Cross-modal Inconsistent Information-based Network for Fake News Detection

Aohan Li, Jiaxin Chen*, Dengyong Zhang, Yuling Liu, Yihui Luo, Xin Liao

***Abstract*—The rapid advancement of social media platforms has significantly reduced the cost of information dissemination, yet it has also led to a proliferation of fake news, posing a threat to societal trust and credibility. Most of fake news detection research focused on integrating text and image information to represent the consistency of multiple modes in news content, while paying less attention to inconsistent information. Besides, existing methods that leveraged inconsistent information often caused one mode overshadowing another, leading to ineffective use of inconsistent clue. To address these issues, we propose a cross-modal inconsistent feature fusion network (CIFF-Net). Inspired by human judgment processes for determining truth and falsity in news, CIFF-Net focuses on inconsistent parts when news content is generally consistent and consistent parts when it is generally inconsistent. Specifically, CIFF-Net extracts semantic and global features from images and texts respectively, and learns consistency information between modes through a multiple feature fusion module. To deal with the problem of modal information being easily masked, we design a single modal feature filtering strategy to capture inconsistent information from corresponding modes separately. Finally, similarity scores are calculated based on global features with adaptive adjustments made to achieve weighted fusion of consistent and inconsistent features. Extensive experimental results demonstrate that CIFF-Net outperforms state-of-the-art methods across public fake news datasets derived from real social medias.**

***Index Terms*—Image forensics, fake news detection, cross-modal information learning.**

I. Introduction

RECENTLY, social media platforms like Weibo and Facebook have basically replaced the traditional information dissemination methods represented by newspapers and magazines. While these platforms provide benefits such as low-cost and real-time information transmission, they have also facilitated the rampant use of fake images and news [1], [2]. Fake news refers to deliberately fabricated information that can be proven false, and its dissemination will erode public trust in public institutions and negatively affect social security. Automated fake news detection technology plays a crucial role in quickly identifying fake news and effectively mitigating their impact on public opinion.

Fake news detection primarily focused on single modal content in news, such as text [3], [4] or images [5], [6]. Early researchers found significant differences in the semantic styles between real and fake news. That is, fake news usually contain fewer transitive verbs and have a large number of emotional words, shorter words, and longer sentences. This study of single modal content used the semantic features of news and considered additional information such as writing style and image processing traces. Although single modal methods is effective in fake news detection, news content spread on social media platforms nowadays often includes multiple interconnected modalities, such as text and visuals. Such methods completely ignored the contribution of the remaining modal content and the correlation between various modalities. Subsequently, researchers shifted their attention to multi-modal feature fusion, which involved combining features from various modalities through concatenation, function mapping [7], [8]. These methods have some performance improvement compared to methods that only exploited single modal content. However, the multi-modal feature fusion approaches are overly simplistic, and the features of each modality are still independent, with no way to facilitate information exchange to promote inter-modal information understanding.

Some fake news detection technologies used different pre-trained models to extract different modal features and integrated them using attention mechanisms [9]–[11]. These technologies analyzed inter-modal consistency, i.e., aligning the semantics of images and text to generate fused features, thereby improving the accuracy of fake news detection. However, Ying et al. [12] found that inter-modal consistency information does not necessarily play a critical role and even sometimes lead to incorrect decisions, especially easy to recognize news with high degree of text-image correlation as real news. Zhang et al. [13] conducted an in-depth study on four types of image-text similarities, namely text similarity, semantic similarity, context similarity and post-training similarity. They found that in most cases, the image-text similarity of fake news was higher than that of real news. Many fake news can mislead readers precisely because of the high correlation between images and text. Moreover, since text and images in fake news may be fabricated, parts of the news that are less or not related between multiple modalities are more likely to contain clues inconsistent with common sense, which can be used to reveal the authenticity of the news. Therefore, while paying attention

This work is supported in part by the New Generation Artificial Intelligence National Science and Technology Major Project under Grant 2025ZD0123601, the National Natural Science Foundation of China under Grants 62402062 and U22A2030, Hunan Provincial Key Research and Development Program under Grant 2025AQ2022, Hunan Provincial Funds for Distinguished Young Scholars under Grant 2024JJ2025, Natural Science Foundation of Hunan Province, China under Grant 2025JJ60415, Research Foundation of Education Bureau of Hunan Province, China under Grant 25B0221, the Science and Technology Innovation Program of Hunan Province under Grant 2025QK2008, National Key Research and Development Program of China under Grant 2024YFF0618800.

Aohan Li, Xin Liao and Yuling Liu are with the College of Cyber Science and Technology, Hunan University, Changsha 410082, China.

Jiaxin Chen and Dengyong Zhang are with the School of Computer Science and Technology, Changsha University of Science and Technology, Changsha 410114, China.

Yihui Luo is with Hunan Department of Human Resources and Social Security, Changsha 410004, China.

*Corresponding author: Jiaxin Chen (jxchen@csust.edu.cn).

(a) Amazing! This dog is split in two, but it's still alive.

(b) The newly built Rose Garden is in a mess. A group of county residents climbed over the railings and got in, wantonly trampling on and picking the rose branches.

Fig. 1. Examples of mimicking human discrimination of news authenticity. (a) is fake news, in which the similarity between the text and images of the news is generally high, as indicated by the green square framing the image. However, the areas marked with red boxes in the image and highlighted in red text indicate parts where the correlation between text and image is low. The presence of a piece of wood reveals that the news is false. (b) is real news, in which the overall similarity between the text and images of the news is relatively low, as indicated by the red square surrounding the image. Nevertheless, the green boxes in the image and the corresponding sections in green text highlight areas with higher similarity. That is, the messy rose petals and broken branches on the ground support the authenticity of the news.

to the inter-modal consistency, the inter-modal inconsistency should also be considered, as these are two features of equal importance in fake news detection.

To address the aforementioned issues, we propose an innovative cross-modal inconsistent feature fusion network (CIFF-Net) that leverages the consistency and inconsistency information of multi-modal data to achieve fake news detection. When manually assessing the authenticity of a news, if the consistency of content across various modalities is high, we will pay more attention to whether there are inconsistencies or less relevant parts between modalities. Conversely, when the content across modalities is less relevant, we rely on the consistency between modalities to determine the news authenticity. For instance, as illustrated in Fig. 1 (a), the similarity between the textual and visual content of this news item is relatively high. However, the region enclosed by the red box and the corresponding text highlighted in red represent areas where the correlation between the image and text is relatively weak. The image depicts a piece of wood, which contradicts the news claim that the puppy was broken into two pieces. This discrepancy indicates that the news is fake news. In contrast, as shown in Fig. 1 (b), the overall similarity between the text and images in this news is relatively low. However, the area marked by the green box and the corresponding text in green font exhibit a strong correlation. These elements depict scattered rose petals and broken branches on the ground, supporting the authenticity of the news.

Thus, CIFF-Net extracts consistency information between modalities and inconsistency information within each modality, and promotes adaptive fusion of feature information through similarity weighting, thereby enhancing the fake news detection performance of CIFF-Net. Specifically, we first exploit the pre-trained SWIN Transformer (SWIN-T) and BERT to capture semantic information from images and texts. Meantime, the pre-trained CLIP is adopted to extract global information from images and texts. Then, we learn inter-modal consistency information through a multiple feature fusion module, and capture inconsistency information of corresponding modalities in two single modal branches. Finally, cosine similarity score of global features is calculated to adaptively adjust the use of each obtained detection information. Notably, the inconsistency captured in CIFF-Net operates at a global level, capturing inconsistency relationships across different components within a modality. In contrast, most existing studies focus on local-level inconsistencies, such as entity-level discrepancies, which are aggregated only at specific entity and thus differ significantly from our CIFF-Net in both scope and granularity level. The contributions are summarized as follows:

- We propose a novel strategy for feature filtering based on intra-modal inconsistency score vectors. Unlike existing methods that aggregate multi-modal information before extracting inconsistent features, CIFF-Net performs inconsistency extraction separately for each modality. This allows effective removal of noise and redundant information within each modality, thereby enhancing the utilization rate of single-modal inconsistency information and preventing potential masking effects among modalities.
- We design a multiple feature fusion module based on a common attention mechanism, which performs fine-grained fusion of semantic information from images and text through multiple pairs of common attention blocks. Meanwhile, it enhances the fused features using the semantic information from each modality to capture consistency features among multiple modalities in news. These consistency features is able to enhance information interaction between modalities, and promote information understanding.
- We conduct extensive experiments on several publicly available datasets and compare our CIFF-Net with various state-of-the-art fake news detection techniques. The experimental results demonstrate that CIFF-Net achieves superior performance. Moreover, ablation studies confirm that both intra-modal inconsistency information and inter-

modal consistency information contribute positively to the overall effectiveness of CIFF-Net, providing robust support for fake news detection.

II. Related Work

A. Single Modal based Detection Method

In the early stages of research on fake news detection, scholars primarily concentrated on single modalities within news content, specifically, either textual information [14] or visual elements [15]. In text-based detection studies, Ma et al. [3] employed sentences from news as input and utilized hidden layer vectors from recurrent neural networks to represent news information. Subsequently, Yu et al. [16] extended their investigation by incorporating convolutional neural networks (CNNs) to model textual content. Nevertheless, news content encompasses more than just textual information, it also includes rich visual components. In image-based detection approaches, Jin et al. [5] identified significant differences in visual characteristics, such as color distribution and scene complexity, between images accompanying real and fake news. These discrepancies serve as a valuable basis for fake news detection. Furthermore, Qi et al. [6] introduced a fake image discriminator that extracted both spatial semantic features and frequency-domain tampering features. By fusing these features through a dual-channel architecture, they achieved enhanced detection performance.

Additionally, pre-trained large models have emerged as powerful tools for fake news detection. For instance, the development of the textual feature extractor BERT [17] and the visual feature extractor Swin-T [18] has enabled researchers to extract high-quality features across modalities with greater efficiency. However, as methods of news fabrication become increasingly sophisticated, the limitations of single-modality detection approaches have become apparent. Specifically, such methods struggle to capture cross-modal consistent patterns, highlighting the need for multi-modal integration.

B. Multi-modal based Detection Method

Although traditional feature concatenation approaches, such as the SpotFake method [19], merely integrate textual features extracted via BERT with visual features obtained through VGG19, they fall short in capturing deep inter-modal correlations. To address this limitation, the authors subsequently proposed an enhanced variant, SpotFake+ [7], which substitutes XLNET for BERT as the text feature extractor and incorporates additional feature transformation procedures to facilitate more effective feature fusion. While this approach achieves a certain degree of improvement in detection performance, its feature fusion mechanism remains confined within a linear framework and is insufficient for fully exploring the intrinsic relationships among multi-modal data.

To better manage multi-modal information, Jin et al. [20] pioneered the use of an attention mechanism to enhance multi-modal representation, aiming to improve the understanding and utilization of inter-modal relationships. Chen et al. [9] introduced a cross-modal alignment auxiliary task to assess modality matching and applies the resulting scores as weights for both single-modal and multi-modal fused features. However, the effectiveness of fake news detection heavily depends on the accuracy of this auxiliary task. Zhou et al. [10] presented a multi-granularity fusion strategy that integrates coarse and fine-grained features from both image and text modalities, thereby capturing comprehensive hierarchical information. Ying et al. [12] proposed a guided multi-view representation learning framework, enabling learning across different modal features through single-modal prediction and cross-modal consensus learning. Zhang et al. [21] developed an adaptive knowledge-aware approach for fake news detection, which constructs an knowledge graph and integrates multi-modal content to model complex semantic and knowledge-level correlations. Zhang et al. [22] mapped multi-modal news content into a unified natural language space to address modality representation discrepancies. However, this approach may lead to significant loss of visual information, as rich semantic and detailed image content cannot be fully captured in textual form. Lao et al. [23] proposed a spectral-based fusion framework that effectively integrates multi-modal features through operations including single-modal spectral compression and cross-modal spectral co-selection. Du et al. [11] instigated a shift in text representation conditioned on auxiliary features to balance efficiency and detection accuracy.

Despite these advancements, most existing methods exhibit a preference for modal consistency and are inadequate in exploiting semantic conflicts between textual and visual modalities. Li et al. [24] attempted to comprehensively capture false information by integrating inconsistent features across different modalities extracted using distinct methods. However, during the fusion process, the inconsistent features of one modality may be overshadowed by those of another, leading to the loss of critical information. In contrast, our proposed method enables adaptive adjustment of the contributions of both consistent and inconsistent features, thereby facilitating the learning of optimal cross-modal discriminative patterns.

III. Network Architecture

A. Overview

The core idea of our CIFF-Net is grounded in cognitive psychological principles governing how humans assess the authenticity of multimodal news. We found that human judgment does not follow a rigid, linear process, rather, it employs a dynamic, adaptive dual-cue strategy, precisely the cognitive capacity enabling robust identification of highly realistic fake news. This judgment comprises two complementary modes: In cases where image and text exhibit high overall semantic consistency, humans do not infer truthfulness outright. Instead, they heighten vigilance and selectively attend to subtle, incongruent, or counterintuitive discrepancies between modalities, because sophisticated disinformation often exploits surface-level alignment to deceive. Identification of such conflict cues is thus critical to exposing forgery. Conversely, when image-text alignment is markedly low, humans shift attention toward identifying shared core semantic elements, specifically, consistent event subjects, temporal anchors, and scene descriptors. Absence of agreement on these foundational elements strongly indicates fabrication.

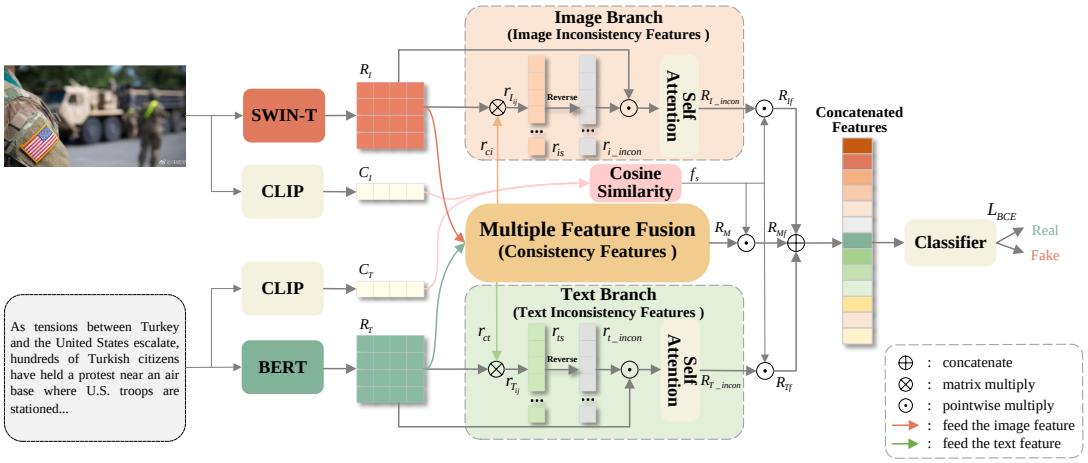


Fig. 2. The overview of our CIFF-Net. The SWIN-T, BERT, and CLIP models are employed to extract single modal features. Then, three parallel branches are utilized to extract the inconsistent and consistent information. Finally, the cosine similarity is computed to regulate the contribution degree of each feature, and the real and fake news are predicted by the classifier.

Conventional multimodal fake news detection methods (such as SpotFake+ [7], MMFN [10], and CFFN [24]) primarily prioritize cross-modal semantic consistency alignment as their central optimization objective, and intra-modal inconsistency is typically treated as ancillary information or, in many cases, entirely disregarded. While certain existing methods (such as CAFE [9] and BMR [12]) do incorporate inconsistency-aware modeling, they commonly extract inconsistency features only after multimodal feature fusion. This sequential design risks modal suppression: dominant features from one modality may obscure or attenuate salient inconsistency signals from another. To mitigate this limitation, our proposed CIFF-Net treats both cross-modal consistency and intra-modal inconsistency as complementary and equally discriminative signals. CIFF-Net jointly leverages consistency cues derived from inter-modal alignment and inconsistency cues arising from intra-modal discrepancies to enhance detection robustness. Moreover, we introduce a single modal feature filtering strategy. This strategy computes inconsistency scores independently within each modality's native feature space and performs feature filtering prior to fusion, thereby preserving the integrity and discriminability of modality-unique inconsistency patterns.

The network architecture of CIFF-Net is depicted in Fig. 2. Firstly, three single modal feature extractors are employed to extract the semantic features $\{R_I, R_T\}$ and global features $\{C_I, C_T\}$ of images and texts. Subsequently, the image inconsistent feature R_{I_incon}, text inconsistent feature R_{T_incon}, and inter-modal consistency feature R_M are learned via the image branch, text branch, and multiple feature fusion module, respectively. Most state-of-the-art methods rely on static feature fusion schemes, wherein fusion weights remain invariant across samples. Our CIFF-Net instead adopts a dynamic, sample-adaptive weighting strategy: the relative contributions of consistency and inconsistency features are modulated in real time based on the text-image semantic alignment score, ensuring that fusion parameters reflect the intrinsic cross-modal coherence (or discordance) of each individual news instance. Subsequently, the weighted cross-modal information is fed into the fake news classifier to enhance the accuracy of determining news authenticity.

B. Single Modal Feature Extraction

Text modal semantic feature extraction: We utilize the BERT pre-trained model [17] to extract text semantic feature R_T of news. BERT is a pre-trained natural language processing model based on the Transformer architecture. Its bidirectional approach allows it to consider both preceding and following text, enabling a more comprehensive and accurate understanding of the input. Additionally, BERT is pre-trained on a large corpus, acquiring rich language representations with strong transferability. The output from BERT's last hidden layer provides a detailed representation of context features, encompassing not only the semantic information of individual tokens but also the surrounding context. This makes BERT highly effective in understanding and processing text. Therefore, we use the BERT pre-trained model to process news text, and the output from BERT's last hidden layer is exploited as the extracted text semantic feature R_T. Specifically, we first convert the news text into a continuous list of words $T = [t_1, t_2, \ldots, t_n]$, where n is the number of words. Subsequently, T is input into the BERT pre-trained model to obtain the text modality feature $R_T = [r_{t_1}, r_{t_2}, \ldots, r_{t_n}]$, where r_{t_i} represents the feature vector corresponding to each word slice t_i output by the last hidden layer of BERT. Through this method, we can effectively extract deep semantic feature of news text, providing reliable text feature support for subsequent fake news detection.

Image modal semantic feature extraction: We use the SWIN-T pre-trained model [18] to extract image semantic feature R_I of news. SWIN-T is a Transformer-based model for visual processing that introduces a hierarchical Transformer to extract features layer-by-layer, and captures both local and global features through a sliding window mechanism, which achieves superior performance in computer vision tasks such

as image classification and semantic segmentation. We introduce SWIN-T to handle the image modal of news, exploiting the output of the last Transformer layer as the image modal semantic feature R_I. Through this method, we can effectively extract multi-scale features of news image, providing reliable visual feature support for fake news detection.

Single modal global feature extraction: We use the CLIP pre-trained model [25] to extract global features C_I and C_T for both image and text modalities simultaneously. The CLIP model achieves text and image embedding through multi-modal feature encoding, mapping them into a unified mathematical space. This feature provides significant advantages when calculating inter-modal consistency. CLIP has been pre-trained on a vast and diverse dataset and demonstrates strong robustness, especially in handling distribution variations and zero-sample learning scenarios. In fake news detection, CLIP can effectively be applied to detect unseen news data [26]. We choose to use the CLIP model to extract text global feature C_T and image global feature C_I for subsequent calculation of inter-modal consistency scores based on cosine similarity. The similarity scores calculated from the two global features mapped to a unified mathematical space adaptively adjust the weights of each feature in fake news detection, which has a significant improvement effect on guiding the subsequent classifier to predict news authenticity.

C. Interactive Learning of Inter-modal Consistency

Since BERT and SWIN-T are not multi-modal models, there exists a considerable gap between text semantic feature R_T and visual semantic feature R_I. If information fusion is conducted directly, the information between the modes cannot be interactively comprehended, and effective information between the modes cannot be learned. To address this issue, we proposed a multiple feature fusion module based on the Co-attention mechanism [27] for extracting inter-modal consistent features. Fig. 3 depicts the structure of the multiple feature fusion module. The Co-attention mechanism mainly consists of a multi-head attention and feed forward layer, with residual connections and layer normalization behind each component. Through multi-head attention, it becomes feasible to focus on the valid information of text and image features simultaneously and establish associations between them. The feed forward layer can further process these correlation features and achieve more efficient information interaction when fusing different modal information. Residual connections and layer normalization can enhance the stability of training and strengthen the representation of features. As a result, by employing the Co-attention mechanism, the text feature R_T extracted by BERT and the image feature R_I extracted by SWIN-T can be effectively fused to enhance the understanding of inter-modal information and better learn inter-modal consistency features. Simultaneously, to prevent overfitting, we share weights between a pair of Co-attention blocks.

When fusing features through the first pair of Co-attention, the multi-head attention of each Co-attention block has H

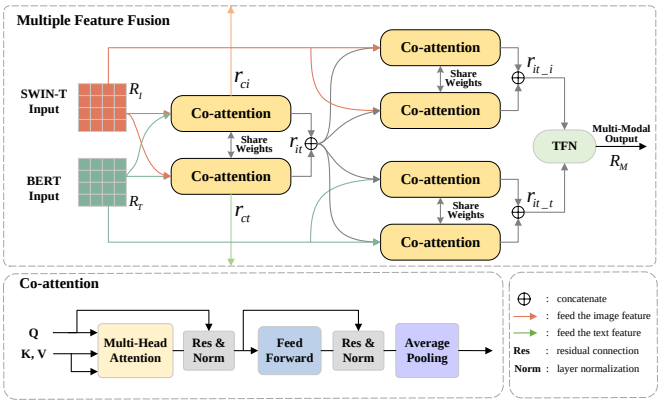


Fig. 3. Illustration of our multiple feature fusion module. From left to right and from top to bottom, the first pair of Co-attention fuses the semantic features of text and image, the second and third pair of Co-attention are used to enhance the fused feature r_{it}, and the final inter-modal consistency information R_M is obtained through the TFN fusion strategy.

heads. Mapping different modal features R_T and R_I to the same feature space by linear transformations, for each head h,

$$\begin{cases} Q_T^h = R_T W_Q^{T,h} & K_T^h = R_T W_K^{T,h} \quad V_T^h = R_T W_V^{T,h} \\ Q_I^h = R_I W_Q^{I,h} & K_I^h = R_I W_K^{I,h} \quad V_I^h = R_I W_V^{I,h} \end{cases} \tag{1}$$

where $Q_T^h, K_T^h, V_T^h \in \mathbb{R}^{n \times d_k}$, $Q_I^h, K_I^h, V_I^h \in \mathbb{R}^{p \times d_k}$, d_k represents the dimensions of the query(Q), key(K), and value(V) vectors for each head in the multi-head attention. $W_Q^{T,h}, W_K^{T,h}, W_V^{T,h} \in \mathbb{R}^{d_t \times d_k}$ and $W_Q^{I,h}, W_K^{I,h}, W_V^{I,h} \in \mathbb{R}^{d_i \times d_k}$ are the learnable weight matrix for each head.

Then, the attention weight of a single attention head is calculated as follow:

$$\begin{cases} A_{TI}^h = softmax(\dfrac{Q_T^h (K_I^h)^T}{\sqrt{d_k}}) \\ A_{IT}^h = softmax(\dfrac{Q_I^h (K_T^h)^I}{\sqrt{d_k}}) \end{cases} \tag{2}$$

where $A_{TI}^h \in \mathbb{R}^{n \times p}$ and $A_{IT}^h \in \mathbb{R}^{p \times n}$.

The attention output of a single attentional head is calculated by the single attentional weight, which can be expressed as:

$$\begin{cases} R_T^h = A_{TI}^h V_I^h \\ R_I^h = A_{IT}^h V_T^h \end{cases} \tag{3}$$

The output of H attention heads in multi-head attention is concatenated and linear transformation is performed to obtain the final output feature of the multi-head attention.

$$\begin{cases} R_T' = concat(R_T^1, R_T^2, \ldots, R_T^H) W_O^T \\ R_I' = concat(R_I^1, R_I^2, \ldots, R_I^H) W_O^I \end{cases} \tag{4}$$

where $W_O^T \in \mathbb{R}^{Hd_k \times d_t}$ and $W_O^I \in \mathbb{R}^{Hd_k \times d_i}$ are linear transformation matrices.

The features R_T' and R_I' are updated through residual connections and layer normalization. Finally, the updated feature is fed to the feed forward layer to strengthen the correlation between features, followed by residual connections, layer normalization, and average pooling. The strengthened

features r_{ct} and r_{ci} are extracted as the output of each Co-attention block. The output of the two Co-attention blocks is fused to obtain the first fused multi-modal feature r_{it}.

Some studies have demonstrated that information enhancement in multi-modal fake news detection can improve detection accuracy and robustness [28]. To further assist CIFF-Net in enhancing its comprehension of the information among the modalities of text and image and to increase the feature richness, we respectively enhance the information of the first integrated multi-modal feature r_{it} with the text feature R_T and the image feature R_I through two pairs of Co-attention blocks (CA).

$$\begin{cases} r_{it\_t} = concat(CA(R_T, r_{it}), CA(r_{it}, R_T)) \\ r_{it\_i} = concat(CA(R_I, r_{it}), CA(r_{it}, R_I)) \end{cases} \tag{5}$$

Meanwhile, the TFN fusion strategy [29] can fully utilize various features for fusion. Hence, we employ the TFN strategy to integrate the two enhanced multi-modal features r_{it_t} and r_{it_i}. Specifically, the dimension expansion of r_{it_t} and r_{it_i} is conducted with 1 respectively, and then the Cartesian product of r_{it_t} and r_{it_i} is calculated as follow,

$$R_M = [r_{it\_t}; 1] \bigotimes [r_{it\_i}; 1] \tag{6}$$

where ";" is concatenation, and $\bigotimes$ is Cartesian product. R_M represents the inter-modal consistency feature output by the final multiple feature fusion module, used for subsequent adaptive weighting and classification.

D. Feature Filtering for Cross-modal Inconsistency Learning

Currently, the majority of fake news detection approaches based on multi-modal information directly concatenated and fused multiple single-modal information, causing information redundancy to a certain extent. Simultaneously, existing methods pay significantly less attention to the inconsistent information among modes compared to the consistent information between them. For the text branch and image branch depicted in Fig. 2, we propose a single modal feature filtering strategy to extract inconsistent features from the image and text modes. This strategy initially computes the inconsistency score and filters the features of the text and image modes based on the self-attention mechanism [30] to eliminate redundant information and useless noise. It should be noted that redundant information refers to the highly consistent features of text and image extracted by the multiple feature fusion module, which is inherently repetitive within the inconsistent feature representation and thus warrants attenuation and suppression, as it contributes no additional discriminative signal for decision-making and may exacerbate feature overfitting. While useless noise comprises isolated, semantically void regions exhibiting anomalously high inconsistency scores, such as image random noise, text segments degraded by compression artifacts, or syntactically peripheral elements (e.g., auxiliary words, punctuation, and non-lexical interjections). Critically, the elevated inconsistency scores in these regions stem not from genuine semantic conflict but from the absence of meaningful semantic alignment with the counterpart modality. Consequently, they constitute spurious signals that risk misleading model inference and must therefore be actively suppressed.

This single modal feature filtering strategy improves the utilization rate of inconsistent information and avoids the shading effect of the inconsistent information in one mode on that in another mode. Specifically, for redundant information filtering, in regions exhibiting strong semantic alignment with the other modality, initial inconsistent feature weights are inherently low. we use the self-attention mechanism to further model inter-regional semantic dependencies, thereby attenuating weights assigned to such redundant regions and diminishing their representational influence. Concurrently, it enhances the weights of genuinely inconsistent regions that reflects authentic cross-modal semantic conflicts, thereby sharpening feature focus on discriminative discrepancies. For useless noise filtering, isolated, semantically incoherent noise regions lack meaningful contextual or semantic relationships with other feature locations. As a result, we assigns them negligible attention weights, directly suppressing the expression of such noise. Conversely, truly semantic conflict regions typically manifest as spatially or sequentially contiguous patterns closely tied to the core event structure, with close semantic correlations between regions. We accordingly assigns higher weights to these regions, reinforcing the representation of semantically grounded inconsistencies.

Taking image branch as an example, as shown in Fig. 2, the similarity matrix $r_{I_{ij}}$ is obtained by multiplying the image semantic feature R_I and the strengthened image feature r_{ci} weighted with text attention. In order to represent the inter-modal consistency score r_{is} of each block in the image feature, the similarity matrix $r_{I_{ij}}$ is summed by column and then normalized by softmax.

$$r_{is} = softmax(\sum_j {R_I}^T r_{ci}) \tag{7}$$

The consistency score vector r_{is} is then reversed to generate the inconsistency score vector, calculated as follows:

$$r_{I\_incon} = 1 - r_{is} \tag{8}$$

The higher the inconsistency score of the corresponding position, the less attention is paid to this area of the image during multiple feature fusion, which means that the corresponding position has inconsistency.

Then, the inconsistency score vector is weighted to the original image feature R_I, and the inconsistency feature representation is calculated. Since there may be some redundant information and interference noise in the obtained inconsistent feature, the self-attention mechanism (SA) is adopted to filter the obtained image inconsistent feature.

$$R_{I\_incon} = SA(r_{I\_incon} \bigodot R_I) \tag{9}$$

where $\bigodot$ is pointwise multiply.

As illustrated in the text branch, the text inconsistent feature R_{T_incon} can be captured in the same way as image inconsistent feature extraction.

E. Weight Adaptation for News Authenticity Classification

To better utilize the image inconsistency feature R_{I_incon}, text inconsistency feature R_{T_incon}, and inter-modal consistency feature R_M in the classifier, we adjust the contribution weights of each feature in fake news detection by computing the consistency score of the text global feature C_T and the image global feature C_I, guiding the classifier to learn useful information. Thus, the classifier can judge the truth of news more effectively. Specifically, we calculate the cosine similarity f_s of the image global feature C_I and the text global feature C_T extracted by the CLIP pre-trained model.

$$f_s = \frac{C_T \cdot (C_I)^T}{\|C_T\| \cdot \|C_I\|} \tag{10}$$

The key idea of CIFF-Net is to imitate human identification of the truth and falseness of news, that is, when the degree of similarity between images and texts is high, the detection model pays more attention to the parts where they are inconsistent, and when the degree of similarity is low, the model focuses more on the parts where they are consistent. Therefore, we first normalize the cosine similarity score f_s to be within the range of 0 and 1.

$$sim = \frac{1 + f_s}{2} \tag{11}$$

Then, we use sim to weight the image inconsistency feature R_{I_incon}, the text inconsistency feature R_{T_incon}, and the inter-modal consistency feature R_M.

$$\begin{cases} R_{If} = sim \cdot R_{I\_incon} \\ R_{Tf} = sim \cdot R_{T\_incon} \\ R_{Mf} = (1 - sim) \cdot R_M \end{cases} \tag{12}$$

where "$\cdot$" indicates element-wise multiplication operation. R_{Mf} is the final consistency feature after adaptive weighting.

Finally, the weighted inconsistency features and consistency feature are concatenated and fed into the fake news classifier, which consists of a fully connected layer and a ReLu activation function. The classifier will output the real-fake prediction result of the news.

The loss function of CIFF-Net is binary cross entropy loss,

$$L_{BCE} = -[ylog(y') + (1 - y)log(1 - y')] \tag{13}$$

where y and y' denote the real and prediction label.

It should be noted that the weight adaptation mechanism proposed in our method is a deterministic, non-learnable computational rule, rather than a trainable parameter optimized end-to-end by the model. Specifically, the weight assigned to the inconsistency feature increases monotonically with the text-image similarity score, whereas the weight assigned to the consistency feature is correspondingly reduced. This design is structurally grounded in the following principle: for news items exhibiting high text-image similarity, the contribution of the consistency feature is deliberately attenuated, while the inconsistency feature is accorded dominant weighting. Consequently, the model cannot exploit the consistency feature as a "shortcut" to minimize loss on high-similarity samples. Instead, it is compelled to learn discriminative cues for fake news detection exclusively from the inconsistency feature. Failure to do so would render accurate classification of such high-similarity instances impossible. In essence, the adaptive weighting mechanism itself imposes a strong architectural constraint: in high-similarity scenarios, reliance on the inconsistency feature is mandatory.

IV. Experimental Results

A. Experimental Setup

Datasets: We validate the performance of our CIFF-Net on five publicly available datasets sourced from social medias: Weibo [20], Weibo-21 [31], GossipCop [32], MCFEND [33], and MMFakeBench [34].

- The Weibo dataset: This is the most widely used Chinese dataset for fake news detection, containing data from verified fake and real news on Sina Weibo between 2012 and 2016. The training set includes 3,783 real news and 3,675 fake news, while the testing set contains 1,685 news.
- The Weibo-21 dataset: This is an updated version of the Weibo dataset published in 2021, containing more recent Weibo news. This dataset includes 4,640 real news and 4,487 fake news, which we divided into a training set and testing set in a 9:1 ratio.
- The GossipCop dataset: This is an English dataset. The training set includes 7,974 real news and 2,036 fake news, and the testing set includes 2,285 real news and 545 fake news.
- The MCFEND dataset: The total number of original samples in this dataset is 23,974, including 17,715 pieces of fake news and 6,074 pieces of real news. To ensure data quality, the original samples were preprocessed and cleaned. Then, 14,770 valid fake news samples and 4,220 valid real news samples were retained. To avoid the interference of category imbalance on the model training effect, 4,770 samples were randomly selected from the valid fake news and combined with all 4,223 valid real news samples to form the experimental dataset. The dataset was randomly divided into a training set and a testing set in a 8:2 ratio after equalization.
- The MMFakeBench dataset: The original sample of this dataset consists of 11,000 entries, among which there are 7,700 pieces of fake news and 3,300 pieces of real news. To ensure data quality, the original samples were preprocessed and cleaned. Then, 6,420 valid fake news and 2,887 valid real news were retained. To avoid the interference of category imbalance on the model training effect, 3,000 samples were randomly selected from the valid fake news and combined with all 2,887 valid real news to form the experimental dataset. Using a random division method, the balanced dataset was split into a training set and a testing set in a 8:2 ratio.

Implement details: We use the Adam optimizer for training, with a batch size set to 16, epochs set to 100, a learning rate set to 10^{-4}. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

TABLE I
COMPARISONS WITH SOTA METHODS ON FIVE REAL SOCIAL MEDIA TRANSMISSION NEWS DATASETS. THE HIGHEST VALUE IS **BOLD**.

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	Spotfake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	BMR	0.918	0.882	**0.948**	0.914	**0.942**	0.870	0.904
	MMFN	0.923	0.921	0.926	0.924	0.924	0.920	0.922
	CFFN	0.901	0.913	0.889	0.889	0.888	0.913	0.900
	FSRU	0.894	0.936	0.849	0.890	0.858	0.940	0.897
	CIFF-Net (our)	**0.930**	**0.951**	0.909	**0.930**	0.910	**0.952**	**0.931**
Weibo-21	Spotfake+	0.851	**0.953**	0.733	0.828	0.786	**0.964**	0.866
	CAFE	0.882	0.857	0.915	0.885	0.907	0.844	0.876
	BMR	0.929	0.908	0.947	0.927	0.946	0.906	0.925
	MMFN	0.930	0.941	0.920	0.930	0.919	0.940	0.929
	CFFN	0.914	0.926	0.899	0.912	0.902	0.928	0.915
	FSRU	0.905	0.941	0.860	0.899	0.875	0.948	0.910
	CIFF-Net (our)	**0.943**	0.934	**0.951**	**0.942**	**0.952**	0.935	**0.943**
GossipCop	Spotfake+	0.858	0.732	0.372	0.494	0.866	0.962	0.914
	CAFE	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	BMR	0.895	0.752	0.639	0.691	0.920	0.965	0.936
	MMFN	0.894	0.799	**0.698**	0.684	0.910	0.964	0.936
	CFFN	0.885	0.749	0.610	0.672	0.894	0.959	0.925
	FSRU	0.872	0.702	0.583	0.637	0.905	0.941	0.923
	CIFF-Net (our)	**0.911**	**0.838**	0.666	**0.742**	**0.924**	**0.969**	**0.946**
MCFEND	Spotfake+	0.840	0.862	0.795	0.827	0.821	0.882	0.850
	CAFE	0.851	0.828	**0.901**	0.863	0.879	0.803	0.839
	BMR	0.881	0.872	0.887	0.879	0.890	0.875	0.882
	MMFN	0.882	0.880	0.878	0.879	0.884	**0.915**	0.899
	CFFN	0.901	0.913	0.889	0.889	0.888	0.913	0.900
	CIFF-Net (our)	**0.903**	**0.893**	0.898	**0.895**	**0.912**	0.907	**0.909**
MMFakeBench	Spotfake+	0.780	0.798	0.726	0.760	0.767	0.831	0.798
	CAFE	0.795	0.768	0.802	0.785	0.821	0.788	0.804
	BMR	0.831	0.820	0.835	0.827	**0.870**	0.827	0.848
	MMFN	0.829	0.827	0.758	0.791	0.831	**0.895**	0.862
	CFFN	0.812	0.810	0.786	0.798	0.814	0.836	0.825
	CIFF-Net (our)	**0.872**	**0.879**	**0.857**	**0.868**	0.865	0.886	**0.875**

TABLE II
COMPARISON OF MODEL EFFICIENCY WITH SEVERAL SOTA METHODS.

Metrics	Spotfake+	CAFE	BMR	MMFN	CFFN	CIFF-Net (our)
Params	124.37M	0.68M	94.39M	202.61M	178.31M	194.61M
FLOPS	30.42G	0.01G	18.42G	43.65G	32.37G	38.95G

SOTA methods: We compare CIFF-Net with several state-of-the-art methods. That is, SpotFake+ [7], CAFE [9], BMR [12], MMFN [10], CFFN [24], and FSRU [23].

B. Comparisons with SOTA Methods

Comparison of fake news detection accuracy. We use accuracy, precision, recall, and F1 score as evaluation metrics for fake news detection. Table I provides the performance comparison results with state-of-the-art methods. It can be seen that our CIFF-Net achieves the highest detection accuracy and F1 score on each dataset, with most other evaluation metrics also ranking first or second. The superiority of our CIFF-Net can be attributed to several reasons. Firstly, we calculate similarity scores based on global features extracted from the CLIP model to adaptively weight the fine-grained semantic features extracted from the BERT and SWIN-T models. Particularly, our method achieves greater accuracy gains on the GossipCop dataset compared to other benchmarks. This is attributable to GossipCop's higher proportion of high text-image similarity news items, and thus greater intrinsic detection difficulty. On datasets dominated by low-similarity samples, consistency-based baselines already attain strong performance, leaving limited room for improvement. Conversely, on datasets enriched with high-similarity instances, the discriminative power of the inconsistency feature is more fully leveraged, yielding proportionally larger performance gains.

Secondly, we extract inconsistency information for each single modal feature by filtering redundant information and interference noise, effectively preventing inconsistencies in one modality from being overwritten and ensuring that consistency and inconsistency information from each modality effectively participate in fake news detection. Finally, in the multiple feature fusion module, after merging the semantic features of image and text modalities, we further enhance the fused feature using the semantic features of the corresponding modality, which promotes the interactive understanding between modality information, and captures more effective consistency information.

Comparison of model efficiency. We also compare the de-

TABLE III
ABLATION STUDY TO VERIFY THE EFFECTIVENESS OF THE CO-ATTENTION BLOCK. THE HIGHEST VALUE IS **BOLD**.

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	with iMMoE	0.910	0.949	0.869	0.907	0.876	0.952	0.912
	with LSTM	0.921	**0.966**	0.875	0.918	0.883	**0.969**	0.924
	with Co-attention (our)	**0.930**	0.951	**0.909**	**0.930**	**0.910**	0.952	**0.931**
Weibo-21	with iMMoE	0.918	0.925	0.906	0.915	0.911	0.929	0.920
	with LSTM	0.932	**0.955**	0.904	0.929	0.912	**0.959**	0.935
	with Co-attention (our)	**0.943**	0.934	**0.951**	**0.942**	**0.952**	0.935	**0.943**
GossipCop	with iMMoE	0.882	0.748	0.583	0.655	0.906	0.953	0.929
	with LSTM	0.903	**0.838**	0.615	0.709	0.914	**0.972**	0.942
	with Co-attention (our)	**0.911**	**0.838**	**0.666**	**0.742**	**0.924**	0.969	**0.946**

tection model efficiency of our CIFF-Net with SOTA methods. As shown in Table II, compared with the MMFN method with the closest detection accuracy, the total number of Params and FLOPS of our CIFF-Net have been reduced. Meanwhile, the average detection accuracy on three public datasets has been improved. This indicates that our parallel branch design does not lead to the accumulation of serial operation overhead. Instead, through structural optimization, it reduces the overall computational load and achieves the optimization of "lower overhead and higher performance". Although the CAFE method has extremely low Params and FLOPs, its accuracy in detecting fake news is on average 2.8% lower than our CIFF-Net, making it difficult to meet the precision requirements for fake news detection in social media scenarios.

Limitations. The fake news detection failures mainly occur in two distinct scenario categories. Firstly, for deeply ambiguous content, this category comprises satirical, ironic, or metaphorical news items. Such cases constitute a distinctive class of disinformation wherein "the literal proposition is true, but the conveyed meaning is false." Current pre-trained language-vision models exhibit inherent limitations in modeling nuanced pragmatic and irony-aware semantics-particularly in multilingual contexts (e.g., Chinese and English), thereby impairing their capacity to jointly reason about both semantic consistency and inconsistency cues. This limitation ultimately leads to erroneous classifications. Secondly, for highly realistic AI-generated content, this refers to synthetic multimodal disinformation in which AI-generated text and images are deliberately engineered for near-perfect semantic and visual coherence. In such instances, inconsistencies are exceptionally subtle, which often manifesting only as minute statistical artifacts or imperceptible rendering anomalies, without overt semantic contradictions. Although our model demonstrates strong performance in detecting conventional inconsistency patterns, its sensitivity to these "high-fidelity forged" cases can be further improved.

C. Ablation Study

Effectiveness of co-attention block. To systematically evaluate the necessity and effectiveness of incorporating Co-attention blocks within the multiple feature fusion module, we conduct ablation studies by replacing all Co-attention blocks with iMMoE networks [12] and LSTM networks [20], respectively. "with iMMoE" denotes the substitution of iMMoE network for Co-attention block, and "with LSTM" denotes the substitution of LSTM network for Co-attention block. Prior research has demonstrated that these two types of networks are effective in fusing multi-modal features for fake news detection. The experimental results are summarized in Table III. As shown in the table, the proposed method employing Co-attention blocks consistently achieves superior performance on the Weibo, Weibo-21, and GossipCop datasets. This provides strong evidence that the Co-attention block contributes more effectively to multi-modal feature fusion.

TABLE IV
ABLATION STUDY TO VERIFY THE EFFECTIVENESS OF EACH MODULE IN THE PROPOSED CIFF-NET. THE HIGHEST VALUE IS **BOLD**.

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	w/o IB	0.918	0.900	**0.944**	0.921	**0.939**	0.892	0.915
	w/o TB	0.904	0.881	0.937	0.908	0.930	0.870	0.899
	w/o FF	0.837	0.896	0.767	0.826	0.791	0.908	0.845
	w/o EF	0.909	**0.953**	0.862	0.905	0.871	**0.957**	0.912
	w/o Sim	0.905	0.897	0.917	0.907	0.912	0.892	0.902
	w/o TFN	0.922	0.949	0.896	0.922	0.899	0.951	0.924
	CIFF-Net	**0.930**	0.951	0.909	**0.930**	0.910	0.952	**0.931**
Weibo-21	w/o IB	0.927	0.917	0.935	0.926	0.936	0.918	0.927
	w/o TB	0.912	0.871	**0.964**	0.915	**0.962**	0.862	0.909
	w/o FF	0.844	0.796	0.920	0.854	0.909	0.772	0.835
	w/o EF	0.922	0.933	0.906	0.919	0.912	0.938	0.925
	w/o Sim	0.919	0.925	0.909	0.917	0.913	0.929	0.921
	w/o TFN	0.927	**0.966**	0.884	0.923	0.896	**0.970**	0.932
	CIFF-Net	**0.943**	0.934	0.951	**0.942**	0.952	0.935	**0.943**
GossipCop	w/o IB	0.901	0.800	0.646	0.715	0.919	0.961	0.940
	w/o TB	0.882	0.683	**0.721**	0.701	**0.933**	0.920	0.926
	w/o FF	0.823	0.544	0.495	0.518	0.882	0.901	0.891
	w/o EF	0.874	**0.859**	0.413	0.558	0.875	**0.984**	0.926
	w/o Sim	0.887	0.746	0.626	0.681	0.914	0.949	0.931
	w/o TFN	0.898	0.834	0.588	0.690	0.908	0.972	0.939
	CIFF-Net	**0.911**	0.838	0.666	**0.742**	0.924	0.969	**0.946**

Effectiveness of each module. We also perform an ablation study to validate the effectiveness of each module in the proposed CIFF-Net. "w/o IB" denotes our model without the image branch, "w/o TB" denotes our model without the text branch, "w/o FF" denotes our model without the multiple

TABLE V
ABLATION STUDY TO VERIFY THAT CIFF-NET OUTPERFORMS MULTI-MODAL BACKBONE NETWORKS.

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Weibo21	with CLIP	0.842	0.880	0.786	0.830	0.813	0.897	0.853
	with B+S	0.857	0.872	0.831	0.851	0.843	0.881	0.862
	with MFF	0.914	0.979	0.842	0.905	0.865	0.983	0.920
	CIFF-Net	0.943	0.934	0.951	0.942	0.952	0.935	0.943

feature fusion module, "w/o EF" denotes our model that does not use single modal features to enhance the multi-modal fused feature in the multiple feature fusion module, "w/o Sim" denotes our model that does not use the cosine similarity calculation. "w/o TFN" denotes our model that simple concatenation is used to replace the TFN fusion in the multiple feature fusion module.

Table IV gives the comparison results. Comparing the experimental results of "w/o IB", "w/o TB", and CIFF-Net, it can be observed that the performance of "w/o IB" and "w/o TB" has significantly decreased, indicating that a single modality's inconsistency information is insufficient to effectively support fake news detection. Meanwhile, the performance of "w/o TB" is slightly lower than that of "w/o IB", suggesting that the inconsistency information from the text modality has a greater impact on fake news detection. Comparing the results of "w/o FF", "w/o EF", and CIFF-Net, it can be seen that "w/o FF" has the largest decrease in performance, indicating that the inter-modal consistency information extracted by the multiple feature fusion module significantly contributes to performance improvement, and inter-modal consistency information is crucial for fake news identification. Comparing "w/o EF" with CIFF-Net, "w/o EF" shows a certain degree of performance drop, suggesting that our design of enhancing multi-modal fused feature with single modal features is effective, and also indicates that the interaction understanding of different modality information is helpful for extracting consistency information and improving detection performance. Comparing the results of "w/o Sim" with CIFF-Net shows that adjusting the weights of inter-modal consistency feature and inconsistency features using cosine similarity score can guide the classifier in learning evidence clue, thereby improving fake news identification. Additionally, Comparing the results of "w/o TFN" with CIFF-Net indicates that the TFN fusion strategy can more effectively integrate cross-modal features and achieve information gain of different modalis compared with simple concatenation.

Comparison of multi-modal backbones. We also conduct three ablation experiments on the Weibo-21 dataset to verify that the advantages of the proposed CIFF-Net are derived from the use of the innovative module rather than multi-modal backbone networks. The CLIP-only baseline ("with CLIP") only uses CLIP model to extract global text and image features. These features are concatenated and fed directly into the classifier. BERT and SWIN-T baseline ("with B+S") concatenate semantic features from text modality (BERT model) and image modality (SWIN-T model) to classify real and fake news. Multiple feature fusion baseline ("with MFF") integrates BERT- and SWIN-T-derived semantic features with the designed multiple feature fusion module to extract cross-modal consistency representations. Note that this baseline does not include our inconsistent branches (Image Branch and Text Branch) and the weight adaption mechanism.

Experimental results shown in Table V reveal that the "with MFF" architecture achieves an accuracy of 0.914, suggesting this value may approximate the upper performance bound attainable using pre-trained feature extractors alone. In contrast, our full CIFF-Net yields statistically significant improvements, attributable exclusively to the synergistic integration of three novel components: the text inconsistency branch, the image inconsistency branch, and the cosine-similarity-guided weight adaption mechanism.

TABLE VI
EXPERIMENTAL RESULTS OF ROBUSTNESS AGAINST GAUSSIAN NOISE.

Dataset	Parameter	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	$\sigma = 10$	0.911	0.883	0.945	0.913	0.942	0.878	0.909
	$\sigma = 25$	0.903	0.867	0.948	0.906	0.945	0.858	0.899
	$\sigma = 0$	0.930	0.951	0.909	0.930	0.910	0.952	0.931
Weibo-21	$\sigma = 10$	0.907	0.868	0.955	0.909	0.952	0.860	0.904
	$\sigma = 25$	0.886	0.844	0.942	0.890	0.937	0.832	0.881
	$\sigma = 0$	0.943	0.934	0.951	0.942	0.952	0.935	0.943
GossipCop	$\sigma = 10$	0.895	0.659	0.943	0.776	0.985	0.884	0.932
	$\sigma = 25$	0.869	0.605	0.919	0.730	0.978	0.857	0.914
	$\sigma = 0$	0.911	0.838	0.666	0.742	0.924	0.969	0.946

TABLE VII
EXPERIMENTAL RESULTS OF ROBUSTNESS AGAINST JPEG COMPRESSION.

Dataset	Parameter	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	quality=50	0.909	0.867	0.962	0.912	0.959	0.858	0.906
	quality=30	0.894	0.860	0.936	0.896	0.927	0.843	0.883
	quality=100	0.930	0.951	0.909	0.930	0.910	0.952	0.931
Weibo-21	quality=50	0.917	0.885	0.955	0.919	0.953	0.879	0.915
	quality=30	0.898	0.857	0.951	0.902	0.947	0.847	0.884
	quality=100	0.943	0.934	0.951	0.942	0.952	0.935	0.943
GossipCop	quality=50	0.888	0.647	0.923	0.761	0.980	0.880	0.927
	quality=30	0.870	0.603	0.901	0.722	0.973	0.859	0.912
	quality=100	0.911	0.838	0.666	0.742	0.924	0.969	0.946

D. Low-quality News Data Detection

To further evaluate the detection performance of CIFF-Net in handling low-quality data, we apply varying degrees of Gaussian noise and JPEG compression to the testing datasets. Table VI presents the robustness against Gaussian noise. It can be observed that as the standard deviation σ of the noise increases, the overall accuracy of our CIFF-Net in detecting fake news declines, while both precision and recall exhibit noticeable fluctuations. This suggests that noise interference negatively affects the ability to assess news authenticity.

However, CIFF-Net still maintains a relatively high level of accuracy in identifying fake news.

Table VII displays the robustness against JPEG compression. Across all datasets, as the compression factor decreases and data quality degrades, the detection accuracy correspondingly diminishes. For fake news, the recall rate remains relatively stable across different compression levels. In the case of real news detection, various evaluation metrics fluctuate with changes in compression quality, nevertheless, the variations are minor. These results demonstrate that CIFF-Net exhibits a considerable degree of robustness against JPEG compression.

The high detection accuracy on low-quality news data is attributed to the two aspects. Firstly, our method jointly leverages two orthogonal yet complementary inference pathways, namely, cross-modal consistency features and intra-modal inconsistency features. This dual-clue architecture enables mutual calibration. When one pathway is compromised by data corruption, the other retains sufficient discriminative capacity to sustain reliable inference. For instance, severe textual noise (e.g., pervasive typos) may impair consistency-based alignment, yet the inconsistency branch remains capable of identifying salient entities or event elements whose semantics contradict the image, providing decisive evidence for detection. Similarly, under severe image degradation (e.g., JPEG compression), consistency signals weaken, but inconsistency features derived from both modalities retain discriminative power through cross-modal discrepancy modeling. This synergistic design substantially enhances model robustness and fault tolerance in low-quality data regimes. Moreover, our weight adaption mechanism operates on CLIP-derived global text-image semantic similarity scores not on local, fine-grained feature activations. By grounding weight assignment in holistic, semantically grounded alignment rather than brittle local correspondences, the mechanism inherently mitigates sensitivity to localized noise, distortions, or spurious correlations, thereby improving stability and generalization under data degradation.

E. Visualization Analysis

To intuitively analyze the role of each module in fake news detection, we employ the T-SNE visualization technique to evaluate the performance of different modules on the Weibo-21 dataset. The results are presented in Fig. 4, where dots of the same color represent news with the same label (green for real news and orange for fake news).

Analysis of single modal feature extraction. It can be observed that compared with "w/o IB", "w/o TB" resulted in greater overlap between labels, with representations of the same class being more dispersed. This suggests that textual inconsistency features contribute more significantly to fake news detection than image-based inconsistency features, which aligns with our earlier quantitative findings.

Analysis of feature fusion. When analyzing the visualization of "w/o FF", although some degree of class discrimination is present, there remains substantial overlap between categories, and its separability is notably lower than that of both "w/o IB" and "w/o TB". This highlights the critical importance

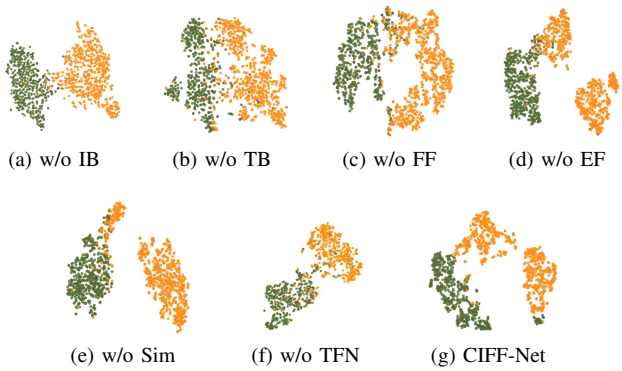


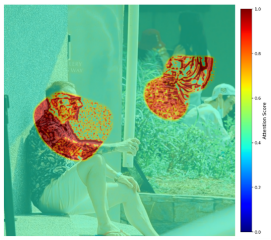
Fig. 4. Visualization results of fake news detection using different modules on the Weibo-21 dataset, where green dots represent real news and orange dots represent fake news.

of cross-modal consistency information in fake news detection. Furthermore, it indicates that intra-modal inconsistency features can complement inter-modal consistency features, thereby enhancing the overall classification capability. In the case of "w/o EF", the learned classification representations appear fragmented within classes, with some overlaps across different labels. This implies that leveraging single-modal features to enhance multi-modal fusion can reduce intra-class distances while increasing inter-class separation, thus improving the model's discriminative power. The visualization of "w/o TFN" reveals a degree of confusion in the learned representations across different labels. Although some separation tendencies exist, the intra-class compactness and inter-class distinction are markedly inferior compared to those achieved by the full CIFF-Net model. This further supports the effectiveness of the TFN fusion strategy in promoting better feature integration, resulting in clearer and more distinguishable classification representations.

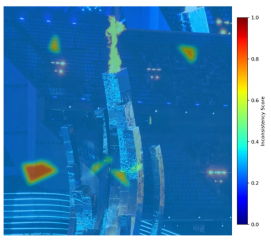
Analysis of similarity calculation. The classification representations generated by "w/o Sim" exhibit a certain level of separability. However, the number of misclassifications has increased significantly. This clearly demonstrates the crucial role of the cosine similarity score in guiding the classifier toward accurate predictions and reducing classification errors.

The classification representations obtained using the complete CIFF-Net demonstrated the highest level of separability and the tightest intra-class aggregation. This strongly validates the effectiveness of the proposed method, which mimics how humans identify fake news. Specifically, when the content across modalities exhibits high consistency, the model focuses more on inconsistent components. Conversely, when consistency is low, the model emphasizes similar parts.

Moreover, we conduct the visualization study on real and fake news examples to prove that the inconsistent features of text and image correspond to the human-perceived conflicting regions. For textual inconsistency R_{T_incon}, we compute token-level inconsistency scores and aggregate them into a text-level heatmap, with high-inconsistency tokens highlighted in red. For visual inconsistency R_{I_incon}, we upsample patch-level inconsistency features extracted from the SWIN-T backbone to match the spatial resolution of the original image, then overlay the resulting heatmap onto the image to localize



(a) A well-known artist was administratively detained by the police on the street for drunk driving. The scene was exposed.



(b) At the opening ceremony of the Hangzhou Asian Games, the moment the main torch tower was lit.

Fig. 5. Visualization results of the inconsistent features between fake (a) and real (b) news.

regions exhibiting high inconsistency.

Fig. 5 (a) presents a representative example of highly consistent fake news. Although surface-level textual and visual elements appear aligned, the core semantic content is fabricated. As shown in Fig. 5 (a), textual inconsistency hotspots concentrate precisely on semantically forged phrases, such as "drunk driving", "administratively detained by the police", and "scene". While visual inconsistency hotspots localize to salient semantic regions, including the artist's face and contextual street elements. This is exactly in line with the conflict point in human perception that "there is no semantic connection between the image and the text description of the 'detention scene' event", which is precisely the clue for humans to identify fake news. Fig. 5 (b) illustrates a real news example with high cross-modal consistency. Both textual and visual inconsistency scores remain uniformly low across the sample, reflecting the intrinsic alignment between image and text of real news. Collectively, these visualizations demonstrate that inconsistency scores are driven primarily by semantic conflict, not by superficial attributes such as visual complexity, background clutter, or token prominence, thereby validating their discriminative utility for multimodal fake news detection.

V. Conclusion

In this paper, we propose an adaptive cross-modal inconsistent feature fusion network for fake news detection. CIFF-Net strengthens the information interaction between news image and text modalities through the proposed multiple feature fusion module, thereby fully learning the inter-modal consistency information. Meanwhile, a single modal feature filtering strategy is designed to obtain inconsistency features from different modalities, which can filter out redundant information and interference noise, and avoid the mutual masking effect of multi-modal information. Furthermore, based on the global features of images and text, a similarity score is computed. Consistency and inconsistency features are weighted with the similarity score to adaptively fuse these features, so as to better guide the classifier in determining the authenticity of news. Experimental results demonstrate that our CIFF-Net achieves greater performance in fake news detection compared with several SOTA methods.

References

[1] W. Zhang, H. Fu, H. Wang, Z. Gong, P. Zhou, and D. Wang, "3a multi-classification division-aggregation framework for fake news detection," *IEEE Transactions on Big Data*, vol. 11, no. 1, pp. 130–140, 2025.

[2] T. Qiao, H. Shao, S. Xie, and R. Shi, "Unsupervised generative fake image detector," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 8442–8455, 2024.

[3] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *Proceedings of The International Joint Conference on Artificial Intelligence*, 2016, pp. 3818–3824.

[4] M. Cheng, S. Nazarian, and P. Bogdan, "Vroc: Variational autoencoder-aided multi-task rumor classifier based on text," in *Proceedings of The Web Conference*, 2020, pp. 2892–2898.

[5] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 598–608, 2016.

[6] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *Proceedings of IEEE International Conference on Data Mining*, 2019, pp. 518–527.

[7] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proceedings of The AAAI Conference on Artificial Intelligence*, 2020, pp. 13 915–13 916.

[8] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multi-modal fake news detection," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2020, pp. 354–367.

[9] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proceedings of The ACM Web Conference*, 2022, pp. 2897–2905.

[10] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang, "Multi-modal fake news detection on social media via multi-grained information fusion," in *Proceedings of The ACM International Conference on Multimedia Retrieval*, 2023, pp. 343–352.

[11] P. Du, Y. Gao, L. Li, and X. Li, "Sgamf: Sparse gated attention-based multimodal fusion method for fake news detection," *IEEE Transactions on Big Data*, vol. 11, no. 2, pp. 540–552, 2025.

[12] Q. Ying, X. Hu, Y. Zhou, Z. Qian, D. Zeng, and S. Ge, "Bootstrapping multi-view representations for fake news detection," in *Proceedings of The AAAI Conference on Artificial Intelligence*, 2023, pp. 5384–5392.

[13] X. Zhang, S. Dadkhah, A. G. Weismann, M. A. Kanaani, and A. A. Ghorbani, "Multimodal fake news analysis based on image-text similarity," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 1, pp. 959–972, 2023.

[14] Z. Shao, Y. Su, Y. Zhou, F. Meng, H. Zhu, B. Liu, and R. Yao, "Ct-net: Arbitrary-shaped text detection via contour transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1815–1826, 2024.

[15] Y. Li, L. Hu, L. Dong, H. Wu, J. Tian, J. Zhou, and X. Li, "Transformer-based image inpainting detection via label decoupling and constrained adversarial training," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1857–1872, 2024.

[16] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A convolutional approach for misinformation identification," in *Proceedings of The International Joint Conference on Artificial Intelligence*, 2017, pp. 3901–3907.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of The IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

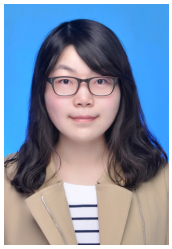
[19] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *Proceedings of IEEE International Conference on Multimedia Big Data*, 2019, pp. 39–47.

[20] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of The ACM International Conference on Multimedia*, 2017, pp. 795–816.

[21] L. Zhang, X. Zhang, Z. Zhou, F. Huang, and C. Li, "Reinforced adaptive knowledge learning for multimodal fake news detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 16 777–16 785.

[22] Q. Zhang, J. Liu, F. Zhang, J. Xie, and Z.-J. Zha, "Natural language-centered inference network for multi-modal fake news detection," in *Proceedings of The International Joint Conference on Artificial Intelligence*, 2024, pp. 2542–2550.
[23] A. Lao, Q. Zhang, C. Shi, L. Cao, K. Yi, L. Hu, and D. Miao, "Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 18 426–18 434.
[24] J. Li, Y. Bin, J. Zou, J. Wei, G. Wang, and Y. Yang, "Cross-modal consistency learning with fine-grained fusion network for multimodal fake news detection," in *Proceedings of The ACM International Conference on Multimedia in Asia*, 2024, pp. 1–7.
[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of International Conference on Machine Learning*, 2021, pp. 8748–8763.
[26] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *Proceedings of International Conference on Learning Representations*, 2022, pp. 1–13.
[27] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1–11, 2019.
[28] S. Xiong, G. Zhang, V. Batra, L. Xi, L. Shi, and L. Liu, "Trimoon: Two-round inconsistency-based multi-modal fusion network for fake news detection," *Information Fusion*, vol. 93, pp. 150–158, 2023.
[29] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of The Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–11, 2017.
[31] Q. Nan, J. Cao, Y. Zhu, Y. Wang, and J. Li, "Mdfend: Multi-domain fake news detection," in *Proceedings of The ACM International Conference on Information and Knowledge Management*, 2021, pp. 3343–3347.
[32] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.
[33] Y. Li, H. He, J. Bai, and D. Wen, "Mcfend: A multi-source benchmark dataset for chinese fake news detection," in *Proceedings of the ACM Web Conference*, 2024, p. 4018–4027.
[34] X. Liu, Z. Li, P. P. Li, H. Huang, S. Xia, X. Cui, L. Huang, W. Deng, and Z. He, "MMFakebench: A mixed-source multimodal misinformation detection benchmark for LVLMs," in *The International Conference on Learning Representations*, 2025, pp. 1–26.

Aohan Li received the B.E. degree in information security from Hunan University of Science and Technology in 2022, and the M.E. degree in electronic information from Hunan University in 2025. His current research interests include fake news detection.



Jiaxin Chen received the B.S. degree from Central China Normal University in 2017, and Ph.D. degree from Hunan University in 2023. She is currently a lecturer with Changsha University of Science and Technology, China. Her current research interests include multimedia forensics and Deepfake detection. She is a member of Asia-Pacific Signal and Information Processing Association (APSIPA) Technical Committee on Media Security and Forensics, and Technical Committee (TC) on Digital Forensics and Security of China Society of Image and Graphics.

Dengyong zhang received the B.S. and M.S. degree from Changsha University of Science and Technology, Changsha, China, in 2003, 2006 respectively. He received Ph.D. degree from Hunan University, Changsha, China, in 2018. Now, He is a Professor at Changsha University of Science and Technology. His current research interests include digital media forensics and image processing.

Yuling Liu received her B.E. degree in computer science and technology, in 2003, and the Ph.D. degree in computer application, in 2008, Hunan University, Changsha, China. She is currently a professor with the College of Cyber Science and Technology of Hunan University, Changsha, China. Her research interests include AI security, information security, natural language processing, multimedia steganography.

Yihui Luo is a Senior Engineer holding professional certifications as Network Planning Designer and Information Security Engineer. He currently serves as the Section Chief of the Network and Data Security Section at the Statistical Information Center, Human Resources and Social Security Department of Hunan Province. Concurrently, he is a Member of the Hunan Provincial Cybersecurity Standardization Technical Committee. His current research interests include network security, data security, and personal information protection.

Xin Liao (Senior Member, IEEE) received the B.E. and Ph.D. degrees in information security from the Beijing University of Posts and Telecommunications in 2007 and 2012, respectively. He is currently a Professor with the College of Cyber Science and Technology, Hunan University, where he also serves as the Vice Dean. He worked as a Post-Doctoral Fellow with the Institute of Software, Chinese Academy of Sciences, and also a Research Associate with The University of Hong Kong. From 2016 to 2017, he was a Visiting Scholar with the University of Maryland, College Park, USA. His current research interests include multimedia forensics, AI security, and watermarking. He is the Vice Chair of Technical Committee (TC) on Multimedia Security and Forensics of Asia-Pacific Signal and Information Processing Association, a member of TC on Computer Forensics of the Chinese Institute of Electronics, and a member of TC on Digital Forensics and Security of the China Society of Image and Graphics. He is serving as an Associate Editor for the IEEE Signal Processing Magazine.