Modality Perception Learning-Based Determinative Factor Discovery for Multimodal Fake News Detection

Boyue Wang, Guangchao Wu, Xiaoyan Li, Junbin Gao, *Associate Member, IEEE*, Yongli Hu, *Member, IEEE*, and Baocai Yin, *Member, IEEE*

***Abstract*— The dissemination of fake news, often fueled by exaggeration, distortion, or misleading statements, significantly jeopardizes public safety and shapes social opinion. Although existing multimodal fake news detection methods focus on multimodal consistency, they occasionally neglect modal heterogeneity, missing the opportunity to unearth the most related determinative information concealed within fake news articles. To address this limitation and extract more decisive information, this article proposes the modality perception learning-based determinative factor discovery (MoPeD) model. MoPeD optimizes the steps of feature extraction, fusion, and aggregation to adaptively discover determinants within both unimodality features and multimodality fusion features for the task of fake news detection. Specifically, to capture comprehensive information, the dual encoding module integrates a modal-consistent contrastive language-image pre-training (CLIP) pretrained encoder with a modal-specific encoder, catering to both explicit and implicit information. Motivated by the prompt strategy, the output features of the dual encoding module are complemented by learnable memory information. To handle modality heterogeneity during fusion, the multilevel cross-modality fusion module is introduced to deeply comprehend the complex implicit meaning within text and image. Finally, for aggregating unimodal and multimodal features, the modality perception learning module gauges the similarity between modalities to dynamically emphasize decisive modality features based on the cross-modal content heterogeneity scores. The experimental evaluations conducted on three public fake news datasets show that the proposed model is superior to other state-of-the-art fake news detection methods.**

***Index Terms*— Adaptive prompt learning, cross-modal analysis, modality perception learning, multimodal fake news detection.**

Manuscript received 16 January 2024; revised 22 May 2024; accepted 12 August 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC3304601; and in part by the National Natural Science Foundation of China under Grant 92370102, Grant 62272015, and Grant U21B2038. *(Corresponding author: Xiaoyan Li.)*

Boyue Wang, Guangchao Wu, Xiaoyan Li, Yongli Hu, and Baocai Yin are with Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing Artificial Intelligence Institute, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: wby@bjut.edu.cn; guangchaowu@emails.bjut.edu.cn; xiaoyan.li@bjut.edu.cn; huyongli@bjut.edu.cn; ybc@bjut.edu.cn).

Junbin Gao is with the Discipline of Business Analytics, The University of Sydney Business School, The University of Sydney, Camperdown, NSW 2006, Australia (e-mail: junbin.gao@sydney.edu.au).

Digital Object Identifier 10.1109/TNNLS.2024.3446030

I. Introduction

IN THE era of digital advancement, the popularity of online social media and the Internet has made the dissemination of information unprecedentedly easy and fast. However, this convenience has also presented an opportunity for the extensive spread of fake news. In such a scenario, people are faced with massive amounts of information and news sources, and sometimes, people lack sufficient judgment in the authenticity of the information, which makes them vulnerable to fake news. The research of fake news detection can efficiently analyze a large number of news posts and help to quickly identify possible fake news. Therefore, fake news detection is an efficient and effective way to provide people with a more reliable information environment.

Fake news is widely defined as intentional and verifiable fake news articles with the potential to mislead readers [1]. The task of fake news detection involves around evaluating the truthfulness of a given news article by examining its news content, social context, and external knowledge. As shown by the previous works [2], [3], [4], the determinative information that can distinguish the reliability of news usually includes exaggerated language styles, manipulated images, inconsistent multimodal contexts, and more. When predicting with these decisive factors, a fake news classifier should obey the logical "AND" operation, meaning that, in spite of voluminous rational information, a bit of unreasonable content should lead to the prediction of the news as fake.

A large number of unimodal fake news detection methods [3], [5], [6] have been studied and innovated, focusing on the utilization of inflammatory language style or manipulated images to promote the widespread dissemination of fake news [7], [8], [9], [10], [11]. For example, Fig. 1(a) demonstrates these kinds of fake news. Examining solely from the image perspective reveals noticeable splicing traces. Meanwhile, an exclusive focus on the text highlights the phrase "gave birth to 11 fetuses at a time" which is deemed unreasonable. These unimodal determinative factors collectively affirm that the news is indeed fake.

When presented with multimodal inputs, fake news detectors examine both single-modality and cross-modality features to make judgments. Previous methodologies have primarily focused on designing sophisticated feature extractors

Fig. 1. Examples of multimodal news explain the MoPeD.[1] (a) "Woman in India gave birth to 11 fetuses at a time, a world record! Congratulations!" (b) "Happy New Year! Auspicious, happy, well-being!"

and multimodal fusion to introduce modality consistency cues between text and visual information, thereby enhancing fake news detection [12], [13], [14], [15], [16]. For instance, EANN [12] directly concatenates text and visual features extracted by the specific encoders for the final news representation. However, this method does not perform any fusion between the two heterogeneous unimodalities. Consequently, significant semantic gaps between the text and image may arise, leading to instances where real news with inconsistent explicit information is incorrectly classified as fake. MFAN [16] incorporates text, visual, and social graph features into a unified framework. These different features undergo direct aggregation and interaction before classification. In this case, the emphasis on cross-modal semantic consistency becomes the sole determinant, potentially overlooking crucial decisive factors hidden in the unimodal information. Other methods [14] simply concatenate unimodal and cross-modal features as the input news representation. However, they assume features from different sources contribute equally to the final prediction, and thus, they may lead to overfitting to specific decisive factors, such as modality consistency in most cases. This contradicts the reality that fake news often involves various subtle forms of deceptive content.

Therefore, both credible and deceptive information coexist within a single piece of fake news. To correctly classify such news, it becomes crucial to emphasize determinative information hidden in different modalities, prioritizing it over potentially misleading information according to specific samples. As shown in Fig. 1(a), the sample is hard to be correctly classified since its textual and visual contents exhibit strong consistency. On the contrary, in Fig. 1(b), the real news shows different semantic meanings between image and text, but the fireworks image and "Happy New Year" text implicitly describe the same happy atmosphere, affirming it as a piece of real news.

[1]Two images in Fig. 1, respectively, come from the Weibo dataset and https://weibo.com/2692054787/NaMlW4Mlm.

To perceive the contributions of different features, recent methods calculate ambiguity scores for different modalities and assign a higher weight to cross-modal features when they exhibit ambiguity, specifically inconsistencies among them [17], [18]. However, these methods adopt independent unimodal feature encoders, resulting in heterogeneous features with notable semantic gaps between modalities. Consequently, extracting certain implicit information, e.g., cross-modal semantic consistency, becomes more challenging in this scenario.

Given the aforementioned challenges, the goal of this article is to comprehensively extract both explicit and implicit information hidden in both unimodal and crossmodal features, highlighting decisive factors for final classification. To achieve this goal, we introduce the modality perception learning-based determinative factor discovery (MoPeD) model. MoPeD is designed to progressively uncover and assimilate determinative information at each stage of feature extraction, fusion, and aggregation. During feature extraction, dual encoders, complemented by a pretrained contrastive language-image pre-training (CLIP) model, are employed to learn rich modal-consistent and modal-specific features, in which the decisive information is hidden. In feature fusion, inspired by the fact that people repeatedly compare text and image to obtain the desired content, the extracted information is fused at multiple levels. This approach aims to foster a deep understanding of the implicit meanings interwoven between different modalities. During feature aggregation, the model achieves an adaptive perception of the modalities' contributions by calculating heterogeneity scores, thereby ensuring that the obtained determinative information can be highlighted. In summary, the MoPeD model effectively addresses the challenges related to insufficient understanding of implicit information and the difficulty in mining decisive factors in multimodal fake news detection.

The main contributions of this article are summarized as follows.

1) The dual encoding module is designed to merge the pretrained CLIP encoder with the modality-specific encoder. The CLIP pretrained encoder helps diminish the disparity between modalities, while the modality-specific encoder focuses on learning the distinctive features of each modality.
2) Inspired by the prompt strategy, the memory information components are designed to learn and adjust the unimodality feature representations, which mines the latent patterns of real and fake news.
3) The multilevel cross-modality fusion module is built to capture the complex correlation between images and texts. This module is dedicated to extracting multimodality deep determinative information, providing a solution for handling the complexities associated with detecting fake news.
4) The proposed modality perception learning module uses the distribution divergence of text and image to adaptively perceive the importance of cross-modality fusion features and unimodality features.

The rest of this article is structured as follows. Section II provides a concise review of related works in the field of

fake news detection. In Section III, we present an elaborate description of the proposed MoPeD model. In Section IV, the experimental results are reported and analyzed in detail. Finally, the main contributions are concluded in Section V.

II. Related Work

In this section, we provide a brief review of several relevant studies on unimodal fake news detection and multimodal fake news detection.

A. Unimodal Fake News Detection

Existing unimodal fake news detection methods can be divided into two main categories: *the text-based detection* and *the visual-based detection*.

For the text-based detection, several methods explore the distinctive linguistic features of fake news, relying on text content information [7], [8]. In the early stages, many works did not distinguish between various news domains or only focused on a single field [2], [9], [19]. SMS [20] extracts text features to assist in evaluating the credibility of news articles and determining the legitimacy of news articles. CNNML [21] introduces a convolutional neural network (CNN) with edge loss to handle the news text and incrementally updates word embeddings during the training stage. DRNN [22] proposes a bidirectional recurrent neural network with only two bidirectional long short-term memory (BiLSTM) layers and two dense layers. SAAD [23] proposes a stacking approach for fake news detection, which evaluates five machine learning models and three deep learning models on two datasets through cross-validation. Recently, researchers have begun to pay attention to multidomain fake news detection. MDFEND [5] uses multiple representations extracted by the mixture-of-experts method and aggregates them with domain gates to identify fake news. Meanwhile, M3FEND [6] proposes the domain adapter and the domain memory bank to address challenges related to domain shift and incomplete domain labeling. For the visual-based detection, certain works emphasize the significance of visual content in posts [3], [10], [11], [24]. MVNN [3] uses the visual information in both the frequency domain and the pixel domain to effectively capture the physics and semantics of image characteristics. However, it primarily focuses on single-modality information.

B. Multimodal Fake News Detection

Existing multimodal fake news detection methods can be broadly divided into the following three categories.

The first category focuses on the multimodal information fusion. Typically, visual information is extracted using a visual feature extractor (VGG, ResNet, etc.), while text features are extracted through a text feature extractor (bidirectional encoder representations from transformers (BERT), Text-CNN, etc.). In Spotfake [25], two types of extracted unimodal features, visual and text, are concatenated as inputs to the classifier. This fusion strategy is further improved in Spotfake+ [26] for full-length article detection. EANN [12] inputs the multimodal features to the event classifier, which learns the invariant features of events and facilitates the detection of newly emerged event. MKEMN [27] combines text, image, and background knowledge to capture the semantic information from multiple modalities, enhancing the accuracy of fake news detection. MCAN [15] stacks multiple co-attention layers together to fuse multimodal features and learn the interdependencies between multiple modalities. HMCAN [13] utilizes multimodal contextual information and a hierarchical encoding network to capture the hierarchical semantics of textual information, thereby enhancing the representation of multimodal news. CARMN [28] combines the cross-modal attention residual network and the multichannel CNN to maintain the unique information of different modals and reduce the noise caused by modal differences. MPFN [29] proposes a novel multimodal progressive fusion network that simultaneously considers both deep and shallow information of image sampling at different levels by Swin transformer, and designs a multilayer perceptron (MLP) mixer to integrate visual features with textual features at different levels.

For the rest, a large number of methods focus on measuring the similarity between modalities. MVAE [30] proposes the multimodal variational autoencoder (VAE), reconstructing two modalities from the text and image representations to find the correlation between the modalities. SAFE [4] defines the relationship between text and visual information through their fine-tuned cosine similarity. EM-FEND [14] uses the embedded text in the image as a supplement to the text information and then calculates the similarity between the text entity and the visual entity. CAFE [17] solves the cross-modal ambiguity learning problem by quantifying the ambiguity between different unimodal features using the distribution divergence. MMFN [31] extracts the multigranularity features and calculates the unimodal similarity to assist multimodal classification adaptively. BMR [32] employs a bootstrapping approach for multiview representation and then separates unimodal and multimodal features through single-view prediction and consistency learning. This process helps reweight and guide the features for enhanced better detection performance.

There are also a few methods that use graph neural networks to fuse external knowledge. KMGCN [33] constructs a heterogeneous information network using text, image, and entity information in the knowledge graph, and then uses graph convolutional network fusion features to obtain news representation. MFAN [16] builds a social graph containing news articles, news publishers, and comments. It is the first try to incorporate text, visual, and social graph features into the same framework.

III. Proposed Model

In this article, we propose the MoPeD model to address the challenge of understanding and discovering determinative information across heterogeneous modalities in multimodal fake news detection. The overall framework of the proposed model is shown in Fig. 2. The MoPeD model is designed to adaptively discover the decisive factors of each input news step by step.

1) The dual encoding module extracts intact information during the feature extraction stage.

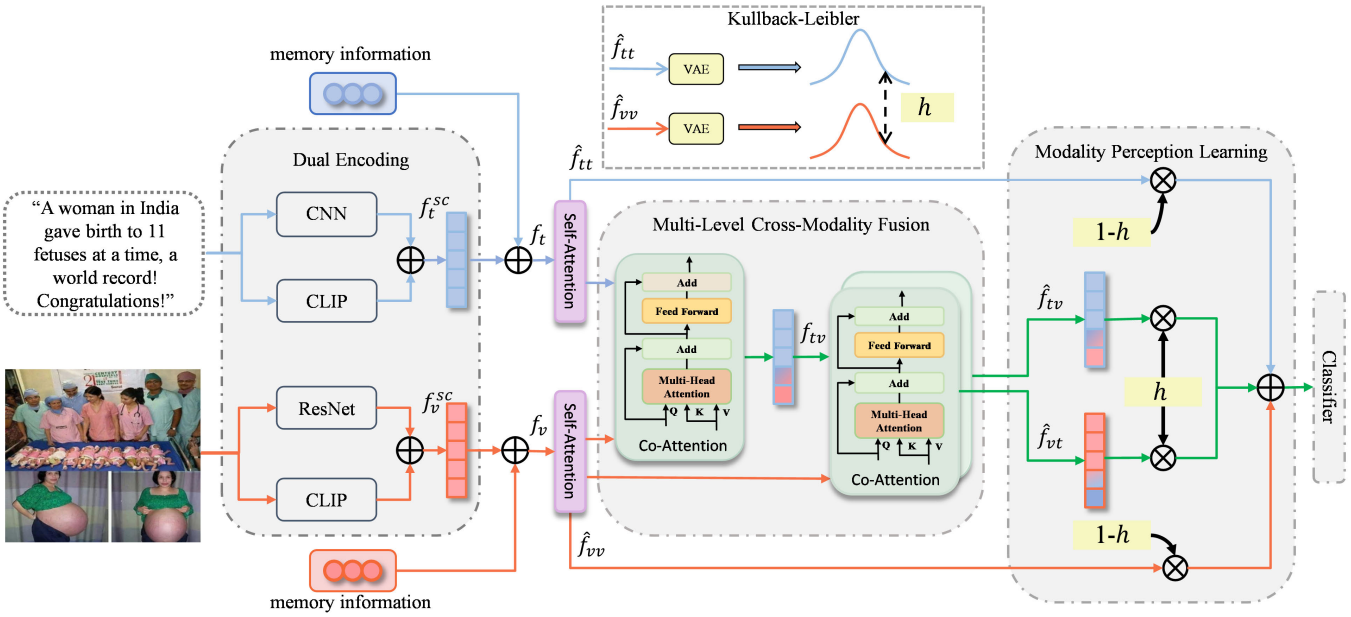


Fig. 2. Architecture of the proposed MoPeD model. We first obtain the unimodal features via the dual encoding module, where CNN and CLIP extract text features, while ResNet and CLIP handle image features. Then, the features undergo MHSA and are fed into the multilevel cross-modal fusion module to thoroughly explore implicit information. The modality perception learning module utilizes the heterogeneity score h, representing the distance between the distribution of text and image, to weight fusion features. The aggregation of features can be dynamically adjusted according to the variation of h.

2) The multilevel cross-modality fusion module is dedicated to uncovering implicit cross-modal information during the feature fusion stage.
3) The modality perception learning module highlights determinative factors in the feature aggregation step.

In Sections III-A–III-D, these modules, the classifier, and the objective function are described in detail.

A. Dual Encoding

To extract comprehensive multimodal information contributing to the classification task, the proposed dual encoding module combines CNNs as a modality-specific encoder and CLIP encoder as the common pretrained encoder.

Formally, the multimodal news dataset is denoted as $P = [T, V]$, where T stands for the set of text inputs and V stands for the set of visual inputs.

In the feature extraction stage, given a news post $p = [t, v] \in P$, two types of encoders are used to extract both modal-specific features and modal-consistent features for text and image as follows:

$$f_t^s = \mathcal{F}_{\text{txt}}^s(t), \quad f_t^c = \mathcal{F}_{\text{txt}}^c(t) \tag{1}$$
$$f_v^s = \mathcal{F}_{\text{vis}}^s(v), \quad f_v^c = \mathcal{F}_{\text{vis}}^c(v) \tag{2}$$

where $\mathcal{F}_{\text{txt}}^s$, $\mathcal{F}_{\text{txt}}^c$, $\mathcal{F}_{\text{vis}}^s$, and $\mathcal{F}_{\text{vis}}^c$ are the modal-specific and modal-consistent encoders for textual and visual modalities, respectively. f_t^s and f_t^c denote the modal-specific and modal-consistent features for the text component t while f_v^s and f_v^c denote the modal-specific and modal-consistent features for the visual component v, respectively. For convenience, in the sequel, we use f_m^s and f_m^c to denote them with the single modal modality $m = t$ or v, respectively.

The final output f_m of the dual encoding module is a concatenation of these features as follows:

$$f_m^{sc} = \text{concat}\left(f_m^s, f_m^c\right) \tag{3}$$

where concat$(\cdot)$ means the concatenation operator. Next, the modal-specific and modal-consistent encoders are illustrated in more detail.

1) Modal-Specific Encoders $\mathcal{F}_{txt}^s/\mathcal{F}_{vis}^s$: The modal-specific encoders for both text and image all exploit CNNs to extract features while preserving the unique attributes of different modalities.

To be specific, the textual information t in each post is extracted by a CNN with pooling [16] for the text-specific feature with different granularities. Each word t^j in the input t is initially represented with word embeddings using an off-the-shelf method [34] and then padded or truncated to the same length as follows:

$$t^{1:L} = \left\{t^1, t^2, \ldots, t^L\right\} \tag{4}$$

in which $t^j \in \mathbb{R}^{d_s}$, d_s is the dimension of word embeddings, and L is the length of tokens. A convolutional layer of size k is applied to the sequence of continuous inputs $t^{1:L}$ to generate a feature map f^k, and a pooling operation is then applied to reduce the dimensionality

$$\mathbf{f}^k = \left\{f^1, f^2, \ldots, f^{L-k+1}\right\}$$
$$p^k = \text{Pooling}\left(\mathbf{f}^k\right). \tag{5}$$

Following the strategy in [16], we obtain the representation with three choices of kernel size $k \in \{3, 4, 5\}$, each with $d_s/3$ heads of filters. Then, we concatenate them as the text-specific feature $f_t^s \in \mathbb{R}^{d_s}$ to consider the word embedding relationships of different ranges

$$f_t^s = \text{concat}\left(p^{k=3}, \ p^{k=4}, p^{k=5}\right). \tag{6}$$

The visual information v in each post is encoded by the ResNet-50 pretrained model [35] to learn the specific representation from images. In this work, we use the output of the second last layer of the ResNet-50 as the input to a fully connected layer to obtain the visual-specific feature $f_v^s \in \mathbb{R}^{d_s}$ as follows:

$$f_v^s = \mathcal{F}_{\text{vis}}^s(v) = \text{ResNet}(v) \tag{7}$$

where d_s is the dimension of the visual embedding the same as the text feature.

2) Modal-Consistent Encoders $\mathcal{F}_{txt}^c/\mathcal{F}_{vis}^c$: The previous modal-specific encoders can extract features with its detailed modality characteristics, such as exaggerated language style and tampered image in news posts. However, encoding the same semantic using different pretrained CNN networks often results in different embeddings for different modalities without any consistent constraint. Therefore, to address the semantic gaps mentioned in this article, a modal-consistent encoder is exploited to generate modal-consistent features. CLIP, being a multimodal pretraining model trained on a substantial amount of data, operates on the core idea of mapping the image and text embeddings into a shared semantic space.

To more effectively extract determinative information from different modalities, we combine the CLIP pretrain encoder and the modality-specific encoders, forming the dual encoding module. The CLIP pretrained text and image features are denoted as $f_t^c \in \mathbb{R}^{d_c}$ and $f_v^c \in \mathbb{R}^{d_c}$, respectively, where d_c is the length of the feature vectors.

Then, the embeddings from two types of encoders are concatenated to form the unimodality feature $f_t^{sc} \in \mathbb{R}^{d_{sc}}$ and $f_v^{sc} \in \mathbb{R}^{d_{sc}}$ for text and visual, as expressed in the following equation:

$$\begin{aligned} f_t^{sc} &= \text{concat}\left(f_t^s, f_t^c\right) \\ f_v^{sc} &= \text{concat}\left(f_v^s, f_v^c\right) \end{aligned} \tag{8}$$

where $d_{sc} = d_s + d_c$ is the dimension of the outputs of the dual encoding module.

3) Memory Information: To enhance the model's understanding of the true or false mode of different modal features in the datasets on this specific task, we introduce a learnable vector representation as memory information $f_t^{\text{mem}} \in \mathbb{R}^{d_{\text{mem}}}$ and $f_v^{\text{mem}} \in \mathbb{R}^{d_{\text{mem}}}$ for text and image, respectively, where d_{mem} is the length of the learnable memory vector. With the backpropagation algorithm, the memory vectors can be updated constantly to adapt to the specific tasks, which learns the latent knowledge from the training set. This information can control and adjust the input of the modalities to a certain extent, thereby improving the flexibility of the model and achieving more accurate results that meet specific needs. Additionally, the memory information builds the relationship between different batches in the training set.

The outputs of the dual encoding module and memory information are concatenated to form the final text and visual representations as follows:

$$\begin{aligned} f_t &= \text{concat}\left(f_t^{sc}, f_t^{\text{mem}}\right) \\ f_v &= \text{concat}\left(f_v^{sc}, f_v^{\text{mem}}\right) \end{aligned} \tag{9}$$

where $f_t \in \mathbb{R}^d$ and $f_v \in \mathbb{R}^d$, and the output dimension $d = d_{sc} + d_{\text{mem}}$.

B. Multilevel Cross-Modality Fusion

The reliability of news is often deeply concealed in its content, making it challenging to classify based solely on explicit information. To address this challenge, we observe that, when people face news posts that are difficult to comprehend, they typically read the news post repeatedly and combine the visual information to aid their understanding. Inspired by this observation, we connect two layers of cross-modal co-attention fusion module to thoroughly grasp the implicit information hidden in text and image. Next, the detail of the multilevel cross-modality fusion module is formally introduced.

Before entering the multilevel cross-modality fusion module, multihead self-attention (MHSA) is applied to enhance the representation of features within each modality. MHSA takes the same embeddings as inputs into the multihead attention (MHA) module, enabling the model to focus on complex relationships in the input sequence. For the feature of modality $m \in \{v, t\}$, f_m is initially projected to serve as queries, keys, and values for the ith head

$$\begin{aligned} Q_i^m &= f_m W_i^{sq} \\ K_i^m &= f_m W_i^{sk} \\ V_i^m &= f_m W_i^{sv} \end{aligned} \tag{10}$$

where Q_i^m, K_i^m, and V_i^m are the query, key, and value embeddings for the ith head, respectively. W_i^{sq}, W_i^{sk}, and $W_i^{sv} \in \mathbb{R}^{d \times d_n}$ are the learnable weight matrices. d is the input feature length and d_n is the output dimension of the projection. Then, the self-attention of each head is calculated as follows:

$$h_i^m = \text{softmax}\left(\frac{Q_i^m K_i^m}{\sqrt{d_h}}\right) V_i^m \tag{11}$$

$$h_m = h_1^m \oplus h_2^m \oplus \ldots \oplus h_n^m \tag{12}$$

$$M_m = h_m W_i^o \tag{13}$$

where $d_h = d_n/n$ and n is the number of heads, $\oplus$ denotes the concatenate operation, and $W_i^o \in \mathbb{R}^{d_n \times d}$ is the output linear transformations.

After being fed into a feedforward network comprising two fully connected (FC) linear layers and a rectified linear unit (ReLU) activation function, the final feature of the self-attention module for modality m is obtained

$$\begin{aligned} \hat{f}_{mm} &= \text{MHA}(f_m, f_m, f_m) \\ &= (M_m + f_m) + \text{FFN}(M_m + f_m) \end{aligned} \tag{14}$$

where modality $m = t$ or v.

Next, the enhanced unimodal features $\hat{f}_{vv}$ and $\hat{f}_{tt}$ are fed into the proposed multilevel cross-modality fusion module to thoroughly assimilate semantic information across modalities. The proposed module is composed of two different cross-modality co-attention layers, with each taking embeddings from different modalities as inputs for fusion.

In the first layer, inspired by the fact that humans usually use images to complement textual information when reading news articles, visual and textural information is input into a single

MHA for low-level fusion. Specifically, the visual feature $\hat{f}_{vv}$ is used to calculate the query matrix, while the text feature $\hat{f}_{tt}$ is projected for the key matrix and value matrix. The inputs of the ith head of the MHA are represented as follows:

$$\begin{aligned} Q_i^{tv} &= \hat{f}_{vv} W_i^{cq} \\ K_i^{tv} &= \hat{f}_{tt} W_i^{ck} \\ V_i^{tv} &= \hat{f}_{tt} W_i^{cv} \end{aligned} \tag{15}$$

where Q_i^{tv}, K_i^{tv}, and V_i^{tv} are the query, key, and value for the ith head of the first cross-modal co-attention layer, and W_i^{cq}, W_i^{ck}, and W_i^{cv} are the learnable projection matrices. The visual-enhanced text feature is obtained through the MHA model

$$f_{tv} = \mathrm{MHA}\big(\hat{f}_{vv},\ \hat{f}_{tt},\ \hat{f}_{tt}\big). \tag{16}$$

When people assess the authenticity of multimodal news, the practice of repeatedly reading the text and observing the details of the pictures can effectively help people combine known implicit information and capture the decisive factors hidden in the post. According to this insight, the second layer of fusion is designed, where two MHAs are performed similar to the above co-attention mechanism. However, in this case, the visual feature $\hat{f}_{vv}$ and the visual-enhanced text feature f_{tv} serve as inputs. The final fusion features are formulated as follows:

$$\begin{aligned} \hat{f}_{tv} &= \mathrm{MHA}\big(\hat{f}_{vv},\ f_{tv},\ f_{tv}\big) \\ \hat{f}_{vt} &= \mathrm{MHA}\big(f_{tv},\ \hat{f}_{vv},\ \hat{f}_{vv}\big). \end{aligned} \tag{17}$$

C. Modality Perception Learning

Following the multilevel cross-modality fusion, four interrelated features from heterogeneous modalities are obtained, i.e., the text feature $\hat{f}_{tt}$, the visual feature $\hat{f}_{vv}$, the visual-enhanced text feature $\hat{f}_{tv}$, and the text-enhanced visual feature $\hat{f}_{vt}$. Each of them significantly influences on the final result, but not all of them contain the determinative information inside. For instance, in a specific case of a fake news post, the text $\hat{f}_{tt}$ is deceptive while other features remain genuine. The correct perception of the necessary determinant factors is crucial. When the text and image show stronger heterogeneity, excessive focus on unimodal features might lead to the misclassification of some real news due to inconsistent text and image content, and the information in unimodality may cause a negative impact on the detection result.

Therefore, a heterogeneity score is calculated, measured by the Kullback-Leibler (KL) divergence of text and visual features, aiming to discern the contribution of unimodal features and multimodal fusion features and highlight the decisive ones.

Specifically, to compute the KL divergence of the text feature $\hat{f}_{tt}$ and visual feature $\hat{f}_{vv}$, the distribution of these data is first learned by utilizing the VAEs [17]. VAE is a generative model that can encode the data into the mean μ and variance σ of the latent space and generate samples by reparameterization. Given the text feature $\hat{f}_{tt}$ and visual feature $\hat{f}_{vv}$, the posterior of their latent variables can be denoted as

$$q\big(z_t|\hat{f}_{tt}\big) = \mathcal{N}\big(z_t|\mu\big(\hat{f}_{tt}\big), \sigma\big(\hat{f}_{tt}\big)\big) \tag{18}$$

$$q\big(z_v|\hat{f}_{vv}\big) = \mathcal{N}\big(z_v|\mu\big(\hat{f}_{vv}\big), \sigma\big(\hat{f}_{vv}\big)\big) \tag{19}$$

where $\mathcal{N}(z|\mu,\sigma)$ denotes the Gaussian distribution of random variable z with mean μ and variance σ, and z_t and z_v are the latent variable underlying the $\hat{f}_{tt}$ and $\hat{f}_{vv}$, respectively. $\mu(\hat{f}_{tt})$, $\sigma(\hat{f}_{tt})$, $\mu(\hat{f}_{vv})$, and $\sigma(\hat{f}_{vv})$ can be obtained from independent MLP layers. The KL divergence of two distributions is calculated and averaged to get the heterogeneity score h

$$\begin{aligned} h_1 &= \mathcal{D}_{\mathrm{KL}}\big(q\big(z_t|\hat{f}_{tt}\big)\|q\big(z_v|\hat{f}_{vv}\big)\big) \\ h_2 &= \mathcal{D}_{\mathrm{KL}}\big(q\big(z_v|\hat{f}_{vv}\big)\|q\big(z_t|\hat{f}_{tt}\big)\big) \\ h &= \mathrm{sigmoid}\left(\frac{h_1+h_2}{2}\right) \end{aligned} \tag{20}$$

where $\mathcal{D}_{\mathrm{KL}}$ represents the KL divergence, and sigmoid$(\cdot)$ is the sigmoid activation function that maps the heterogeneity score to a value between 0 and 1.

The heterogeneity score is served as a weight to dynamically perceive the importance of unimodal features and multimodal fused features

$$\begin{aligned} \hat{f}_{tt}{}' &= (1-h)\hat{f}_{tt} \\ \hat{f}_{vv}{}' &= (1-h)\hat{f}_{vv} \\ \hat{f}_{tv}{}' &= h\hat{f}_{tv} \\ \hat{f}_{vt}{}' &= h\hat{f}_{vt}. \end{aligned} \tag{21}$$

A relatively large heterogeneity score indicates a substantial gap between text and visual distributions. Given this inconsistency, relying solely on a single modality becomes less reliable for prediction. In such cases, greater attention should be paid to the multimodal features after fusion, which may encapsulate deep implicit decisive factors. On the contrary, when the heterogeneity score is relatively small, unimodal features describing its detailed modality property should be assigned a larger weight.

D. Classifier and Objective Function

The weighed unimodal features and multimodal features are concatenated as the final news representation

$$\hat{f} = \mathrm{concat}\Big(\hat{f}_{tt}{}', \hat{f}_{vv}{}', \hat{f}_{tv}{}', \hat{f}_{vt}{}'\Big) \tag{22}$$

then the feature $\hat{f}$ is fed into the classifier

$$\hat{y} = \mathrm{softmax}\big(\mathrm{FCs}\big(\hat{f}\big)\big) \tag{23}$$

where FCs$(\cdot)$ is the fake news classifier consisting of five fully connected layers with Relu$(\cdot)$ activation functions except for the last FC layer.

The cross-entropy loss function is adopted as the objective function to correctly predict real and fake news

$$\mathcal{L}_{\mathrm{cls}} = y\log\big(\hat{y}\big) + (1-y)\log\big(1-\hat{y}\big) \tag{24}$$

where y is the ground truth, with 1 representing fake news and 0 representing real news. In addition, as a common practice [16], we conduct the PGD [36] adversarial training on text embedding to enhance the robustness of the model and achieve more accurate predictions. During this process, we compute the adversarial sample loss and backpropagate gradients to update parameters.

TABLE I
COMPARISON BETWEEN MOPED AND THE STATE-OF-THE-ART MULTIMODAL FAKE NEWS DETECTION METHODS ON WEIBO, WEIBO-19, AND PHEME

	Method	Accuracy	Precision	Recall	F1-score	Fake News			Real News		
						Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	EANN	0.770	0.770	0.770	0.770	0.774	0.757	0.766	0.766	0.782	0.774
	MVAE	0.715	0.723	0.715	0.712	0.765	0.616	0.682	0.682	0.812	0.742
	SpotFake	0.767	0.772	0.767	0.766	0.734	0.834	0.780	0.810	0.701	0.752
	SpotFake+	0.731	0.732	0.731	0.731	0.716	0.760	0.737	0.748	0.702	0.725
	SAFE	0.812	0.823	0.812	0.810	0.884	0.715	0.790	0.763	**0.907**	0.829
	CARMN	0.825	0.826	0.825	0.825	0.837	0.816	0.826	0.814	0.835	0.824
	MFAN	0.844	0.845	0.844	0.844	0.863	0.816	0.839	0.828	0.872	0.849
	CAFE	0.849	0.848	0.849	0.849	**0.898**	0.810	0.852	0.798	0.891	0.842
	COOLANT	0.843	0.843	0.843	0.843	0.826	0.853	0.839	0.859	0.834	0.846
	MoPeD	**0.883**	**0.883**	**0.883**	**0.883**	0.879	**0.887**	**0.883**	**0.887**	0.878	**0.883**
Weibo-19	EANN	0.827	0.831	0.827	0.828	0.760	0.819	0.788	0.876	0.832	0.854
	MVAE	0.722	0.718	0.722	0.719	0.663	0.595	0.627	0.754	0.804	0.778
	SpotFake	0.712	0.711	0.712	0.696	0.707	0.459	0.555	0.714	0.877	0.787
	SpotFake+	0.722	0.718	0.722	0.718	0.667	0.586	0.624	0.751	0.810	0.780
	SAFE	0.854	0.855	0.854	0.855	0.807	0.828	0.817	0.886	0.872	0.879
	CARMN	0.864	0.864	0.864	0.864	0.845	0.817	0.831	0.877	0.897	0.887
	MFAN	0.878	0.880	0.878	0.878	0.828	0.871	0.849	0.913	0.883	0.898
	CAFE	0.898	0.899	0.898	0.898	0.864	0.879	0.872	0.921	0.911	0.916
	COOLANT	0.881	0.881	0.881	0.881	0.853	0.846	0.850	0.899	0.904	0.902
	MoPeD	**0.929**	**0.929**	**0.929**	**0.929**	**0.899**	**0.922**	**0.911**	**0.949**	**0.933**	**0.941**
Pheme	EANN	0.782	0.774	0.782	0.777	0.649	0.558	0.600	0.826	0.875	0.850
	MVAE	0.792	0.784	0.792	0.784	0.685	0.540	0.604	0.824	0.897	0.859
	SpotFake	0.769	0.758	0.769	0.760	0.636	0.496	0.557	0.808	0.882	0.844
	SpotFake+	0.823	0.822	0.823	0.822	0.706	0.681	0.694	0.870	0.882	0.876
	SAFE	0.813	0.807	0.813	0.803	0.753	0.540	0.629	0.829	0.926	0.875
	CARMN	0.805	0.831	0.805	0.814	0.549	0.721	0.623	0.912	0.829	0.869
	MFAN	0.875	0.882	0.875	0.877	0.756	**0.850**	0.800	0.934	0.886	0.909
	CAFE	0.891	0.891	0.892	0.891	0.826	0.796	0.810	0.917	0.930	0.923
	COOLANT	0.894	0.895	0.894	0.894	0.796	0.833	0.814	**0.934**	0.917	0.925
	MoPeD	**0.904**	**0.903**	**0.904**	**0.903**	**0.851**	0.814	**0.833**	0.924	**0.941**	**0.932**

TABLE II
STATISTICS OF THREE DATASETS

Dataset	Real news	Fake news	Images
Weibo	3345	3499	6844
Weibo-19	877	590	1467
Pheme	1428	590	2018

IV. EXPERIMENT

In this section, we evaluate the performance of the proposed model on three public datasets. The corresponding experiment results are shown in Table I.

A. Experimental Setup

1) Datasets: Three public real-world datasets are used, i.e., Weibo [37], Weibo-19 [38], and Pheme [39]. Table II shows their statistics.

1) Weibo is one Chinese dataset, collected from the most popular social media and verified by Xinhua News Agency. In our experiments, 2435 real news and 2356 fake news are used for training, 463 fake news and 221 real news are used for validation, and 680 fake news and 689 real news are used for testing.
2) Weibo-19 is one simplified version of Weibo dataset. In our experiments, 617 real news and 409 fake news are used for training, 146 posts are used for validation, and 295 news are used for testing.
3) The Pheme dataset consists of tweets from the Twitter platform, covering five breaking news. In our experiments, the training set contains 999 real news and 413 fake news, the validation set contains 221 posts, and the testing set contains 385 news.

For the original Weibo dataset, the training set contains 3749 fake news and 3783 real news; the testing set contains 1000 fake news and 996 real news. In our experiment, we filter out the news articles lacking complete text or image, and remove the duplicated and low-quality texts to ensure the dataset quality. Therefore, the amount of data in the Weibo dataset is different from the dataset used in other methods.

2) Implementation Details: In modal-specific encoders, the provided word vectors [34] are employed for word embeddings with a dimension size of 300. The input image is resized into 224 $\times$ 224. For modal-consistent encoders, the text is truncated or padded to a length of 77 tokens. The pretrained

Algorithm 1 MoPeD for Multimodal Fake News Detection

Input: The multimodal news data $\mathbf{P} = [\mathbf{T}, \mathbf{V}]$ including text $\mathbf{T}$ and image $\mathbf{V}$,the label of news y, and the iteration number $epoch$.

Output: The classifying results $\hat{y}$.

for $i = 1$ to $epoch$ **do**
- Generate the representations f_t^{sc} and f_v^{sc} of existing modality data by Eq.(1), Eq.(2) and Eq.(3);
- Concate the text and viual features with each memory information as the final representations f_t and f_v;
- Generate the enhanced unimodal representations $\hat{f}_{vv}$ and $\hat{f}_{tt}$ by Eq.(14);
- Obtain the visual-enhanced text feature f_{tv} by Eq.(16);
- Generate the final multi-modal fused features $\hat{f}_{tv}$ and $\hat{f}_{vt}$ by Eq.(17);
- Calculate the heterogeneity score h by Eq.(20);
- Obtain the weighed uni-modal features and multi-modal features $\hat{f}_{tt}{}', \hat{f}_{vv}{}', \hat{f}_{tv}{}', \hat{f}_{vt}{}'$ by Eq.(21)
- Concate the four features as the final news representation $\hat{f}$ by Eq.(22);
- Predict the category of the news by Eq.(23);
- Update the entire network by minimizing the objective function Eq.(24).

end for

CLIP model used is "ViT-B/32" with a dimension size of 512. The length of the learnable memory information is set to 50.

These three datasets are split for training, validation, and testing with a ratio of 7 : 1 : 2. Each dataset is individually tested by the proposed model, resulting in three sets of experiments conducted for the three datasets. For the Weibo-19 and Pheme datasets, the MFAN [16] model provides high-quality preprocessed data, in which each news article consists of one unique text and one image. For the original Weibo dataset, a news article's text may be accompanied by multiple images or no images. To construct standard news samples, we retain only one high-quality image for each news article or filter out articles without images. Following the text-preprocessed strategy in MFAN, we use the stopword list to eliminate frequently occurring words that do not carry significant semantics in the text, which highlights the critical words and reduces the negative interference to the model. Subsequently, we use the Chinese word segmentation dictionary to segment the text and then obtain word vectors of the same length through the specified Word2Vec model. If one word is not found in the given Word2Vec model, we use randomly generated word vectors instead.

The batch size is set to 64 for training and 50 for testing, the model using Adam with a learning rate of 0.001 for the Weibo dataset and 0.002 for the Weibo-19 and Pheme datasets. The learning rate setting refers to MFAN [16]. The head number of the MHA mechanism is set to 8. In all experiments, the model is trained for 20 epochs on a single NVIDIA RTX 2080 GPU, and the random seed is fixed to ensure result reproducibility. The code is available at https://github.com/littlesunnywgc/MoPeD.

3) Baselines: In order to make a fair comparison, we exclusively consider methods with publicly available source code. The selected representative multimodal methods are as follows.

1) EANN [12] utilizes event classification as an auxiliary task to improve the classification results for urgent news.
2) MVAE [30] adopts VAE to reconstruct different modalities, effectively distinguishing news by comparing the differences between the original and the reconstructed data.
3) Spotfake [25] concatenates visual features extracted by VGG-19 and text features extracted by BERT and inputs the fusion representation to the classifier to judge the authenticity of news.
4) Spotfake+ [26] further improves this fusion strategy, which uses the pretrained XLNet to handle the full-length article detection.
5) CARMN [28] combines the cross-modal attention residual network and the multichannel CNN to maintain the unique information of different modalities and reduce the noise caused by modal differences.
6) SAFE [4] transforms visual information into text information and calculates the text and visual cosine similarity to aid in identifying fake news.
7) MFAN [16] incorporates text, visual, and social graph features into the same framework based on graph attention network (GAT) to improve classification accuracy.
8) CAFE [17] proposes an ambiguity-aware multimodal method and uses the unimodalities common semantic sharing auxiliary tasks to address the misclassification caused by modal inconsistency.
9) COOLANT [18] utilizes contrastive learning and cross-modal consistency learning tasks to effectively align visual and linguistic representations, thus establishing more accurate fake news detection across different modalities.

B. Performance Comparison

The MoPeD model is compared with the state-of-the-art multimodal fake news detection methods, and the results are presented in Table I. Reproduced versions of EANN, MVAE, Spotfake, Spotfake+, SAFE, CARMN, MFAN, CAFE, and COOLANT, which provide publicly available source code, are used for the comparison across three datasets. The evaluation metrics include accuracy, precision, recall, and $F1$-score. The results indicate that MoPeD outperforms all the compared methods in terms of accuracy on the three datasets, achieving the highest accuracy of **0.883**, **0.929**, and **0.904** on the Weibo, Weibo-19, and Pheme, respectively. These results demonstrate a superiority of **3.4**%, **3.1**%, and **1.0**% over the state-of-the-art methods. In addition, the MoPeD ranks first or second in all tests in terms of accuracy, precision, recall, and $F1$ score, demonstrating its effectiveness in fake news detection.

The observed inferior performance of EANN, MVAE, Spotfake, and Spotfake+ methods can be attributed to their usage of different encoders for text and image feature extraction, leading to a significant semantic gap between modalities.

Similarly, SAFE and MFAN adopt straightforward feature extraction and aggregation. However, MFAN outperforms SAFE by constructing a social graph that includes additional information such as additional posts, comments, and users information, introducing a lot of crucial information to the entire model. The alignment module in MFAN also reduces the heterogeneity gap between features, contributing to its superior performance. CAFE and COOLANT achieve the best performance on our three datasets because they both align the semantic space across modalities to a certain extent by cross-modal alignment module and contrastive learning, respectively. Additionally, the cross-modal ambiguity learning used in these two methods enables weighted aggregation for different modal features. However, CAFE and COOLANT extract unimodal features solely through their mode-specific encoders, inherently introducing heterogeneity into the features used for alignment and weighting.

The superiority of MoPeD over the state-of-the-art methods can be attributed to several key factors. First, the utilization of the CLIP pretrain encoders in the dual encoding module can extract text and image features with content consistency in the same semantic space. The combination of these two encoders ensures that the determinative information is comprehensively captured. Second, the multilevel cross-modality fusion module facilitates deep interaction between multimodal information, thereby enhancing the understanding of intrinsic meaning. Lastly, the heterogeneity score, calculated through the KL divergence of unimodal distributions, enables the model to automatically perceive the importance of determinative factors in different features.

C. Ablation Study

To further assess the impact of each important module in MoPeD on overall performance, seven sets of experiments on these three datasets are conducted. For each experiment, a different component is removed and the model is retrained. The compared variants of MoPeD are implemented as follows.

1) *MoPeD w/o P:* The modality perception learning module is removed, and the heterogeneity score is not used. The unimodal features and multimodal features are directly concatenated for classification.
2) *MoPeD w/o U:* The modality perception learning module is not utilized, and the unimodal features are removed for concatenation. Only the multimodal fusion features are concatenated as the input to the classifier.
3) *MoPeD w/o C:* The CLIP-related modal-consistent encodes are removed. The classification process is performed using only modal-specific features.
4) *MoPeD w/o F:* The multilevel cross-modality fusion module is removed, and a single-level fusion is used to obtain multimodal features for modality perception learning.
5) *MoPeD w/o M:* The memory information is removed, and the unimodal features are only represented by the dual encoding module.
6) *MoPeD w/o S:* The MHSA module is removed, and the unimodal features are directly utilized for the multilevel fusion and modality perception learning.
7) *MoPeD w/o C+M:* The CLIP-related modal-consistent encoders and the memory information are removed together.

TABLE III
ABLATION STUDY ON THE ARCHITECTURE DESIGN OF MOPED ON THREE DATASETS

	Method	Accuracy	F1-score	
			Fake News	Real News
Weibo	MoPeD	**0.883**	**0.883**	0.883
	w/o P	0.863	0.859	0.866
	w/o U	0.862	0.852	0.87
	w/o C	0.831	0.817	0.842
	w/o F	0.871	0.868	0.875
	w/o M	0.882	0.878	**0.886**
	w/o S	0.877	0.871	0.883
	w/o C+M	0.828	0.818	0.836
Weibo-19	MoPeD	**0.929**	**0.911**	**0.941**
	w/o P	0.909	0.873	0.928
	w/o U	0.898	0.870	0.916
	w/o C	0.868	0.843	0.886
	w/o F	0.919	0.896	0.933
	w/o M	0.922	0.901	0.936
	w/o S	0.902	0.874	0.919
	w/o C+M	0.847	0.816	0.870
Pheme	MoPeD	**0.904**	**0.833**	**0.932**
	w/o P	0.896	0.821	0.93
	w/o U	0.885	0.804	0.919
	w/o C	0.852	0.776	0.889
	w/o F	0.862	0.780	0.9
	w/o M	0.899	0.825	0.929
	w/o S	0.855	0.786	0.890
	w/o C+M	0.844	0.725	0.891

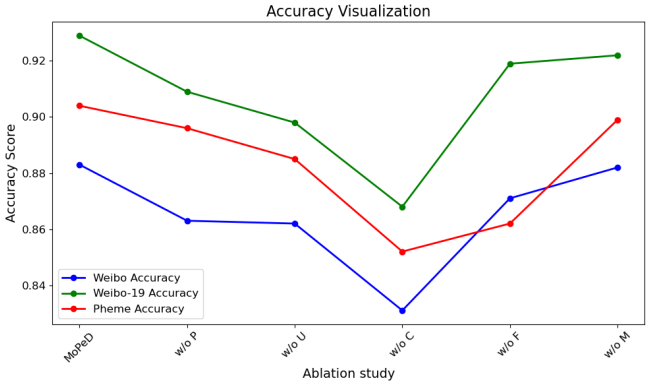


Fig. 3. Statistical analysis of the accuracy of ablation experiments on three datasets.

The results of the ablation study are reported in Table III, and the accuracy visualization is shown in Fig. 3. It can be observed that removing any module results in a varying degree of deterioration in the model's performance, which demonstrates the effectiveness of each component.

Table III reveals that MoPeD w/o C records the poorest results across all datasets, underscoring the effectiveness of employing dual encoders to extract comprehensive

Fig. 4. Examples of multimodal news explain the MoPeD. (a) "Woman in India gave birth to 11 fetuses at a time, a world record! Congratulations!" (b) New species of fish found at Arkansas. (c) "Happy New Year! Auspicious, happy, well-being!"

information, and some determinative information, e.g., modality consistency, has a crucial impact on the classification results, requiring the CLIP pretrained encoder to mine. MoPeD w/o M shows its decreased performance. This effectively verifies the advantage of the memory information module, which can remember the discriminative mode of real/fake news. MoPeD w/o F yields weaker performance, proving that deeply understanding the complex implicit information via multilevel fusion can also help improve the performance significantly.

To analyze the effect of the modality perception learning module on the performance of MoPeD, the results of MoPeD w/o P and MoPeD w/o U are compared with the complete model. It can be found that if the heterogeneity score is not calculated to adaptively weight the unimodal features and the multimodal features, the detection effect is significantly degraded, and the result has become worse if the unimodal features are not applicable to classification. This indicates that if the contributions of features from different sources are treated equally, misleading information may dominate the decision-making process.

V. Conclusion

In this article, we propose MoPeD, a model designed to extract comprehensive information for multimodal fake news detection and adaptively perceiving the contribution of determinative factors within unimodal and multimodal features. The feature extraction process incorporates the CLIP pretrain model as a modal-consistent encoder, working with the modal-specific encoders to extract complete information. Furthermore, we use memory information as a supplement to improve the adaptability of the unimodal features. The multilevel cross-modality fusion module is then implemented during feature fusion to delve into the implicit meaning hidden in modalities. Finally, the modality perception learning module orchestrates the aggregation of various features according to the heterogeneity score. Experiments on three widely used datasets demonstrate that MoPeD consistently outperforms the state-of-the-art methods by a large margin.

Appendix

By carefully analyzing the contents of text and image in the news, we design the modality perception learning module to highlight the decisive factor or modality. For instance, this module enables us to promptly identify fake news attributed to manipulated images or blatant rumors.

1) *Textual Modality*: In Fig. 4(a), classifying the news as truth is straightforward due to the striking resemblance between its textual and visual contents. However, the exaggerated text expressions suggest that it is fake news, thereby indicating that the decisive factor should be the unimodal text information.
2) *Visual Modality*: In Fig. 4(b), the image demonstrates clear traces of tampering, so the decisive factor should be unimodal visual information. Regarding the high similarity between the textual and visual contents, relying solely on a single modality may be more reliable for judging.
3) *Cross-Modality*: In Fig. 4(c), while the firework image and the "Happy New Year" text convey different semantics, they collectively evoke a shared atmosphere of happiness, confirming it as genuine news. In this instance, the decisive factor should be the combination of multiple modalities.

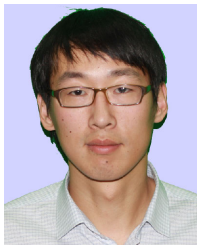
References

[1] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the role of visual content in fake news detection," in *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*. Cham, Switzerland: Springer, 2020, pp. 141–161.

[2] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors on Twitter by promoting information campaigns with generative adversarial learning," in *Proc. World Wide Web Conf.*, May 2019, pp. 3049–3055.

[3] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 518–527.

[4] X. Zhou, J. Wu, and R. Zafarani, "Similarity-aware multi-modal fake news detection," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2020, pp. 354–367.
[5] Q. Nan, J. Cao, Y. Zhu, Y. Wang, and J. Li, "MDFEND: Multi-domain fake news detection," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manag.*, 2021, pp. 3343–3347.
[6] Y. Zhu et al., "Memory-guided multi-view multi-domain fake news detection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 7178–7191, Jul. 2022.
[7] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? Using satirical cues to detect potentially misleading news," in *Proc. 2nd Workshop Comput. Approaches Deception Detection*, 2016, pp. 7–17.
[8] S. De Sarkar, F. Yang, and A. Mukherjee, "Attending sentences to detect satirical fake news," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 3371–3380.
[9] T. Chen, X. Li, H. Yin, and J. Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2018, pp. 40–52.
[10] Y. Li, S. Wang, Q. Tian, and X. Ding, "A survey of recent advances in visual feature detection," *Neurocomputing*, vol. 149, pp. 736–751, Feb. 2015.
[11] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017.
[12] Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 849–857.
[13] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 153–162.
[14] P. Qi et al., "Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1212–1220.
[15] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 2560–2569.
[16] J. Zheng, X. Zhang, S. Guo, Q. Wang, W. Zang, and Y. Zhang, "MFAN: Multi-modal feature-enhanced attention networks for rumor detection," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 2413–2419.
[17] Y. Chen et al., "Cross-modal ambiguity learning for multimodal fake news detection," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2897–2905.
[18] L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, and S. Wang, "Cross-modal contrastive learning for multimodal fake news detection," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5696–5704.
[19] J. Ma et al., "Detecting rumors from microblogs with recurrent neural networks," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 3818–3824.
[20] S. S. Birunda and R. K. Devi, "A novel score-based multi-source fake news detection using gradient boosting algorithm," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 406–414.
[21] M. H. Goldani, R. Safabakhsh, and S. Momtazi, "Convolutional neural network with margin loss for fake news detection," *Inf. Process. Manage.*, vol. 58, no. 1, Jan. 2021, Art. no. 102418.
[22] T. Jiang, J. P. Li, A. U. Haq, and A. Saboor, "Fake news detection using deep recurrent neural networks," in *Proc. 17th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2020, pp. 205–208.
[23] T. Jiang, J. P. Li, A. U. Haq, A. Saboor, and A. Ali, "A novel stacking approach for accurate detection of fake news," *IEEE Access*, vol. 9, pp. 22626–22639, 2021.
[24] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: Characterizing and identifying fake images on Twitter during hurricane sandy," in *Proc. World Wide Web Conf.*, 2013, pp. 729–736.
[25] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 39–47.
[26] 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," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13915–13916.
[27] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1942–1951.
[28] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Inf. Process. Manage.*, vol. 58, no. 1, Jan. 2021, Art. no. 102437.
[29] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Inf. Process. Manage.*, vol. 60, no. 1, Jan. 2023, Art. no. 103120.
[30] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, May 2019, pp. 2915–2921.
[31] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang, "Multimodal fake news detection via CLIP-guided learning," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2023, pp. 2825–2830.
[32] Q. Ying, X. Hu, Y. Zhou, Z. Qian, D. Zeng, and S. Ge, "Bootstrapping multi-view representations for fake news detection," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 5384–5392.
[33] Y. Wang, S. Qian, J. Hu, Q. Fang, and C. Xu, "Fake news detection via knowledge-driven multimodal graph convolutional networks," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 540–547.
[34] C. Yuan, Q. Ma, W. Zhou, J. Han, and S. Hu, "Jointly embedding the local and global relations of heterogeneous graph for rumor detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 796–805.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[36] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–23.
[37] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 795–816.
[38] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun, "CED: Credible early detection of social media rumors," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3035–3047, Aug. 2021.
[39] A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumor detection in social media," in *Proc. Int. Conf. Social Informat.*, 2017, pp. 109–123.



Boyue Wang received the B.Sc. degree in computer science from Hebei University of Technology, Tianjin, China, in 2012, and the Ph.D. degree from Beijing University of Technology, Beijing, China, in 2018.

He is currently an Associate Professor with Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, Beijing University of Technology. His current research interests include multimodal analysis, knowledge graph, manifold learning, and clustering.

Guangchao Wu received the B.Sc. degree from Hebei University, Hebei, China, in 2023. She is currently pursuing the master's degree with Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, Beijing University of Technology, Beijing, China.

Her research interests include multimodal fake news detection, computer vision, and pattern recognition.

Xiaoyan Li received the B.Sc. degree from Tianjing University, Tianjin, China, in 2014, the M.S. degree from Brown University, Providence, RI, USA, in 2016, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2022.

She is currently a Post-Doctoral Fellow with Beijing University of Technology, Beijing. Her research interests are mainly computer vision, weakly supervised learning, 2-D and 3-D object detection, semantic segmentation, and knowledge graph.

Junbin Gao (Associate Member, IEEE) received the B.Sc. degree in computational mathematics from the Huazhong University of Science and Technology (HUST), Wuhan, Hubei, China, in 1982, and the Ph.D. degree from Dalian University of Technology, Dalian, China, in 1991.

From 1982 to 2001, he was an Associate Lecturer, a Lecturer, an Associate Professor, and a Professor with the Department of Mathematics, HUST. From 2001 to 2005, he was a Senior Lecturer and a Lecturer of computer science with the University of New England, Armidale, NSW, Australia. He was a Professor of computer science with the School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW, Australia. He is currently a Professor of big data analytics with The University of Sydney Business School, The University of Sydney, Camperdown, NSW, Australia. His main research interests include machine learning, data analytics, Bayesian learning and inference, and image analysis.

Yongli Hu (Member, IEEE) received the Ph.D. degree from Beijing University of Technology, Beijing, China, in 2005.

He is currently a Professor with the Faculty of Information Technology, Beijing University of Technology, where he is also a Researcher with Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology. His research interests include computer graphics, pattern recognition, and multimedia technology.

Baocai Yin (Member, IEEE) received the Ph.D. degree from Dalian University of Technology, Dalian, China, in 1993.

He is currently a Professor with the Faculty of Information Technology, Beijing University of Technology, Beijing, China, where he is also a Researcher with Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology. His research interests cover multimedia, multifunctional perception, virtual reality, and computer graphics.

Dr. Yin is a member of China Computer Federation.