymbol{L}, \boldsymbol{O})$, which can be expressed as follows:

$$\begin{aligned} &\boldsymbol{C} = \boldsymbol{M}_{BLIP}\left(I\right)\\ &\boldsymbol{L} = \boldsymbol{M}_{VinVL}\left(I\right) = \left\{l_1, ..., l_i\right\},\\ &where\ l_i = \left(w_0^{attr}, ..., w_n^{attr}, w^{obj}\right)\\ &\boldsymbol{O} = \boldsymbol{M}_{EasyOCR}\left(I\right) = \left\{w_0^{ocr}, ..., w_j^{ocr}\right\} \end{aligned} \tag{1}$$

**Background Knowledge Description.** News background knowledge can supply rich evidential information, which is beneficial for both understanding news content and improving the interpretability of fake news detection model. Therefore, in order to construct news background knowledge, we first extract news entities from the news content, after which we perform entity search in the Wikidata [Vrandečić and Krötzsch, 2014] database to obtain the specific description of each entity. The details of the process are as follows:

Regarding visual entities, we chose to extract visual entities from the visual context with the entity linking tool TagMe. Since the image object labels $\boldsymbol{L}$ extracted by the VinVL model tend to be more redundant, and their main labels are often the same as the entities in the image caption, we extracted the visual entities only using the content of the image caption $\boldsymbol{C}$ and the image embedded text $\boldsymbol{O}$. Thus, we can obtain the set of visual entities $\{\boldsymbol{E}_V^{\mathrm{v}}\}$. Regarding textual entities, we also employ the TagMe tool to obtain the textual entity set $\{\boldsymbol{E}_T^{\mathrm{t}}\}$ from the news text.

After obtaining the news visual entity set $\{\boldsymbol{E}_V^{\mathrm{v}}\}$ and the textual entity set $\{\boldsymbol{E}_T^{\mathrm{t}}\}$, we search the Wikidata database for each entity individually to acquire its corresponding entity description. Meanwhile, we form a complete sentence with the entity and its contextual description, *e.g.*, "Tony Abbott : prime minister of Australia from 2013 to 2015.", as well as treating this as a news background knowledge item description. After searching for all entities, we have completed the construction of the news visual background knowledge $\boldsymbol{V}_b$ and textual background knowledge $\boldsymbol{T}_b$.

## 4.3 Multi-modal Feature Reasoning

We utilize the Flan-T5 [Chung *et al.*, 2022] model as the backbone for the multi-modal feature reasoning module, *i.e.*, employ its text encoder and decoder as our unified-modal context encoder and inference decoder. Additionally, the CLIP [Radford *et al.*, 2021] model is applied to extract the news complementary multi-modal features. To avoid excessive training parameters and ensure that the pre-trained language model effectively adapts to the multi-modal fake news detection task, we employ low-rank adaptation (LoRA) [Hu *et al.*, 2021] to efficiently fine-tune the Flan-T5 model, which freezes the pre-trained Flan-T5 model weights and injects trainable rank decomposition matrices into each layer of the transformer architecture, significantly reducing the number of trainable parameters. The forward propagation process of the modified model can be expressed as follows:

$$\boldsymbol{h} = W_0\boldsymbol{x} + \Delta W\boldsymbol{x} = W_0\boldsymbol{x} + BA\boldsymbol{x} \tag{2}$$

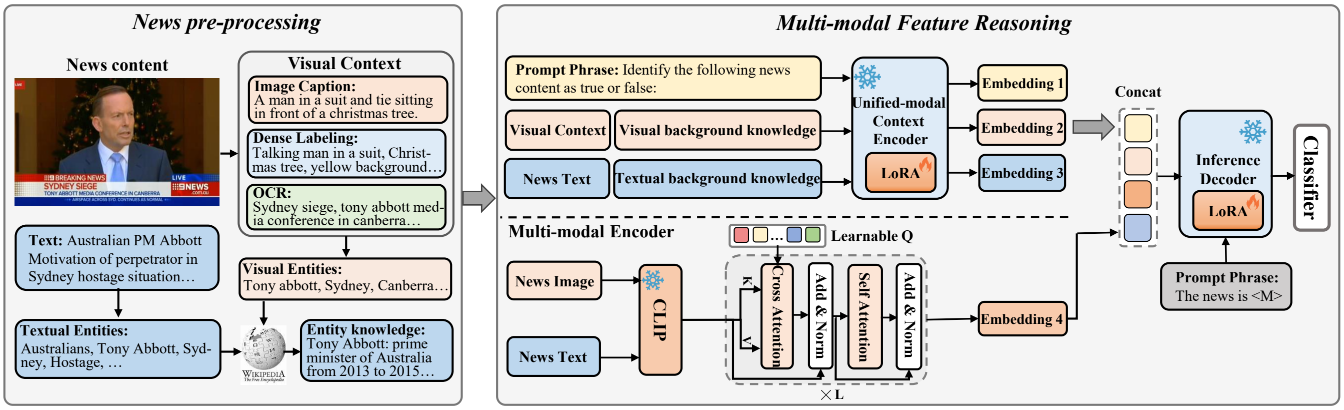

News pre-processing
News content
BREAKING NEWS
SYDNEY SIEGE
TONY ABBOTT MEDIA CONFERENCE IN CANBERRA
LIVE
9NEWS
Visual Context
Image Caption: A man in a suit and tie sitting in front of a christmas tree.
Dense Labeling: Talking man in a suit, Christmas tree, yellow background…
OCR: Sydney siege, tony abbott media conference in canberra…
Text: Australian PM Abbott Motivation of perpetrator in Sydney hostage situation…
Visual Entities: Tony abbott, Sydney, Canberra…
Textual Entities: Australians, Tony Abbott, Sydney, Hostage, …
WIKIPEDIA
Entity knowledge: Tony Abbott: prime minister of Australia from 2013 to 2015…
Multi-modal Feature Reasoning
Prompt Phrase: Identify the following news content as true or false:
Visual Context
Visual background knowledge
News Text
Textual background knowledge
Unified-modal Context Encoder
LoRA
Embedding 1
Embedding 2
Embedding 3
Concat
Inference Decoder
LoRA
Classifier
Multi-modal Encoder
Learnable Q
News Image
News Text
CLIP
K
V
Cross Attention
Add & Norm
Self Attention
Add & Norm
× L
Embedding 4
Prompt Phrase: The news is <M>

Figure 3: The overall architecture of the proposed NLIN. It consists of two modules: news pre-processing module and multi-modal feature reasoning module. The news pre-processing module is responsible for unifying news content into natural language space. And the multi-modal feature reasoning module is used to reason about the authenticity of news content.

Where $W_0$ represents the frozen parameters of the original pre-trained Flan-T5 model, $A$ is a parameter matrix initialized using random gaussian initialization and $B$ is a parameter matrix initialized with zeros.

**Unified-modal Context Encoder.** We concatenate the visual context and visual background knowledge as visual text input $\boldsymbol{T}_v$, which is denoted as <Caption $\boldsymbol{C}$> + <Dense labelling $\boldsymbol{L}$> + <OCR $\boldsymbol{O}$> + <Visual background knowledge $\boldsymbol{V}_b$>. Additionally, the news text and textual background knowledge are concatenated as news text input $\boldsymbol{T}_t$, denoted by <News text $\boldsymbol{T}$> + <Textual background knowledge $\boldsymbol{T}_b$>. Apart from this, to make the pre-trained Flan-T5 model comprehend the purpose of multi-modal fake news detection task and maximise its potential, we design a prompt phrase $\boldsymbol{T}_p$, *i.e.*, "Identify the following news content as true or false:". Then, we encode each of the above three contexts with unified-modal context encoder to obtain the prompt phrase feature $\boldsymbol{f}_P \in R^{l \times d}$, the visual context feature $\boldsymbol{f}_V \in R^{m \times d}$ and the textual context feature $\boldsymbol{f}_T \in R^{n \times d}$, where $l$, $m$ and $n$ correspond to their text lengths and $d$ refers to the embedding dimension.

**Multi-modal Encoder.** In order to prevent the loss of original news information, we utilize the CLIP model to extract multi-modal feature from news image and text, and employ a multi-modal mapping network that maps it to the same space as the text embedding, thus obtaining multi-modal embedding of the news. The details are described as follows:

We firstly apply the image encoder and text encoder of the CLIP model to extract the news image and text features respectively, which are denoted as $\boldsymbol{f}_{clip-V}$ and $\boldsymbol{f}_{clip-T}$. After that, we construct a news multi-modal feature $\boldsymbol{f}_{clip}$, which is the concatenation of the image and text features.

**Multi-modal Mapping Network.** In order to map the news multi-modal feature $\boldsymbol{f}_{clip}$ to the text embedding space, we employ a multi-modal mapping network, $Map_m$, to reconstruct the original multi-modal feature representation. Within the multi-modal mapping network, we apply a set of $M$ learnable prompts $\boldsymbol{P}_m$ to facilitate the mapping, which will be used as input along with $\boldsymbol{f}_{clip}$ to obtain the news multi-modal embedding $\boldsymbol{f}_M$.

$$\boldsymbol{f}_M = Map_m \left(\boldsymbol{P}_m, \boldsymbol{f}_{clip}\right) \tag{3}$$

Specifically, the multi-modal mapping network consists of L blocks, each containing a cross-attention layer and a self-attention layer. Each block is computed as follows:

$$\boldsymbol{f}_M^l = SA_l \left(CA_l \left(\boldsymbol{P}_m, \boldsymbol{f}_M^{l-1}\right)\right) \tag{4}$$

where $l = 1, ..., L$, and if $l = 1$, then the input $\boldsymbol{f}_M^{l-1} = \boldsymbol{f}_{clip}$. Each block in the multi-modal mapping network $Map_m$ first murders the multi-modal feature $\boldsymbol{f}_{clip}$ to an intermediate feature via the cross-attention $CA_l$, and then obtains $\boldsymbol{f}_M^l$ via the self-attention $SA_l$. In the cross-attention, $\boldsymbol{P}_m$ serves as the query, and $\boldsymbol{f}_M^{l-1}$ serves as the key and value. Thus, we iteratively extract information from the news multi-modal feature $\boldsymbol{f}_{clip}$ into the potential features and finally output the news multi-modal embedding $\boldsymbol{f}_M$.

**Inference Decoder.** To further construct the global representation of the news, we concatenate the chains of features previously extracted by the unified-modal context encoder and the multi-modal encoder to obtain the news global feature $\boldsymbol{f}_N$ of the news, and input it into the inference decoder, as shown below:

$$\boldsymbol{f}_N = Concat \left[\boldsymbol{f}_P : \boldsymbol{f}_V : \boldsymbol{f}_M : \boldsymbol{f}_T\right] \tag{5}$$

Additionally, existing research [Chen *et al.*, 2022b] has shown that prompt engineering has a significant impact on model performance. Therefore, we adopt the prompt phrases "The news is <M>" as the input to the inference decoder, which echoes the prompt phrase $\boldsymbol{T}_p$ of the unified-modal context encoder. Based on these prompt phrases, we can stimulate the powerful in-context learning ability of Flan-T5 model to inference the truthfulness of news content.

## 4.4 Model Optimization

We can transform the generative problem into a binary classification problem by taking the output $\boldsymbol{H}_{<M>}$ of the inference decoder for <M> token as the final feature used for

| Method | Pheme | | | | CFND | | | | Weibo | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Prec | Rec | F1 | Acc | Prec | Rec | F1 | Acc | Prec | Rec | F1 |
| MVAE [Khattar *et al.*, 2019] | 0.776 | 0.735 | 0.723 | 0.728 | 0.812 | 0.807 | 0.811 | 0.806 | 0.824 | 0.828 | 0.822 | 0.823 |
| SAFE [Zhou *et al.*, 2020] | 0.807 | 0.787 | 0.789 | 0.791 | 0.795 | 0.789 | 0.804 | 0.796 | 0.851 | 0.849 | 0.849 | 0.849 |
| SpotFake [Singhal *et al.*, 2019b] | 0.845 | 0.809 | 0.836 | 0.822 | 0.830 | 0.825 | 0.841 | 0.833 | 0.873 | 0.873 | 0.874 | 0.873 |
| CAFE [Chen *et al.*, 2022a] | 0.832 | 0.796 | 0.794 | 0.795 | 0.826 | 0.827 | 0.846 | 0.837 | 0.840 | 0.840 | 0.841 | 0.840 |
| MCAN [Wu *et al.*, 2021] | 0.861 | 0.830 | 0.840 | 0.835 | 0.845 | 0.831 | 0.784 | 0.807 | 0.899 | 0.899 | 0.899 | 0.899 |
| KDIN [Sun *et al.*, 2021] | 0.846 | 0.815 | 0.804 | 0.809 | 0.847 | 0.813 | 0.846 | 0.830 | 0.893 | 0.894 | 0.892 | 0.893 |
| LIIMR [Singhal *et al.*, 2022] | 0.870 | 0.848 | 0.831 | 0.839 | 0.852 | 0.817 | 0.834 | 0.826 | 0.900 | 0.882 | 0.823 | 0.847 |
| BMR [Ying *et al.*, 2023] | 0.884 | 0.872 | 0.840 | 0.855 | 0.859 | 0.834 | 0.815 | 0.824 | 0.918 | 0.912 | 0.909 | 0.910 |
| **NLIN** | **0.903** | **0.875** | **0.883** | **0.879** | **0.874** | **0.848** | **0.841** | **0.844** | **0.922** | **0.917** | **0.922** | **0.919** |

Table 2: Performance comparison to the state-of-the-art methods on Pheme, CFND and Weibo datasets.

classification. Specifically, we convert this embedding feature $\boldsymbol{H}_{<M>}$ through a fully-connected network layer with a sigmoid activation function to predict the probability of fake news occurrence. The formulation is described below:

$$\overset{\wedge}{y} = \sigma\left(\boldsymbol{W}_c\left(\boldsymbol{H}_{<M>}\right) + \boldsymbol{b}_c\right) \tag{6}$$

where $\boldsymbol{W}_c$ and $\boldsymbol{b}_c$ are the parameters of the fully-connected network layer, and $\sigma$ refers to the sigmoid function.

After that we utilize the cross entropy loss function as the loss of the whole model with the formula:

$$\mathcal{L}_p = -y\log\left(\overset{\wedge}{y}\right) - (1-y)\log\left(1-\overset{\wedge}{y}\right) \tag{7}$$

## 5 Experiments

### 5.1 Experimental Settings

**Dataset.** To evaluate the effectiveness of the proposed NLIN, we compare it with the state-of-the-art methods on the CFND, Pheme [Zubiaga *et al.*, 2017] and Weibo [Jin *et al.*, 2017] datasets. In particular, the CFND dataset is detailed as described in section 3. The Pheme dataset is constructed based on five tweets related to breaking news on the Twitter platform, each containing text with labels and corresponding images. In addition, the Weibo dataset is from Xinhua News Agency and Weibo platform, which contains a large number of labeled texts and images. For the Pheme and Weibo datasets, we divide them into training, validation and testing sets according to the ratio of 6:2:2.

**Implementation Details.** For the unified-modal context encoder, the input lengths for both the visual text input $\boldsymbol{T}_v$ and the news text input $\boldsymbol{T}_t$ are set to a maximum of 200 words. For the multi-modal mapping network, the number of blocks is 2 and the length of the learnable prompt $\boldsymbol{P}_m$ is 16. For other hyperparameters, we set the batch size to 16, the number of epochs to 60, the learning rate to 5e-4, employ the Adam optimizer, and apply the learning rate warm-up strategy. We implement all models in PyTorch and experimented on a Tesla V100 GPU.

**Evaluation Metrics.** We utilize the accuracy metric Acc as our evaluation metric. Moreover, considering the category imbalance problem, we apply precision, recall and F1 score as supplementary evaluation metrics.

| Methods | Acc | Prec | Rec | F1 |
|---|---|---|---|---|
| w/o I-T | 0.882 | 0.851 | 0.870 | 0.860 |
| w/o K-T | 0.886 | 0.856 | 0.873 | 0.864 |
| w/o L-E | 0.879 | 0.849 | 0.863 | 0.856 |
| **NLIN** | **0.903** | **0.875** | **0.883** | **0.879** |

Table 3: Ablation study of different components of NLIN's news pre-processing module on the Pheme dataset.

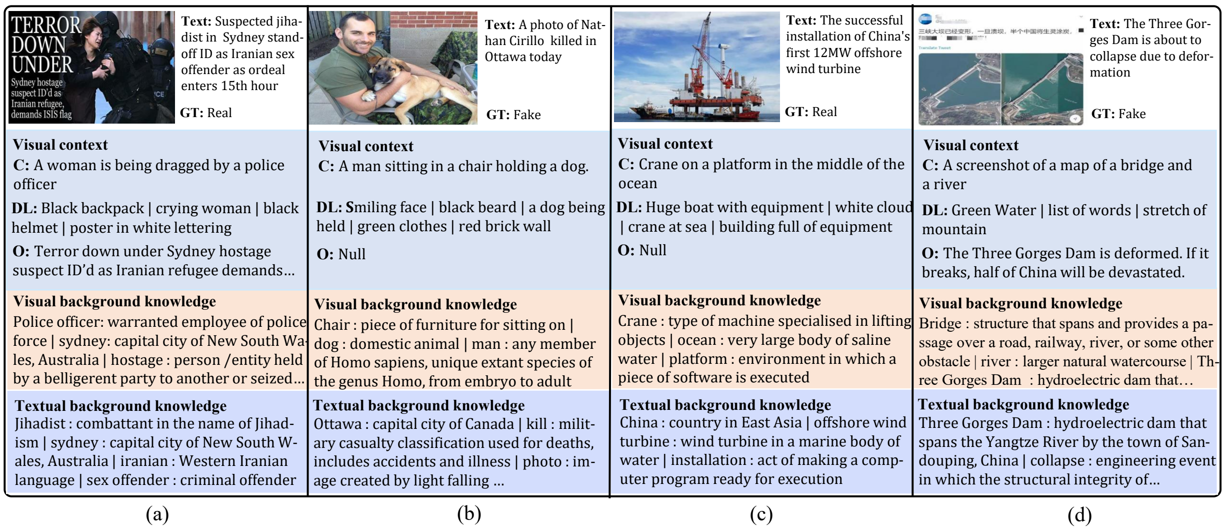
### 5.2 Results and Discussion

The results of the comparison are shown in Table 2. In all three datasets, our method NLIN outperforms other comparative methods on all evaluation metrics, demonstrating its superior performance.

Among these comparative methods, MVAE and SAFE both apply multi-modal information such as image and text, but they perform weakly as the other methods, which may be due to the fact that they apply Text-CNN and Bi-LSTM with weakly learned textual representations, suggesting that textual representation plays an important role in multi-modal fake news detection. In addition, KDIN achieves good performance, which suggests that the introduction of external knowledge information is effective in multimodal fake news detection. The BMR model achieves the second best results on all datasets, which shows that both unimodal and multimodal perspectives on news content contribute to the detection of fake news. NLIN outperforms other methods, demonstrating that unifying news content into textual modalities can avoid the gap between different modal representations of news. And utilizing the powerful in-context learning capability of the LLM to infer news authenticity can enable the NLIN model to achieve stronger detection performance with fewer training parameters.

### 5.3 Ablation Studies

The ablation experiments in Table 3 investigate the effect of different components of NLIN's news pre-processing module on the fake news detection performance. Specifically, *w/o I-T* means that the input of the unified-modal context encoder removes the visual text input $\boldsymbol{T}_v$. *w/o K-T* refers to the non-transformation of news background knowledge into textual modality, *i.e.*, the input removes the news visual and textual background knowledge. *w/o L-E* means that, in addition to the contents of the image caption and OCR text, the image


TERROR
DOWN
UNDER
Sydney hostage suspect ID'd as Iranian refugee, demands ISIS flag
三峡大坝已经变形，一旦溃坝，半个中国将生灵涂炭，
Translate Tweet
(a)
(b)
(c)
(d)

**Text:** Suspected jihadist in Sydney standoff ID as Iranian sex offender as ordeal enters 15th hour

**GT:** Real

**Visual context**
**C:** A woman is being dragged by a police officer
**DL:** Black backpack | crying woman | black helmet | poster in white lettering
**O:** Terror down under Sydney hostage suspect ID'd as Iranian refugee demands…

**Visual background knowledge**
Police officer: warranted employee of police force | sydney: capital city of New South Wales, Australia | hostage : person /entity held by a belligerent party to another or seized…

**Textual background knowledge**
Jihadist : combattant in the name of Jihadism | sydney : capital city of New South Wales, Australia | iranian : Western Iranian language | sex offender : criminal offender

**Text:** A photo of Nathan Cirillo killed in Ottawa today

**GT:** Fake

**Visual context**
**C:** A man sitting in a chair holding a dog.
**DL:** **S**miling face | black beard | a dog being held | green clothes | red brick wall
**O:** Null

**Visual background knowledge**
Chair : piece of furniture for sitting on | dog : domestic animal | man : any member of Homo sapiens, unique extant species of the genus Homo, from embryo to adult

**Textual background knowledge**
Ottawa : capital city of Canada | kill : military casualty classification used for deaths, includes accidents and illness | photo : image created by light falling ...

**Text:** The successful installation of China's first 12MW offshore wind turbine

**GT:** Real

**Visual context**
**C:** Crane on a platform in the middle of the ocean
**DL:** Huge boat with equipment | white cloud | crane at sea | building full of equipment
**O:** Null

**Visual background knowledge**
Crane : type of machine specialised in lifting objects | ocean : very large body of saline water | platform : environment in which a piece of software is executed

**Textual background knowledge**
China : country in East Asia | offshore wind turbine : wind turbine in a marine body of water | installation : act of making a computer program ready for execution

**Text:** The Three Gorges Dam is about to collapse due to deformation

**GT:** Fake

**Visual context**
**C:** A screenshot of a map of a bridge and a river
**DL:** Green Water | list of words | stretch of mountain
**O:** The Three Gorges Dam is deformed. If it breaks, half of China will be devastated.

**Visual background knowledge**
Bridge : structure that spans and provides a passage over a road, railway, river, or some other obstacle | river : larger natural watercourse | Three Gorges Dam : hydroelectric dam that…

**Textual background knowledge**
Three Gorges Dam : hydroelectric dam that spans the Yangtze River by the town of Sandouping, China | collapse : engineering event in which the structural integrity of...

Figure 4: Four examples of NLIN models making correct predictions, where (a) and (b) are from the Pheme dataset and (c) and (d) are from the CFND dataset. In the content of visual cotext: **C** represents image caption, **DL** represents dense labelling, **O** represents OCR text.

| Methods | Acc | Prec | Rec | F1 |
|---|---|---|---|---|
| w/o Clip | 0.875 | 0.845 | 0.861 | 0.853 |
| w/o Map | 0.895 | 0.868 | 0.880 | 0.874 |
| w/o LoRA | 0.893 | 0.865 | 0.882 | 0.873 |
| w/o Prompt | 0.845 | 0.810 | 0.840 | 0.825 |
| **NLIN** | **0.903** | **0.875** | **0.883** | **0.879** |

Table 4: Ablation study of different components of NLIN's multi-modal feature reasoning module on the Pheme dataset.

object labels $\boldsymbol{L}$ extracted by the VinVL model are also used to extract visual entities. Experimental results of these variants of NLIN illustrate that: (1) unifying the news images and background knowledge into textual modality avoids the gap among news different modality representations, and news background knowledge contributes to a more comprehensive understanding of news content by the model. (2) When there are more descriptions of irrelevant news entities, it will bring certain noise information to the model, which in turn destroys the semantics of the original information. This has a greater impact on the model performance than losing some of the image information.

The ablation experiments in Table 4 examine the impact of NLIN's multi-modal feature reasoning module on the performance of fake news detection. Specifically, *w/o Clip* refers to not using the CLIP to supplement multi-modal feature. *w/o Map* means not mapping news features $\boldsymbol{f}_{clip}$ to the text space with multi-modal mapping networks. *w/o LoRA* refers to not using LoRA to fine-tune the pre-trained language model. *w/o Prompt* represents the non-use of prompt phrases, while the model applies the text generation pattern directly to produce words such as "true" or "false". These variants perform significantly worse than the original NLIN, and *w/o Prompt* performs the worst of these variants. This indicates that employing prompt phrases can significantly tap into the potential of pre-trained Flan-T5 model, while utilizing the CLIP model to extract multi-modal features can effectively complement the raw information present in news content.

### 5.4 Visualization Results

Figure 4 gives four examples of correct recognition from the Pheme and CFND datasets and provides the extracted visual context, visual background knowledge and textual background knowledge. We can observe that the content of the visual context can comprehensively present the global and local semantic information embedded in the image and effectively provide background knowledge information for the news. For instance, in the example (a), the news content is relatively obscure, but through the descriptions of entities such as 'jihadist' and 'sydney', the model can fully understand the news content, which in turn facilitates to recognize fake news. In addition, referring to the example (b), by transforming the image and news background knowledge to the textual modality, the discrepancy between image and textual content is strengthened, *i.e.*, the inconsistency information between the image and text is more reflected, which undoubtedly brings benefits for performing fake news detection.

## 6 Conclusion

In this paper, we propose a novel Natural Language-centered Inference Network (NLIN) for multi-modal fake news detection. The proposed NLIN not only radically handles the huge gap among news different modality representations, but also introduces an encoder-decoder architecture equipped with prompt phrases for fully comprehending the news in-context and inferring its authenticity. In addition, we produce a challenging large scale, multi-platform, multi-domain multi-modal Chinese Fake News Detection (CFND) dataset. Extensive experiments show that our dataset is challenging and our NLIN outperforms the state-of-the-art methods.

## Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) under Grants 62106245, 62225207 and U19B2038.

## References

[Abdelnabi *et al.*, 2022] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multimodal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14940–14949, 2022.

[Boididou *et al.*, 2018] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

[Chen *et al.*, 2022a] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, pages 2897–2905, 2022.

[Chen *et al.*, 2022b] Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z Pan, Ningyu Zhang, and Wen Zhang. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. In *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, pages 20–29, 2022.

[Chung *et al.*, 2022] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[EasyOCR, ] EasyOCR. ocr. https://github.com/JaidedAI/EasyOCR.

[Han *et al.*, 2021] Yi Han, Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Knowledge enhanced multi-modal fake news detection. *arXiv preprint arXiv:2108.04418*, 2021.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Jin *et al.*, 2017] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 795–816, 2017.

[Khattar *et al.*, 2019] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921, 2019.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[Liu *et al.*, 2023] Jiawei Liu, Jingyi Xie, Yang Wang, and Zheng-Jun Zha. Adaptive texture and spectrum clue mining for generalizable face forgery detection. *IEEE Transactions on Information Forensics and Security*, 2023.

[Ma *et al.*, 2015] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1751–1754, 2015.

[Nan *et al.*, 2021] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347, 2021.

[Narayan *et al.*, 2022] Kartik Narayan, Harsh Agarwal, Surbhi Mittal, Kartik Thakral, Suman Kundu, Mayank Vatsa, and Richa Singh. Desi: Deepfake source identifier for social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2858–2867, 2022.

[Pérez-Rosas *et al.*, 2017] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Shu *et al.*, 2020] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

[Singhal *et al.*, 2019a] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47. IEEE, 2019.

[Singhal *et al.*, 2019b] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47, 2019.

[Singhal *et al.*, 2022] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference 2022*, pages 726–734, 2022.

[Sun *et al.*, 2021] Mengzhu Sun, Xi Zhang, Jianqiang Ma, and Yazheng Liu. Inconsistency matters: A knowledge-guided dual-inconsistency network for multi-modal rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1412–1423, 2021.

[Tseng *et al.*, 2022] Yu-Wun Tseng, Hui-Kuo Yang, Wei-Yao Wang, and Wen-Chih Peng. Kahan: Knowledge-aware hierarchical attention network for fake news detection on social media. In *Companion Proceedings of the Web Conference 2022*, pages 868–875, 2022.

[Vrandečić and Krötzsch, 2014] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[Wang *et al.*, 2021] Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. Multimodal emergent fake news detection via meta neural process networks. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3708–3716, 2021.

[Wu *et al.*, 2015] Ke Wu, Song Yang, and Kenny Q Zhu. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering*, pages 651–662. IEEE, 2015.

[Wu *et al.*, 2021] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569, 2021.

[Ying *et al.*, 2023] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. Bootstrapping multi-view representations for fake news detection, 2023.

[Zhang *et al.*, 2021] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[Zhang *et al.*, 2023a] Fanrui Zhang, Jiawei Liu, Qiang Zhang, Esther Sun, Jingyi Xie, and Zheng-Jun Zha. Ecenet: Explainable and context-enhanced network for muti-modal fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1231–1240, 2023.

[Zhang *et al.*, 2023b] Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zheng-Jun Zha. Hierarchical semantic enhancement network for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3424–3433, 2023.

[Zhou *et al.*, 2020] Xinyi Zhou, Jindi Wu, and Reza Zafarani. : Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*, pages 354–367. Springer, 2020.

[Zhou *et al.*, 2022] Yangming Zhou, Qichao Ying, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Multimodal fake news detection via clip-guided learning. *arXiv preprint arXiv:2205.14304*, 2022.

[Zubiaga *et al.*, 2017] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, pages 109–123. Springer, 2017.