HVLM: Hierarchical Visual-Language Models are Excellent Decision-makers for Multimodal Fake News Detection

Anonymous ACL submission

Abstract

Existing multimodal fake news detection methods based on traditional small models are prone to learn superficial features while struggling to perform knowledge-based reasoning and truly perceive fine-grained image-text consistency. Recently, fueled by large language models and multimodal pretraining techniques, large vision-language models (LVLMs) has seen significant progress in these aspects, which motivate us to transfer them for multimodal fake news detection. Specifically, barely a small LVLM (sLVLM) *Qwen2-vl-2b* as the multimodal fusion module even significantly outperforms existing methods. However, we still find two weaknesses within it:1) insufficient learning of low-level visual features; 2) difficulty in knowledge-based reasoning from a macro perspective. For the former problem, we employ an additional smaller VLM, i.e., the CLIP, as a visual-enhanced module to mitigate the weakness of the sLVLM in visual perception. For the latter problem, multi-perspective prompts are used to elicit high-level rationales from a larger un-tuned LVLM *Qwen2-vl-72B*, which are then explicitly concatenated into the input of the sLVLM as supplementary features. The three-tier framework of *CLIP-sLVLM-LVLM* forms our proposed **H**ierarchical **V**isual-**L**anguage **M**odels (HVLM). Extensive experiments on three public datasets demonstrate the significant effectiveness and generalization ability of our proposed framework.

1 Introduction

Multimodal fake news detection aims to use both news text and the corresponding image to determine the authenticity of a given news. This is a challenging task that requires the model to have two key capabilities: 1) deep semantic understanding and knowledge-based reasoning, and 2) perception of fine-grained image-text consistency. However, we point out that existing multimodal fake news detection methods still struggle to develop

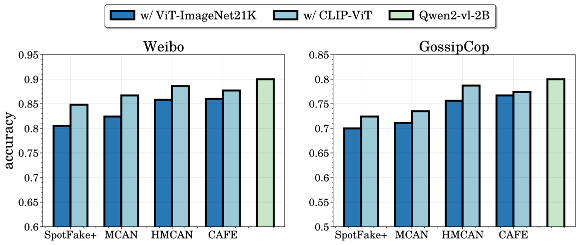


Figure 1: Test performance comparison between existing methods with different visual encoders and domain-specific fine-tuned *Qwen2-vl-2B* on Weibo and GossipCop dataset. The significant improvement indicates the vital role of semantic understanding and image-text alignment abilities for multimodal fake news detection.

these abilities. Due to limitations in model capacity and pretraining datasets, previous traditional small models (Wu et al., 2021; Qian et al., 2021) typically only learn superficial features during fine-tuning rather than understand the true meaning behind fake news. Furthermore, although many image-text alignment modules have been proposed (Chen et al., 2022; Ying et al., 2023), without pretraining on large-scale multimodal instruction data, we emphasize that these methods are unable to truly capture fine-grained image-text alignment.

Nowadays, Vision-Language Models (VLMs) based on large-scale image-text pretraining paradigms are demonstrated to have better semantic understanding and image-text matching capabilities (Radford et al., 2021). As shown in the Figure 1, simply replacing the traditional ImageNet-pretrained ViT (Dosovitskiy, 2020) with a CLIP-pretrained one as visual encoder results in significant improvements across existing methods (Singhal et al., 2020; Wu et al., 2021; Qian et al., 2021; Chen et al., 2022), even surpassing the performance boost brought by model designs.

Even more exciting is the rise of Large Vision-Language Models (LVLMs) (Li et al., 2023; Liu et al., 2024c) recently. Compared to CLIP, LVLMs, which further incorporate powerful Large Language Models (LLMs) and more complex image-text alignment tasks, possess stronger capabilities of deep semantic understanding and fine-grained image-text alignment. Therefore, in this paper, we attempt to transfer the advanced LVLMs for multimodal fake news detection to benefit from their massive advantages.

Similar to the findings of ARG (Hu et al., 2024) in exploring the performance of LLMs in text-only fake news detection tasks, for multimodal fake news detection, we find that un-tuned LVLMs can generate reasonable analysis from high-level perspective like common-sense reasoning but still lag behind small models in overall accuracy, indicating the necessity of domain-specific fine-tuning to fully unlock its potential. Considering memory and time overhead during fine-tuning, we use a relatively small Large Vision-Language Model (sLVLM[1]) *Qwen2-vl-2B*[2] as a superior multimodal fusion module for multimodal fake news detection, which already outperforms existing baselines as shown in Figure 1.

However, we still find two issues within the fine-tuned sLVLM: 1) It suffers from insufficient learning of low-level visual features including local patterns or photoshop traces, which are also important for fake news detection (Qi et al., 2019). Typically, LVLMs adopt a *visual encoder-projector-LLM decoder* architecture, where low-level visual features gradually merge with the text input and the internal parameters of the LLM during the forward pass, resulting in significant loss of information. 2) It has difficulty in knowledge-based reasoning from a macro perspective. Typically, traces for identifying fake news can be multi-level, including high-level clues like common-sense errors, mid-level clues like emotional features, or lower-level patterns or statistical features. Since the datasets only contain binary labels without fine-grained guidelines, the model may prone to rely on mid- and low-level features during the fine-tuning, while hard to capture high-level features.

To this end, we propose our **H**ierarchical **V**isual-**L**anguage **M**odels (HVLM), which fully leverages the advantages of large, medium, and small-scale VLMs for multimodal fake news detection. To compensate for the failure of sLVLM on the visual side, we additionally use a smaller VLM, specifically the CLIP-pretrained ViT, to extract individual visual features and concatenate them with the multimodal features obtained from sLVLM to enhance visual representation. To address the insufficient learning of high-level features and to fully leverage the advantages of LLM's world knowledge and reasoning abilities, we use a larger LVLM *Qwen2-vl-72B* (Bai et al., 2023) as an agent model for rationale augmentation. Specifically, we use carefully designed prompts to guide the agent to providing high-level rationales from various perspectives. Afterwards, we further prompt the agent to extract key statements from all the analyses, which not only reduces the model's complexity but also filters out noisy information. Finally, the refined rationales are explicitly concatenated into the input of the sLVLM as supplementary chain-of-thoughts, thereby injecting deeper insights into the model's training.

Extensive experiments conducted on three widely used real-world benchmark datasets consistently demonstrate the superior effectivene of our method, which outperforms all baseline methods by a large margin. Furthermore, HVLM can serve as a plug-and-play module, which could be easily integrated into future LVLMs. To summarize, the main contribution of this work is threefold:

- We are the first to comprehensively explore the capabilities of LVLMs for multimodal fake news detection task, both in fine-tuned and un-tuned scenarios, and have extensively analyzed its limitations and potentials.
- We propose a novel framework HVLM, which fully leverages the advantages of VLMs of different sizes. It comprehensively captures both micro and macro-level features, achieving optimal overall performance for multimodal fake news detection.
- We conducted extensive experiments on three well-known public datasets. The empirical results validate the significant superiority of our proposed framework.

2 Related Work

2.1 Multimodal Fake News Detection

With the growing popularity of multimodal news online, multimodal fake news detection has gained much attention in recent years (Hu et al., 2022b). In general, these methods first use individual unimodal feature encoders to separately extract textual and visual features, and then design various

[1]For clarity, LVLM refers to models with 7B(+) parameters, while sLVLM refers to models with 2B(-) parameters.
[2]http://huggingface.co/Qwen/Qwen2-VL-2B-Instruct

cross-modal fusion strategies to combine the features and output the final prediction (Jin et al., 2017; Wang et al., 2018; Khattar et al., 2019; Song et al., 2021; Qi et al., 2021; Zheng et al., 2022; Zhou et al., 2023; Liu et al., 2024a). To capture fine-grained correlations across modality, HMCAN (Qian et al., 2021) uses a multimodal contextual attention network to model both inter-modality and intra-modality features. MCAN (Wu et al., 2021) extracts both spatial-domain and frequency-domain features from image and then fuse them with textual features using multiple co-attention layers. BMR (Ying et al., 2023) individually trains each uni-modal counterparts and then adaptively aggregates them based on MOE network. Furthermore, many methods also consider the cross-modal consistency degree as an important indicator for detecting fake news (Zhou et al., 2020; Xue et al., 2021; Chen et al., 2022; Hu et al., 2022a; Wang et al., 2023; Sun et al., 2023; Wu et al., 2023; Ma et al., 2024). However, we point out that although these methods have achieved some success, the limitations of model capacity and pretraining tasks keep these methods still stuck at the stage of capturing superficial features, lacking deep image-text understanding and fine-grained cross-modal alignment abilities, which in turn limits the performance potential of the models.

2.2 Large Vision-Language Models

In recent years, the development of Large Vision-Language Models (LVLMs) has seen significant progress (Alayrac et al., 2022; Li et al., 2022; Zhu et al., 2023; Li et al., 2023; Liu et al., 2024c). By combining visual encoder with powerful LLMs and further pretraining using multimodal instruction data, these models have shown impressive performance across a range of tasks (Liu et al., 2024b). Previous studies have applied LLMs to fake news detection, but research on using LVLMs for multimodal fake news detection remains scarce. In the text-only domain, ARG (Hu et al., 2024) finds that un-tuned LLMs perform worse than fine-tuned traditional small models when making decisions independently. It proposes to use the analysis of LLMs to assist in training small models through knowledge distillation. LeRuD (Liu et al., 2024d) employs LLMs to extract key traces in user comments to effectively identify fake news. DELL (Wan et al., 2024) decomposes the fake news detection task into multiple sub-tasks and uses LLMs to handle them separately and integrate the final decisions. GenFEND (Nan et al., 2024) uses LLMs to simulate user behaviors and generates user comments to enhance the model performance. However, these methods mainly focus on the text-only field while lacking exploration into multimodal fake news detection. Additionally, they all control the LLMs' behavior via prompts, either to assist small models, or to make decisions independently. In this paper, we aim to explore a new paradigm based on both fine-tuned small LVLMs and un-tuned large LVLMs to fully utilize their capabilities for the multimodal fake news detection.

3 Preliminaries

3.1 Problem Formulation

Given a multimodal dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1,\dots,n}$ with each sample contains text, a corresponding image, i.e., $X_i = (X_{i,t}, X_{i,v})$ and a ground-truth label $y_i \in \{0, 1\}$. As a binary classification problem, the goal of multimodal fake news detection is to learn a set of features i.e., uni-modal features and cross-modal features and finally output the prediction $\hat{y} = 1$ for the fake news and $\hat{y} = 0$ for the real news respectively.

Table 1: Zero-shot test accuracy on Weibo dataset (1641 samples) of several un-tuned LVLMs under different prompts with details presented in appendix B. $*$ denotes accuracy on a subset of samples.

Model	Acc.	
	$\mathcal{P}_1$	$\mathcal{P}_2^*$
Qwen2-vl-7B	0.709	0.732 (487/665)
Qwen2-vl-72B	0.802	0.886 (542/612)
Llava-ov-7B	0.738	0.745 (1035/1390)
Llava-ov-72B	0.814	0.873 (958/1097)
Qwen-vl-max	0.731	0.880 (478/573)
SpotFake+	0.848	

3.2 Zero-shot Performance of VLMs

In this section, we first investigate the multimodal fake news detection performance of un-tuned LVLMs. To achieve a comprehensive evaluation, we design two types of prompts, i.e., $\mathcal{P}_1$ and $\mathcal{P}_2$, and select both open-source and closed-source LVLMs at different scales for testing. The specific prompts and settings are detailed in Appendix B. For comparison, we also report the performance of fine-tuned small modal SpotFake+ (Singhal et al., 2020). The results are shown in Table 1, from

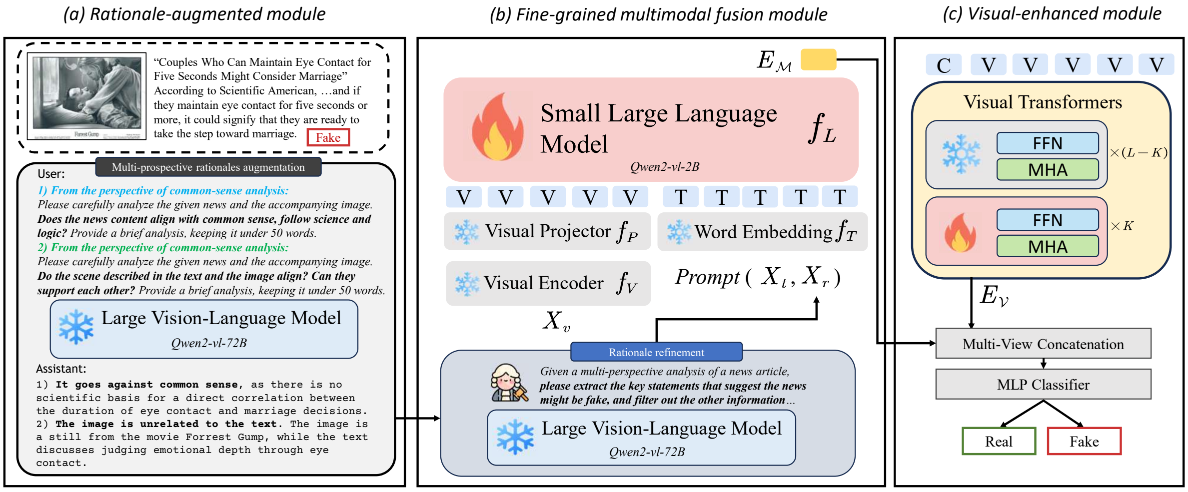


Figure 2: An illustration of our HVLM framework, which consisis of three main parts. (a) Multi-perspective prompts are used to elicit high-level rationales from a larger frozen LVLM *Qwen2-vl-72B*, which are further refined to obtain X_r (b) The news image X_v, text X_t and rationales X_r are then integrated and fused by sLVLM *Qwen2-vl-2B* to benefit from its deep image-text underestaning and fine-grained alignment abilities. (c) An additional visual-enhanced module is utilized to mitigate the weakness of the sLVLM in low-level visual perception. The multimodal features $E_{\mathcal{M}}$ and visual-enhanced features $E_{\mathcal{V}}$ are then concatenated for the final prediction.

which we can obtain the following observations: 1) The performance of prompt $\mathcal{P}_1$ indicates that relying solely on the un-tuned LVLM is insufficient, as its performance still falls short compared to the fine-tuned small model. 2) However, in $\mathcal{P}_2$, we don't require the LVLM to give explicit predictions for all samples. Instead, it only predicts for those having clear common-sense or scientific errors. The unexpected performance, even surpassing the fine-tuned small model, suggests that an un-tuned LVLM may not be capable of detecting fake news for all samples, but it can indeed be effective to detect specific cases from a high-level perspective. 3) From $\mathcal{P}_1$ to $\mathcal{P}_2$, the 7B models show only a marginal improvement, indicating models at this scale still lack sufficient knowledge-based reasoning and instruction following abilities. In contrast, the 72B models achieve a more significant improvement. This highlights the importance of the LVLM's inherent capabilities to reason from a high-level perspective, with the prompt serving merely as a tool to activate specific abilities.

4 Method

In this section, we introduce our proposed HVLM framework in detail, as depicted in Figure 2.

4.1 Transfer LVLMs for Fake News Detection

We first discuss how to fine-tune the LVLM for multimodal fake news detection to benefit from its superior semantic understanding and fine-grained cross-modal alignment abilities. In general, existing LVLMs follow the similar paradigm, i.e., *visual encoder-visual projector-LLM decoder*, which is first introduced by Llava (Liu et al., 2024c). Therefore, we emphasize that our method can serve as a plug-and-play module, which could be easily integrated into future LVLMs.

Given a news with an image $X_v \in \mathbb{R}^{H\times W\times 3}$ and a text X_t, where H and W are the origin resolution. First, the input image X_v is partitioned into 2d patches $P_v = \left[p_v^1, p_v^2, ..., p_v^{N_p}\right] \in \mathbb{R}^{N_P\times C}$, where $N_P = \frac{H\times W}{P^2}$. N_P represents the sequence length of visual tokens and P is the patch size. Visual encoder f_V is designed to encode them into visual features $F_v \in \mathbb{R}^{N_P\times C}$. Then A visual projector f_P, consisting of two linear layers with a GELU activation function, is used to map F_v into the embeddings $H_v \in \mathbb{R}^{N_P\times D}$ in text embedding space, where D represents the embedding dimensions of LLM decoder.

Next, we turn to the text input. Since we need to fine-tune the LLM for the classification task and the fake news detection datasets only contain binary labels, we require the LLM to output a single token representing the prediction, without generating additional information. To achieve this, we use the prompt $\psi_{\mathcal{C}_1}$="*<image> You need to act as a fake news detection model. Given a news article and a related image, you need to determine the authenticity of the news. Output 0 for real news and*

1 for fake news. News content: <text> ", which guides the LLM to directly provide the prediction.

After that, the text embedded in the template $\psi_{\mathcal{C}_1}(X_t)$ is tokenized and then projected to textual features $H_t \in \mathbb{R}^{N_t \times D}$ using word embedding layer f_T, where N_t represents the sequence length of textual tokens. Subsequently, multimodal pretrained LLM decoder is able to achieve a unified understanding of both visual and textual information and gradually fuse them through attention mechanism. we concatenate the visual tokens and textual tokens together as input for the LLM. The forward process of the LLM can be formulated as:

$$\mathrm{x}_0 = [H_v, H_t], \tag{1}$$

$$\mathrm{x}'_\ell = \mathrm{MHA}\left(\mathrm{LN}\left(\mathrm{x}_{\ell-1}\right)\right) + \mathrm{x}_{\ell-1}, \ell = 1...L, \tag{2}$$

$$\mathrm{x}_\ell = \mathrm{FFN}\left(\mathrm{LN}\left(\mathrm{x}'_\ell\right)\right) + \mathrm{x}'_\ell, \ell = 1...L, \tag{3}$$

$$E_{\mathcal{M}} = \mathrm{LN}\left(\mathrm{x}_L^{[-1]}\right). \tag{4}$$

The LLM is composed of stacked multi-head attention (MHA) and feed-forward neural networks (FFN). Layer normalization (LN) and residual connections are also applied between the modules. Originally, the LLM would use a fully connected layer *lm_head* to project $\mathrm{LN}\left(\mathrm{x}_L\right)$ into probability distributions over the vocabulary tokens for generation. However, to ensure the model outputs valid content for the binary classification, we train a new classification head $f_{\mathcal{C}}$, consisting of multiple fully connected layers, on top of the hidden state of the last token in the final layer. The whole process is presented as:

$$E_{\mathcal{M}} = f_L\left(f_P\left(f_V\left(\mathrm{X}_v\right)\right), f_T\left(\psi_{\mathcal{C}_1}\left(\mathrm{X}_t\right)\right)\right), \tag{5}$$

$$y_{\mathcal{M}} = f_{\mathcal{C}}\left(E_{\mathcal{M}}\right), \tag{6}$$

where $E_{\mathcal{M}}$ is the multimodal features after modality fusion through the LLM decoder f_L, and $y_{\mathcal{M}}$ is the binary prediction output by $f_{\mathcal{C}}$.

4.2 Rationale-Augmented Module

Due to the lack of fine-grained supervisory signals in fake news detection datasets, and the limited capacity of the sLVLM, it is difficult for the model to uncover the high-level features of the news during fine-tuning. In Section 3.2, we have already demonstrated that the un-tuned LVLM can provide valuable judgments from some high-level perspectives. Therefore, we propose to guide a larger LVLM to act as an agent model to output high-level rationales, which are then explicitly concatenated into the input of the sLVLM as supplementary chain-of-thoughts. Different from Section 3.2, considering that the un-tuned LVLM may not cover all possible clues, we do not require it to output judgments but instead only provide analysis from a given angle.

Specifically, to maximize the advantages of the agent model while avoiding redundancy, we guide it to generate rationales from two perspectives: common-sense analysis and image-text coherence. The former aims to analyze whether the news violates common sense, logic, or science, while the latter focuses on examining whether image and text of the news corroborate each other from an overall perspective. In contrast, we do not use the agent to analyze the writing style or emotional tone of the news, as these features can be learned by the sLVLM during fine-tuning. The detailed prompts and more discussions are presented in Appendix C. The multi-perspective rationale-augmentation process can then be represented as:

$$\mathcal{R}_i = \mathrm{LVLM}\left(X_v, \psi_{\mathcal{R}_i}\left(X_t\right)\right), i = 1, ..., N_r, \tag{7}$$

where $\mathcal{R}_i$ represents the rationale generated under specific prompt $\psi_{\mathcal{R}_i}$, and $N_r = 2$. In addition, a piece of fake news may contain common-sense errors, but the image and text might match, as the image could have been manipulated through Photoshop. This can create conflicting analysis, leading to ambiguity in the model's judgment. Therefore, we further use the agent model to streamline the multi-perspective analysis, filtering out noisy information, which can be represented as:

$$X_r = \mathrm{LVLM}\left(\psi_{\mathcal{S}}\left(\Sigma\mathcal{R}_i\right)\right), \tag{8}$$

where X_r represents the final rationale summarized by the agent model under prompt $\psi_{\mathcal{S}}$, which is presented in Appendix D with detailed discussions. After that, we further use a new classification prompt $\psi_{\mathcal{C}_2}$ to aggregate it into the model input for the sLVLM, where $\psi_{\mathcal{C}_2}$ = *"<image> You need to act as a fake news detection model. Given a news article and a related image, you need to determine the news' authenticity. Output 0 for real news and 1 for fake news. News content: <text> Analysis: <rationale>."*. After adding the rationale-augmented module, the multimodal features $E_{\mathcal{M}}$ output by the sLVLM can be further formalized as:

$$E_{\mathcal{M}} = f_L\left(f_P\left(f_V\left(\mathrm{X}_v\right)\right), f_T\left(\psi_{\mathcal{C}_2}\left(\mathrm{X}_t, X_r\right)\right)\right). \tag{9}$$

4.3 Visual-Enhanced Module

Apart from high-level features, multimodal fake news detection also heavily relies on low-level visual features like local patterns or Photoshop traces. However, despite the fine-tuned sLVLM already outperforming traditional models, we find that the it still suffers from insufficient learning of uni-modal features of visual modality. Due to the inherent modality imbalance caused by the model structure of the LVLM, low-level visual information is severely lost during the LLM's forward process.

To evaluate the model's uni-modal performance, we additionally test the model with input from single modality during training. For visual modality, we remove the textual input and use a new classification prompt $\psi_{\mathcal{C}}^{v}$ ="*<image> You need to act as a fake news detection model. Given a news image, you need to determine the news' authenticity. Output 0 for real news and 1 for fake news.*". For textual modality, we remove the visual input and use the prompt $\psi_{\mathcal{C}}^{t}$ ="*You need to act as a fake news detection model. Given a news article, you need to determine the news' authenticity. Output 0 for real news and 1 for fake news. News content: <text>*".

For comparison, we also report the performance of the individually trained ViT model as the baseline performance for visual modality. As shown in Figure 3, the visual performance of sLVLM lags significantly behind that of the ViT model, despite both using ViT as the visual encoder. To mitigate the weakness of the sLVLM in visual perception while avoiding disrupting its forward process, we introduce an additional ViT to extract pure visual features:

$$\mathrm{z}_0 = \left[p_v^{\mathrm{cls}}, p_v^1\mathrm{W}, p_v^2\mathrm{W}, ..., p_v^{N_p}\mathrm{W}\right], \quad (10)$$

$$\mathrm{z}_\ell' = \mathrm{MHA}\left(\mathrm{LN}\left(\mathrm{z}_{\ell-1}\right)\right) + \mathrm{z}_{\ell-1}, \ell = 1...L, \quad (11)$$

$$\mathrm{z}_\ell = \mathrm{FFN}\left(\mathrm{LN}\left(\mathrm{z}_\ell'\right)\right) + \mathrm{z}_\ell', \ell = 1...L, \quad (12)$$

$$E_{\mathcal{V}} = \mathrm{LN}\left(\mathrm{z}_L^{[0]}\right), \quad (13)$$

where $\mathrm{W} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is a linear projector. The ViT is also composed of stacked MHA and FFN blocks. The last hidden state of `[cls]` token is used as visual-enhanced feature. The above process can be simplified as $E_{\mathcal{V}} = f_{\mathcal{V}}\left(\mathrm{X_v}\right)$. We then use simple concatenation to fuse $E_{\mathcal{M}}$ and $E_{\mathcal{V}}$ and output the prediction with classification head $f_{\mathcal{C}}$:

$$y_{\mathcal{F}} = f_{\mathcal{C}}\left(E_{\mathcal{M}} \oplus E_{\mathcal{V}}\right). \quad (14)$$

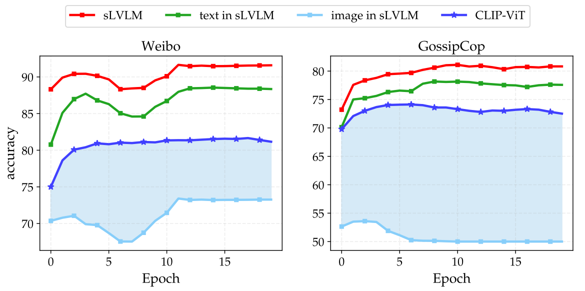


Figure 3: Performance comparison in test accuracy. The sLVLM is trained with multimodal input while tested with multimodal, text-only, image-only input. Compared to individual ViT model, the shaded area indicates the severe under-optimization of visual modality.

4.4 Model Training

Finally, we train the model using Binary Cross-Entropy loss, which can be formulated as:

$$\mathcal{L} = -y \log\left(y_{\mathcal{F}}\right) - (1-y)\log\left(1-y_{\mathcal{F}}\right). \quad (15)$$

For sLVLM, we only fine-tune the LLM decoder f_L while keeping other parameters fixed. LoRA (Low-Rank Adaptation) (Hu et al., 2021) is employed to prevent overfitting and save memory overhead. In particular, for the specified linear layer $W \in \mathbb{R}^{d \times m}$ in f_L, we fix the original parameters W and instead train two low-rank matrices, $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times m}$, for updates, i.e., $W' = W + AB$, where r is the rank and much smaller than d and m. Similarly, to avoid overfitting, we only fine-tune the last K layers in $f_{\mathcal{V}}$, where K is a hyper-parameter.

5 Experiments

5.1 Experimental Settings

We employ the widely used Chinese dataset Weibo (Jin et al., 2017) and the English dataset GossipCop (Shu et al., 2020) for evaluation. In addition, we additionally use the Chinese dataset Weibo21 (Nan et al., 2021) to study the generalization ability of our method. To validate the superior effectiveness of our proposed method, we also conduct experiments on several most representative fake news detection methods for comparison. The uni-modal methods include: 1)BERT; 2)CLIP-ViT. The multi-modal methods include: 3)SpotFake+; 4)MCAN; 5)HMCAN; 6)CAFE. Other details of experiment settings can be found in appendix A.

5.2 Main Results

The overall performance of our proposed HVLM and baseline methods is shown in Table 2, from

Table 2: Performance comparison between HVLM and other baseline methods in terms of Accuracy, Precision, Recall and F1 Score. The best performance is highlighted **in bold**.

Datasets	Models	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	BERT	0.818	0.863	0.790	0.825	0.773	0.851	0.810
	CLIP-ViT	0.766	0.752	0.771	0.762	0.779	0.761	0.770
	SpotFake+	0.848	0.839	0.852	0.846	0.856	0.843	0.850
	MCAN	0.867	0.875	0.860	0.868	0.859	0.874	0.867
	HMCAN	0.886	0.885	0.886	0.885	0.887	0.886	0.887
	CAFE	0.877	0.866	0.884	0.875	0.887	0.870	0.879
	HVLM	**0.939**	**0.930**	**0.945**	**0.938**	**0.947**	**0.932**	**0.939**
GossipCop	BERT	0.722	0.666	0.750	0.706	0.778	0.700	0.737
	CLIP-ViT	0.706	0.708	0.706	0.707	0.705	0.707	0.706
	SpotFake+	0.724	0.741	0.716	0.729	0.706	0.732	0.719
	MCAN	0.735	0.721	0.742	0.731	0.749	0.729	0.738
	HMCAN	0.787	0.745	0.814	0.778	0.829	0.765	0.796
	CAFE	0.774	0.760	0.783	0.771	0.789	0.766	0.778
	HVLM	**0.832**	**0.774**	**0.876**	**0.822**	**0.890**	**0.798**	**0.841**

which we could have the following key points:

- Compared to uni-modal methods, multimodal methods achieve better performance, demonstrating the importance of modality collaboration.
- Although the introduction of CLIP improves the performance of existing baseline methods, HVLM still outperforms them on both Weibo and GossipCop datasets by a large margin, achieving improvements of 6.0% and 5.7% in classification accuracy, respectively. This demonstrates that the introduction of LVLM significantly enhances the model's capabilities in deep semantic understanding, fine-grained image-text alignment, and knowledge-based reasoning, further pushing the performance ceiling. The three-tier framework of *CLIP-sLVLM-LVLM* can comprehensively capture both micro and macro-level features, achieving optimal overall performance for multimodal fake news detection.
- The textual modality, as the dominant modality, actually plays a more important role for both Weibo and GossipCop datasets. However, we find that introducing a visual-enhanced module to improve the learning of visual features can still boost the model's overall performance, underscoring the necessity of fully utilizing all types of features.

5.3 Ablation Study

In this section, to evaluate the effectiveness of each component of our proposed HVLM, we remove each module from the entire framework for comparison. Specifically, the compared variants of HVLM are implemented as follows: *-w/o Text*: This variant only uses news text as input but removes image. *-w/o Image*: This variant only uses news image as input but removes text. *-w/o VE*: This variant removes vision-enhanced module. *-w/o RA*: This variant removes rationale-augmented module.

The experimental results are shown in Table 3 and we have the following observations: 1) If we only use uni-modal input, the model's performance will considerably decline, which indicates both textual and visual modalities of news help improve the model's overall performance. 2) If we remove the vision-enhanced module (VE), there is a significant decrease in model performance on all datasets. This demonstrates the sLVLM has severe insufficient learning of visual features while the vision-enhanced module well mitigates this problem. 3) If we remove the rationale-augmented module (RA), the model's performance also declines. This indicates the analysis from a larger un-tuned LVLM can provide insights from a higher perspective, addressing the shortcomings of sLVLM in the ability of reasoning and the width of world knowledge.

5.4 Impact of the backbone sLVLM

In this section, we further conduct experiments based on a new backbone sLVLM *Llava-onevision-0.5B*[3], which is the latest model in *Llava* series, to explore the impact of using different sLVLMs for

[3]https://huggingface.co/llava-hf/llava-onevision-qwen2-0.5b-ov-hf

Table 3: Performance comparison between HVLM and its several variants for ablation study.

Models	Weibo		GossipCop	
	Acc	F1	Acc	F1
HVLM	0.939	0.939	0.832	0.832
-w/o Text	0.821	0.820	0.747	0.747
-w/o Image	0.899	0.899	0.797	0.797
-w/o VE	0.922	0.921	0.817	0.816
-w/o RA	0.927	0.927	0.826	0.827
-w/o VE+RA	0.918	0.917	0.811	0.811

our method. We emphasize that our HVLM can serve as a plug-and-play module which could be easily integrated into any sLVLM backbone. The experimental results are shown in Table 4, from which we could draw the following conclusions: 1) LVLM, by combining powerful LLM and multimodal pre-training techniques, exhibits strong potential for multimodal fake news detection. Both *Qwen2-vl* and *Llava-onevision* consistently outperform traditional small models by a large margin. 2) By enhancing the model's visual feature learning and common-sense reasoning abilities, HVLM consistently improves the performance of *Llava-onevision*, proving the wide applicability of our method.

Table 4: Impact of backbone sLVLM. Test performance on Weibo dataset is reported with *Qwen2-vl* replaced by *Llava-onevision* in the HVLM model.

Models	Acc	Fake News			Real News		
		P	R	F1	P	R	F1
repl.Llava-ov	0.929	0.911	0.945	0.928	0.948	0.915	0.931
-w/o VE	0.915	0.906	0.921	0.913	0.924	0.908	0.916
-w/o RA	0.920	0.912	0.927	0.919	0.928	0.914	0.921
-w/o VE+RA	0.910	0.938	0.889	0.912	0.884	0.935	0.908

5.5 Generalization Study

In this section, we explore the generalization ability of our proposed method. To eliminate the influence of language, we choose Weibo and Weibo21 datasets for our experiment. Specifically, we first train the model on one of the datasets and then test the trained model on the other dataset. We also conduct experiments on baseline methods for comparison. As the results shown in Figure 4, HVLM consistently outperforms the baseline methods, demonstrating that the knowledge learned by HVLM can generalize to new datasets. This also highlights the vital importance of the rationale-augmented module in helping build a more robust fake news detection system, as it can avoid the influence of bias in the training data and provide valuable analysis from a neutral standpoint.

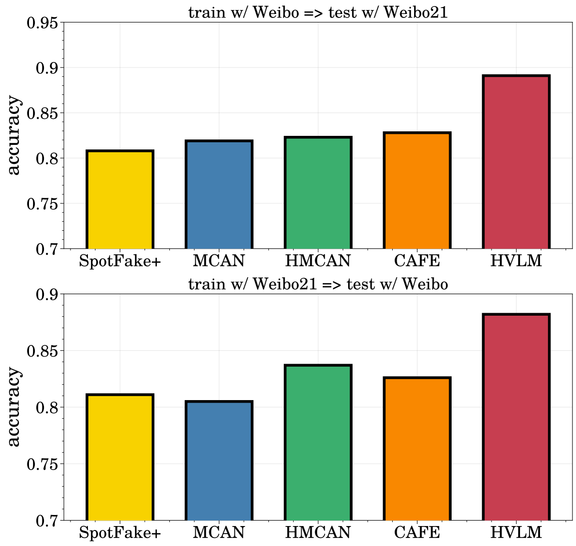


Figure 4: Generalization study on the Weibo and Weibo21 datasets. We first train the models on one of the datasets and then report the test accuracy of them on the other dataset.

5.6 Case Study

In this section, we present some samples to demonstrate that the rationale-augmented module can provide valuable analysis for multimodal fake news detection. Please refer to Appendix E for more details.

6 Conclusion

In this work, we attempt to transfer the advanced Large Vision-Language Models (LVLM) for multimodal fake news detection to benefit from their massive advantages. We first investigate potentials and limitations of LVLMs in both fine-tuning and non-tuning scenarios. Then, we propose our novel HVLM, comprising a three-level hierarchy of large, medium, and small-scale VLMs, to comprehensively capture both micro and macro-level features, thereby achieving optimal performance. Extensive experiments on three public datasets demonstrate the significant effectiveness and generalization ability of our proposed framework.

7 Limitations

Although our proposed HVLM has achieved outstanding performance, we acknowledge that our work still has several limitations: 1) Although we have found that larger-scale models can have stronger scene understanding and instruction-following abilities, controlling the agent model's behavior solely through prompts may be insufficient. This is because simple instructions can not cover all possible scenarios, leading the model to produce incorrect conclusions for some hard samples. To solve this, it might require constructing high-quality datasets for diverse cases in each predefined aspect and then injecting fine-grained guidelines into the agent model through in-context learning or fine-tuning. 2) The knowledge stored in LVLM may not be extensive enough and could become outdated. Therefore, employing retrieval-augmented generation (RAG) techniques (Lewis et al., 2020) to fetch relevant knowledge from the web could also improve the quality of the agent model's outputs. 3) How to leverage LVLMs for more robust and general multimodal fake news detection still presents many problems unsolved. For instance, LVLMs inherently possess multilingual understanding capabilities. Exploring how to leverage cross-lingual datasets to train a more generalized fake news detector is a potential research direction.

8 Ethics Statements

Social Impact Our work aims to detect multimodal fake news, as fake news can lead to significant social consequences, including the spread of misinformation, political polarization, and harm to public trust. Therefore, our work contributes positively to social harmony and stability. We are committed to ensuring that the methods developed are not misused, and that the research adheres to the highest ethical standards in promoting truthful and responsible information dissemination. However, we must still be mindful of the risks of our approach being misused. For example, attackers may develop attack algorithms based on our model as a surrogate model and then target the fake news detection model deployed online. This could lead to the online model being unable to effectively detect manipulated fake news, causing negative impacts on society. Therefore, we suggest enhancing the online model's robustness through model ensemble techniques.

Data Privacy We emphasize that the datasets we use are all publicly available, and we strictly adhere to the relevant regulations during their use. All the data used in this study are carefully processed through appropriate data anonymization techniques to protect the privacy of individuals.

Informed Consent This study does not involve direct interaction with human participants.

Bias and Fairness We recognize that the fake news detection algorithms could potentially exhibit biases based on the training data. To address this, we take steps to ensure that the datasets used are diverse and representative. Furthermore, we remain committed to continuously evaluating and mitigating bias within our model to ensure fairness and accuracy in detecting misinformation.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

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

Linmei Hu, Ziwei Chen, Ziwang Zhao Jianhua Yin, and Liqiang Nie. 2022a. Causal inference for leveraging image-text matching bias in multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022b. Deep learning for fake news detection: A comprehensive survey. *AI open*, 3:133–155.

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

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Guofan Liu, Jinghao Zhang, Qiang Liu, Junfei Wu, Shu Wu, and Liang Wang. 2024a. Uni-modal event-agnostic knowledge distillation for multimodal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Qiang Liu, Xiang Tao, Junfei Wu, Shu Wu, and Liang Wang. 2024d. Can large language models detect rumors on social media? *arXiv preprint arXiv:2402.03916*.

Ilya Loshchilov, Frank Hutter, et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.

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

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1732–1742.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1212–1220.

Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 518–527. IEEE.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

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

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. 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.

Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13915–13916.

Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1):102437.

Mengzhu Sun, Xi Zhang, Jianqiang Ma, Sihong Xie, Yazheng Liu, and S Yu Philip. 2023. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Longzheng Wang, Chuang Zhang, Hongbo Xu, Yongxiu Xu, Xiaohan Xu, and Siqi Wang. 2023. Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 5696–5704.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Lianwei Wu, Pusheng Liu, Yongqiang Zhao, Peng Wang, and Yangning Zhang. 2023. Human cognition-based consistency inference networks for multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2560–2569.

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 5384–5392.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In *IJCAI*, volume 2022, pages 2413–2419.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2825–2830. IEEE.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Detailed Experimental Settings

A.1 Baselines

To validate the superior effectiveness of our proposed method, we also conduct experiments on several most representative fake news detection methods, including both uni-modal methods and multimodal methods. All the baselines are open-source and we use the code published to conduct the experiments.

- **BERT** (Devlin et al., 2018) employs a bidirectional transformer encoder pre-trained with masked language modeling to capture deep contextual semantics, enabling effective transfer learning for various NLP tasks such as text classification and question answering. In our work, we use it to extract textual features and then additionally train MLPs for classification.

- **CLIP-ViT** (Radford et al., 2021) integrates Vision Transformer (ViT) as the visual encoder in the CLIP framework, aligning visual and textual features via contrastive learning to improve performance on tasks like zero-shot prediction and image-text matching. In our work, we use it to extract visual features and then additionally train MLPs for classification.

- **SpotFake+** (Singhal et al., 2020) can be regarded as a vanilla multimodal baseline, it first extract textual features and visual features from uni-modal pre-trained models, which are then concatenate after being projected the same dimension. After that, MLPs are used to fuse the multimodal feature and yield the final prediction.

- **MCAN** (Wu et al., 2021) extracts both spatial-domain and frequency-domain features from image and then fuse them with textual features using multiple co-attention layers to learn the fine-grained correlation across modalities.
- **HMCAN** (Qian et al., 2021) utilizes hierachical hidden states of pre-trained BERT and pad-level features of the image to enhance the uni-modal representation, and further capture the inter-modality and intra-modality relationships by a contextual attention network.
- **CAFE** (Chen et al., 2022) reveals the inherent ambiguity across modalities, i.e., predictions from different modalities may contradict with each other. It dynamically adjusts the weights of uni-modal features and cross-modal features for the final decision based on the intensity of cross-modal ambiguity.

A.2 Data Statistics

We employ the widely used Chinese dataset Weibo (Jin et al., 2017) and the English dataset GossipCop (Shu et al., 2020) for evaluation. In addition, we additionally use the Chinese dataset Weibo21 (Nan et al., 2021) to study the generalization ability of our method. For Weibo, it contains 2776 real news and 3275 fake news for training, 825 real news and 816 fake news for testing. For GossipCop, considering the significant imbalance between positive and negative samples in GossipCop (more than 80% are real news), we retained all fake news and performed down-sampling on real news to achieve a balanced data distribution. After that, GossipCop contains 2036 real news and 2036 fake news for training, 545 real news and 545 fake news for testing. We keep the original train-test split for both Weibo and GossipCop. For Weibo21, it contains 4640 real news and 4487 fake news without original train-test split. Therefore, we randomly split it into training set and testing set in an 8:2 ratio.

A.3 Implementation Details

We use *clip-vit-large-patch14* for visual-enhanced module and both *Llava-onevision-0.5B* and *Qwen2-vl-2B-Instruct* as the backbone sLVLM. For rationale-augmented module, we employ model *Qwen2-vl-72B-Instruct*. We use the batch size of 8 and train the model using AdamW (Loshchilov et al., 2017) with an learning rate of 1e-4. The model is trained for 100 epochs with an early stop strategy to avoid overfitting. Low-Rank Adaptation (LoRA) (Hu et al., 2021) is utilized when fine-tuning the backbone sLVLM with the rank $r = 8$. The fine-tuning layer num K in the visual-enhanced module is set to 3. In addition, for fair comparison and to mitigate the limitations of outdated uni-modal pre-trained feature extractor on the baseline models' capability upper bound, we uniformly employ the same pre-trained model to extract the preprocessed textual and visual features and adaptively align the dimensions of the features with each baseline's original requirements. For textual modality, we utilize the 'bert-base-chinese' model for Weibo and Weibo21 and the 'bert-base-uncased' model for GossipCop. For visual modality, we also use the *clip-vit-large-patch14*. Both pre-trained models are kept frozen during the training. All the baseline methods are open-source and we use the code published to conduct the experiments. All the methods are implemented on Pytorch (Paszke et al., 2019) and trained on the NVIDIA RTX 3090 GPU.

B Specific Settings in Section 3.2

First, we will introduce the selected large vision-language models for evaluation. For open-source models, we choose the most advanced versions of the *Qwen-vl* and *Llava* series, i.e., *Qwen2-vl* (Wang et al., 2024) and *Llava-onevision* (Li et al., 2024). Additionally, we also test the closed-source model *Qwen-vl-max* (Bai et al., 2023). Then, we introduce the specific prompts we used, we use prompt $\mathcal{P}_1$ to evaluate the overall zero-shot performance of LVLMs for multimodal fake news detection. Specifically, $\mathcal{P}_1$ = *"Your task is to act as a fake news detection model. Given a news article and a related image, you need to determine the authenticity of the news. Output 0 for real news and 1 for fake news. Please only output your prediction without any additional information. News image:<image>, News content:<text>. Next, please output your prediction directly:"*

We use $\mathcal{P}_2$ to guide the LVLMs to only give a explicit prediction for samples with definitive clues from a predefined high-level perspective. Specifically, $\mathcal{P}_2$ = *"You need to act as a fake news detection model. Given a news article and a related image, you need to assess the authenticity of the news* *based on the following criteria: whether the news description aligns with common sense, science, and logic. Output 0 for real news, 1 for fake news, and 2 if uncertain. Note that we require a very high accuracy, and if there are no clear clues, output 2. Please only output your prediction without any additional information. News image:<image>, News content:<text>. Next, please output your prediction directly:"*

C Generating Multi-perspective Rationales

We first discuss how to choose LVLM as the agent model. Regarding model size, the results in the section 3.2 indicate that a 7B model still lacks sufficient knowledge-based reasoning and instruction-following abilities. Therefore, we need to select a larger model. Additionally, due to the security restrictions of closed-source commercial models, some news samples may trigger the model's security mechanisms, preventing it from generating valid content. As a result, we ultimately use the open-source model *Qwen2-vl-72B* as the agent model to generate the analysis.

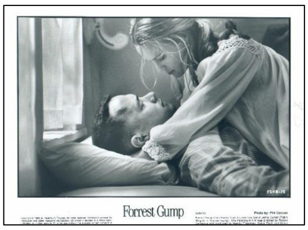
Regarding the design of specific prompts, considering that the fine-tuned model excels at capturing features from a micro perspective, we guide the agent model to complement it from a macro perspective. In order to fully utilize the agent model's wide knowledge base and its deep image-text understanding capability, we propose guiding the agent model to generate rationales from two angles: common-sense analysis and image-text coherence. Specifically, we have the common-sense analysis prompt $\psi_{\mathcal{R}_1}$ = *"<image> News content: <text> Please carefully analyze the given news and the accompanying image. Does the news content align with common sense, follow science and logic? Provide a brief analysis, keeping it under 50 words."*. From the perspective of image-text coherence, we have the prompt $\psi_{\mathcal{R}_2}$ = *"<image> News content: <text> Please carefully analyze the given news and the accompanying image. Do the scene described in the text and the image align? Can they support each other? Provide a brief analysis, keeping it under 50 words."* Although analyses from these two aspects may not cover all higher-level clues for fake news detection, we point out that they are more universal and general compared to other aspects. In the appendix E, we present several output examples from the agent model.

D Prompts for Abstracting Rationales

For a given fake news, it may contain obvious common-sense errors, but the image and text might match, as the image could have been manipulated through Photoshop. This can create conflicting analysis, leading to ambiguity in the model's judgment. At the same time, the initial analysis output by the agent model may still contain redundant content, which adds extra inference burden for the whole system. Therefore, to streamline the analysis and avoid introducing noise, we further use the prompt $\psi_{\mathcal{S}}$ to extract key statements that suggest the news might be fake while filtering out other information. Specifically, we have $\psi_{\mathcal{S}}$ = *"Given a multi-perspective analysis of a news article, please extract the key statements that suggest the news might be fake, and filter out the other information. If there are no such statements, output 'None'. Below is the news analysis: (1) common-sense analysis: <*$\mathcal{R}_1$*> (2) image-text coherence: <*$\mathcal{R}_2$*>".*

E Case Studies

In this section, we present several output examples of the agent model, as illustrated in Figures 5 to 10. In Cases 1 and 2, we demonstrate that the agent model effectively detects fake news and provides sound rationales based on both common-sense reasoning and image-text coherence. We highlight that this requires high demands on the model's image-text understanding ability and knowledge storage capacity. In Case 1, the model can detect errors from a scientific perspective as well as identify issues from the lack of relevance between the image and text. In Case 2, the model can deduce errors from contradictions between the news and established facts, as well as from the lack of relevance between the image and text. In Cases 3 and 4, we showcase how the rationale refinement module filters out noisy analysis. Even when the news articles include images that align with the text due to Photoshop manipulation, the model can still identify issues from a common-sense perspective. Finally, In Cases 5 and 6, we show the modal's output for real news and we find that the model can also provide high-quality analysis.



News content: *"Couples Who Can Maintain Eye Contact for Five Seconds Might Consider Marriage" According to Scientific American, a series of human physiological studies suggest that if a man and a woman make eye contact for less than one second, it indicates a lack of mutual affection. Eye contact lasting two seconds shows there is some liking; three seconds may suggest the beginning of romantic feelings; four seconds indicates a deep emotional connection; and if they maintain eye contact for five seconds or more, it could signify that they are ready to take the step toward marriage.*
Common-sense analysis: It goes against common sense, as there is no scientific basis for a direct correlation between the duration of eye contact, emotional depth, and marriage decisions.
Image-text coherence: The image is unrelated to the text. The image is a still from the movie Forrest Gump, while the text discusses judging emotional depth through eye contact.
Summary: (1) It goes against common sense, as there is no scientific basis for a direct correlation between the duration of eye contact, emotional depth, and marriage decisions. (2) The image is unrelated to the text. The image is a still from the movie Forrest Gump, while the text discusses judging emotional depth through eye contact.

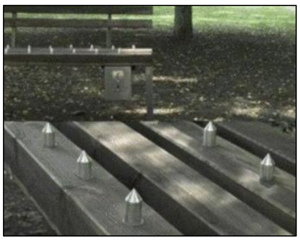
Figure 5: Rationales of Case 1, which is a fake news.

News content: *Nicolas Cage died due to a skiing accident on January 17, 2013.*
Common-sense analysis: Nicolas Cage actually did not pass away in 2013, and this information does not align with public records. Therefore, the news content goes against common sense and the facts.
Image-text coherence: The image and text are inconsistent. The person in the image matches the description of Nicolas Cage mentioned in the news headline, but the news states that he passed away due to a skiing accident, while the image does not show any scenes related to skiing or an accident.
Summary: (1) This information does not align with public records. Therefore, the news content goes against common sense and the facts. (2) The image and text are inconsistent, the news states that he passed away due to a skiing accident, while the image does not show any scenes related to skiing or an accident.

Figure 6: Rationales of Case 2, which is a fake news.

News content: *Yesterday, the Vienna subway caught its one-millionth fare evader, and to celebrate, they gave this person 100,000 euros!! After deducting the fine for the fare evasion, he still have 99,087 euros left! The fare evader was crying on the spot! Couldn't stop the tears of joy, right? The mayor, in a TV interview, claimed: "Fare evaders have made significant contributions to Vienna's economy. For example, Line 5 of the subway was funded by fines, and it's time we give something back."*
Common-sense analysis: The news content goes against common sense and logic, as fare evaders are typically fined, not rewarded. Moreover, using fines to fund public projects does not mean fare evaders should be rewarded.
Image-text coherence: The image and the text in the news confirm each other. The image shows a Vienna subway station and subway, which aligns with the news about the Vienna subway catching its one-millionth fare evader. The text describes the celebration and reward details, which are consistent with the content depicted in the image.
Summary: The news content goes against common sense and logic, as fare evaders are typically fined, not rewarded.

Figure 7: Rationales of Case 3, which is a fake news.



News content: *Yantai Park has coin-operated benches. If you insert a coin before sitting, the nails will retract.*
Common-sense analysis: It goes against common sense. Parks are public facilities, and it is unusual to charge for using basic amenities like benches. This design is neither practical nor in line with the principles of openness and shared use in public spaces.
Image-text coherence: The news text describes coin-operated benches at a park in Yantai, where you need to insert coins to sit down, and the image shows benches with protrusions resembling nails, which aligns with the description in the text.
Summary: It goes against common sense. Parks are public facilities, and it is unusual to charge for using basic amenities like benches.

Figure 8: Rationales of Case 4, which is a fake news.

News content: *A swarm of crayfish washed ashore, turning the seawater red in New Zealand. According to foreign media reports, recent sightings by residents near Dunedin on the Otago Peninsula revealed waters that were vividly red. Upon closer inspection, an astonishing number of crayfish was found, washed ashore by the tides.*
Common-sense analysis: The news content aligns with common sense, as it is a possible natural phenomenon for crayfish swarms to be washed ashore, turning the seawater red.
Image-text coherence: The news and the image confirm each other. The image shows a red area in the seawater, which aligns with the description in the news of crayfish swarms being washed ashore, causing the water to turn red.
Summary: None.

Figure 9: Rationales of Case 5, which is a real news.

News content: *According to French media, a series of shootings occurred on the evening of November 13th at locations including the Bataclan Theatre and a restaurant in Paris, and an explosion took place near the Stade de France. At least 60 people were killed, and it is estimated that several hundred hostages were taken inside the Bataclan Theatre. French President François Hollande said this was an unprecedented terrorist attack in France and declared a nationwide state of emergency.*
Common-sense analysis: The news content aligns with common sense and logic, following scientific and logical principles. A state of emergency is a special measure taken by the government in response to major crises, and France declaring a nationwide state of emergency after a severe terrorist attack is a reasonable response.
Image-text coherence: The image shows a scene outside a stadium, with a large crowd gathering, which aligns with the explosion incident near the Stade de France mentioned in the news. The scene in the image may be depicting the event described in the news.
Summary: None.

Figure 10: Rationales of Case 6, which is a real news.