Covariance Matrix-Driven Image Channel Allocation for Multimodal Fake News Detection

Zongliang Han[1,2], Wenyu Guo[1], Guoqing Jin[1]*, Yang Liu[2]*,
Yan Song[3], Dong Yu[1], Min Wang[1]
[1]State Key Laboratory of Communication Content Cognition, People's Daily Online, China
[2]University of Chinese Academy of Sciences, China
[3]University of Science and Technology of China, China
{guowenyu, jinguoqing, yudong, wangmin}@people.cn
liuyang22@ucas.ac.cn, hanzongliang22@mails.ucas.ac.cn, clksong@gmail.com

Abstract

With the widespread proliferation of the Internet, the spread of fake news has accelerated significantly, evolving from single-text content to multimodal forms that include images and videos. The task of Multimodal Fake News Detection (MFND) takes both text and relevant images as input for fake news identification. However, issues such as image noise and inaccurate focus of visual features often lead to insufficient attention to critical information within images during multimodal fusion. To effectively address these challenges, we propose a covariance matrix-driven image channel allocation method. This method first expands the number of original channel maps, then evaluates the importance of image channels through the covariance matrix and assigns importance scores to the expanded channel maps, thereby redirecting the focus of visual features. Subsequently, we design a multimodal fusion strategy based on a multilayer Co-Attention mechanism to achieve dynamic fusion across modalities. Finally, a contrastive learning loss is introduced to enhance the alignment between textual and visual modalities. Extensive experiments demonstrate that our method achieves state-of-the-art performance on three public multimodal fake news detection benchmark datasets.

1 Introduction

With the rapid expansion of social media, online platforms such as Twitter and Weibo have become crucial information sources for the public. Although these platforms facilitate the dissemination of news, they provide fertile ground for the proliferation of fake news. Consequently, fake news detection has attracted significant attention from the research community. In comparison to plain text, multimodal information is more intuitive and can elicit stronger emotional responses from users (Shu et al., 2017). It is frequently used as a medium for news dissemination (Wu et al., 2015) and has therefore become a vital basis for fake news detection. Multimodal Fake News Detection (MFND) simultaneously takes the text and a relevant image as input for fake news detection, has emerged as a research focus and is gaining increasing attention.

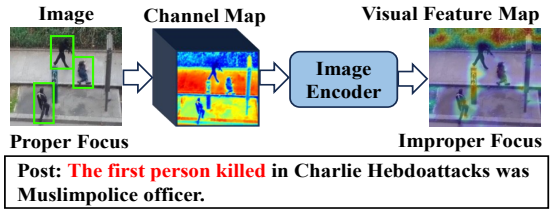

Figure 1: An example illustrating the image noise and improper focus of visual feature in multimodal samples.

Most existing MFND methods focus on designing an efficient multimodal fusion framework to bridge the semantic gap between images and texts. Previous studies have fed images and texts directly into multimodal encoders, employing fusion methods such as concatenation (Jin et al., 2017; Wang et al., 2018) and attention mechanisms (Wu et al., 2021; Wang and Sui, 2021), followed by fake news detection. However, these methods overlook the interference caused by noise in the images, which can disrupt fake news detection. Some studies attempt to alleviate this issue by extracting salient visual entities and feeding them into the image encoder (Qi et al., 2021). Although this method reduces noise to some extent, it inevitably overlooks some important features beyond the visual entities in the image, which also play a significant role in fake news detection. We summarize two main shortcomings of current methods:

Noise of Multimodal Samples It is often challenging to acquire an image that precisely corresponds to the textual information. As shown on the left of Figure 1, the objects enclosed in the green box of the image correspond to the phrase "*Police*

* Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 1101–1115

forces take position" in the post sentence, representing valuable visual information for fake news detection. In the task of multimodal fake news detection (MFND), visual noise primarily originates from background elements irrelevant to the text, such as shrubs, roads, and street lights. These elements may introduce cognitive bias and degrade detection performance. Therefore, exploring effective image denoising techniques is essential for improving the accuracy of the MFND task.

Improper Focus of Visual Feature Map Although existing methods have enabled models to focus on relevant visual information during the multimodal fusion process, they have not fundamentally addressed the issue at the visual feature level. As illustrated in Figure 1, the image is mapped into channel maps and fed into the image encoder to obtain visual features. However, the visual feature assigns weights to the irrelevant background region, resulting from improper image channel allocation of the channel map. This can mislead the attention assignment of image encoders and interfere with the multimodal fusion process. We posit that if the model can identify salient information from the perspective of image channels and modulate its focus on the size distribution of the channel map, it can optimize the encoded visual features without sacrificing global visual information. This approach would thereby achieve the goal of filtering image noise at the visual feature level.

In order to address the aforementioned challenges, we propose a **Co**variance **M**atrix-Driven **I**mage **C**hannel **A**llocation method (abbreviated as **CoM-ICA**) to effectively guide visual features to focus on salient information. Firstly, we expand the original image channel maps into richer channel representations through convolutional neural networks. Secondly, we evaluate the importance of each channel through the calculation of channel scores based on channel covariance matrix. Thirdly, we allocate the calculated channel scores to the expanded channel maps, and reconstruct the channel maps back to their original dimensions through another convolutional neural network. For the multimodal fusion method, we propose an image-feature-dominated multilayer Co-Attention Transformer to dynamically and selectively integrate text and visual features. Contrastive learning loss is utilized to enhance multimodal alignment. Overall, we employ a covariance matrix-based method to calculate importance scores to image channels, enabling the model to focus on crucial visual information while minimizing information loss. By optimizing both the visual feature and cross-modal fusion, we enhance the performance of multimodal fake news detection.

The main contributions of our work are as follows:

- We propose a covariance matrix-based channel importance evaluation method that models the correlations and linear relationships between channels to accurately assess the importance of each image channel, thereby quantifying its contribution to the overall feature representation.
- A channel allocation method is proposed to enhance visual representation by expanding image channels and assigning importance scores to the expanded channels. Additionally, we design a multi-layer fusion approach to facilitate multimodal collaboration.
- Extensive experiments are conducted on three public datasets, and the results demonstrate the effectiveness and superiority of the proposed method.

2 Related Work

2.1 Multimodal Fake News Detection

Multimodal Fake News Detection models mainly achieve fusion by simply concatenating image and text features (Jin et al., 2017; Wang et al., 2018). To learn shared representations of multimodal information, MVAE (Khattar et al., 2019) proposes a multimodal variational autoencoder, reconstructing multimodal representations from a learned probabilistic latent model. Ma et al. (2024) integrate visual manipulation, textual emotion and multimodal inconsistency at event-level for fake news detection. Chen et al. (2023) conduct causal graphs to mitigate image biases. Some works utilize the consistency between modalities to promote fake news detection (Chen et al., 2022; Fung et al., 2021). Furthermore, considering the complexity of real-world scenarios, Zeng et al. (2025) propose the IMOL framework to address the challenges of multi-domain distribution and incomplete modalities (e.g., missing audio or text) in news videos. However, these models either assume the integrity of all modalities or do not specifically address the noise issues in images.

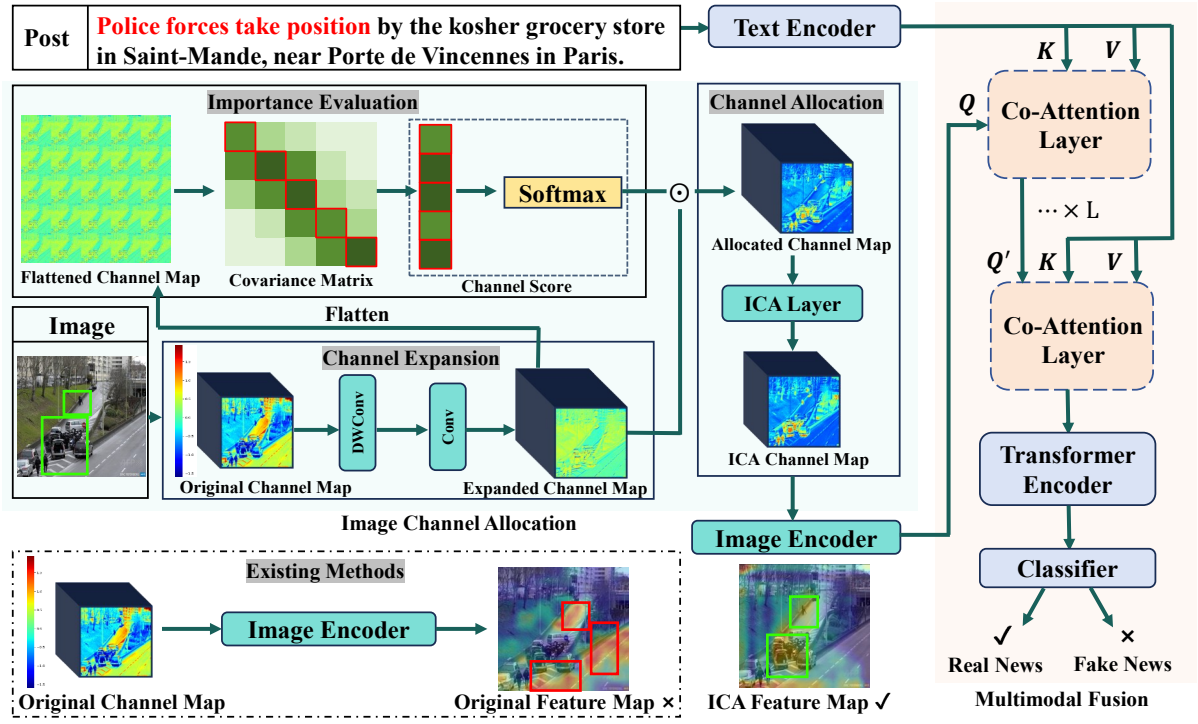

Figure 2: The overall architecture of CoM-ICA, which consists of two main modules: the Image Channel Allocation module and the Multimodal Fusion module. The Image Channel Allocation module enhances the focus of visual features, while the Multimodal Fusion module dynamically integrates multimodal information.

2.2 Image Denoising

The noise in news images primarily stems from irrelevant information. Fang and Feng (2022) model the semantic information in visual features to obtain a text-guided visual feature representation, while Ye et al. (2022) employ a position-aware masking mechanism to mask the noise components in visual features. However, these methods filter out noise in the multimodal fusion level using the text modality. As for the visual feature level, a direct approach involves extracting visual entities by identifying key objects or regions in news images (Qi et al., 2021). Futeral et al. (2023) introduce prior knowledge of text-image alignment to filter out noisy visual information. These methods may lead to information loss and be constrained by external knowledge. To address these issues, we propose an importance-based channel allocation method, filtering the noise at the visual feature level without information loss.

3 Methodology

Figure 2 illustrates the overall architecture of CoM-ICA, which mainly consists of an image channel allocation module and a multimodal fusion module. In the image channel allocation module, we expand the number of channels, evaluate their importance using a covariance matrix, assign weights in a high-dimensional space, and then restore the channel dimension to its original size, thereby guiding visual attention to focus on critical information. In the multimodal fusion module, we introduce an image-feature-dominated multilayer Co-Attention Transformer architecture to fuse textual and visual information, combined with a contrastive learning loss to enhance cross-modal alignment. Finally, a classifier is used for fake news detection.

3.1 Image Channel Allocation(ICA)

Considering that improper allocation of image channels can disrupt the focus of visual features, we assign different weights to channels based on their importance. This module consists of three stages: The first stage is Channel Expansion, where the channel map is expanded from $3 \times 244 \times 244$ to $3n \times 244 \times 244$ to enhance feature representation. The second stage is Importance Evaluation, where channel scores are calculated using the covariance matrix, and channels with larger variances are regarded as more critical. The third stage is Channel Allocation, where the scores are used to weight the expanded channel map, assigning differ-

ent weights to guide visual features to focus on the key information.

Channel Expansion In this stage, an expanded depthwise convolution (Chollet, 2017) with a kernel size of 3×3 is used to encode local information $\mathbf{I}_h$ from the original channel map $\mathbf{I}_o$. Then, a convolution followed by a GELU activation function is applied for cross-channel linear combination, generating the expanded channel map $\mathbf{I}_d$:

$$\mathbf{I}_h = \text{DWConv}_{3\times3}\left(\mathbf{I}_o\right), \tag{1}$$

$$\mathbf{I}_d = \text{Conv}_{1\times1}\left(\phi(\mathbf{I}_h)\right), \tag{2}$$

where $\mathbf{I}_h \in \mathbb{R}^{H\times W\times C}$, $\mathbf{I}_d \in \mathbb{R}^{H\times W\times C'}$, and $C' = nC$. $\phi(\cdot)$ denotes the GELU activation function.

Importance Evaluation Mathematically, variance measures the dispersion of a single random variable. For two or more variables, this dispersion is described by covariance. For random variables X and Y, covariance is defined as:

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])\left(Y - \mathbb{E}\left[Y\right]\right)\right]. \tag{3}$$

In the multivariate case, covariance generalizes to the covariance matrix, which describes the correlation and linear relationships among variables. For an n-dimensional variable $\mathbf{X} = (X_1, \ldots, X_n)^T$, the covariance matrix is:

$$\mathbf{C} = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix}, \tag{4}$$

where the diagonal elements represent the variances of each variable:

$$c_{ii} = \text{Var}(X_i), \tag{5}$$

indicating the fluctuation level of variable X_i.

In the ICA module, we leverage this property to model channel importance. Specifically, the expanded channel map $\mathbf{I}_d$ is flattened into $\mathbf{I}_D \in \mathbb{R}^{C'\times(H\cdot W)}$, and channel-wise mean subtraction is performed to center the data, resulting in $\mathbf{I}_C$,

$$\mathbf{I}_C = \mathbf{I}_D - \frac{1}{H\cdot W}\sum_{i=1}^{H\cdot W}\mathbf{I}_D[:, i]. \tag{6}$$

We first compute the covariance matrix for the centered channel map $\mathbf{I}_C = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{C'})$, then extract its diagonal elements and normalize them using the Softmax function to obtain the importance scores α_i for each channel,

$$\mathbf{C} = \frac{1}{H\cdot W - 1}\mathbf{I}_C \cdot \mathbf{I}_C^T, \tag{7}$$

$$v_i = \text{Var}(Y_i), \quad i = 1, 2, \ldots, C', \tag{8}$$

$$\alpha_i = \text{softmax}(v_i) = \frac{e^{v_i}}{\sum_{j=1}^{C'} e^{v_j}}, \tag{9}$$

under this mechanism, channels with larger variances contain more information and are thus assigned higher weights, while channels with smaller variances carry less information and receive correspondingly lower weights. This effectively suppresses noise and enhances useful features. Based on this, we can accurately assess the importance of each channel by modeling the variability of its features, assigning higher weights to the key information with greater variability.

Channel Allocation Then, based on the channel scores $\boldsymbol{S} = [\alpha_1, \alpha_2, \ldots \alpha_{C'}]$, the expanded channel map $\mathbf{I}_d$ is weighted and scaled to enhance the weights of important channels while suppressing the influence of less important channels:

$$\mathbf{I}_s = \mathbf{I}_d \odot \mathbf{S} \in \mathbb{R}^{H\times W\times C'}, \tag{10}$$

where $\odot$ denotes element-wise multiplication. Finally, a 1×1 convolution is applied to the weighted channel map $\mathbf{I}_s$ to restore it back to the 3-channel ICA channel map $\mathbf{I}'$:

$$\mathbf{I}' = \text{Conv}_{1\times1}\left(\mathbf{I}_s\right) \in \mathbb{R}^{H\times W\times C}. \tag{11}$$

3.2 Multimodal Fusion

Text Encoder As for the text encoder, we feed the raw Text into the pre-trained RoBERTa model (Liu et al., 2019) to generate text features:

$$\mathbf{T} = \text{RoBERTa}\left(\text{Text}\right) \in \mathbb{R}^{B\times L_T\times d_T}, \tag{12}$$

where B is the batch size, d_T is the dimension of the text features, and L_T represents the length of the post sentence.

Image Encoder As for the image encoder, we feed the ICA channel map $\mathbf{I}'$ into the ResNet-50 model (He et al., 2016) and extract the visual features through flattening and a linear layer with a ReLU activation function:

$$\mathbf{I} = \varphi\left(w_1 \cdot \text{Flatten}\left(\text{ResNet50}\left(\mathbf{I}'\right)\right) + b_1\right), \tag{13}$$

where $\mathbf{I} \in \mathbb{R}^{B \times d_I}$, $d_I = d_T$ and $\varphi(\cdot)$ denotes the ReLU activation function. Flatten represents the operation of flattening features into a one-dimensional vector, w_1 is the weight matrix of the fully connected layer, and b_1 is the bias term.

Multimodal Fusion Method The Multimodal Transformer (Xu et al., 2023) integrates image and text features through multi-head attention, achieving good results in previous studies (Qian et al., 2021; Zhang et al., 2023). However, these methods still lack sufficient modality alignment and fine-grained fusion, and the visual features tend to be weakened during fusion. To address this, we propose an image-feature-dominated multilayer Co-Attention Transformer that progressively refines the alignment between visual and textual features using multiple layers of Co-Attention (MHCA). In this model, visual and textual features interact through multiple Co-Attention layers, each incorporating residual connections and layer normalization (LN). The computation is as follows:

$$\mathbf{I}^{(l)} = \mathrm{LN}\left(\mathrm{MHCA}^{(l)}(\mathbf{I}^{(l-1)}, \mathbf{T}) + \mathbf{I}^{(l-1)}\right), \quad (14)$$

where $l \in [1, L]$, and $\mathrm{MHCA}^{(l)}$ denotes the Multi-Head Co-Attention at the l-th layer. The queries are derived from the image features, while the keys and values come from the textual features.

The output of the final Co-Attention layer $\mathbf{I}^{(L)}$ is fed into a multi-head self-attention (MHA) module and a feed-forward network (FFN) for feature refinement. Residual connections and layer normalization are applied after each sub-layer to stabilize training and preserve information,

$$\mathbf{I}_a^{(L)} = \mathrm{LN}\left(\mathrm{MHA}\left(\mathbf{I}^{(L)}\right) + \mathbf{I}^{(L)}\right), \quad (15)$$

$$\mathbf{F} = \mathrm{LN}\left(\mathrm{FFN}\left(\mathbf{I}_a^{(L)}\right) + \mathbf{I}_a^{(L)}\right), \quad (16)$$

where $\mathbf{F}$ is the output of our proposed multimodal fusion method.

3.3 Training Objects

After obtaining the fused feature vector, we input it into a multi-layer perceptron (MLP) for classification:

$$\hat{y} = \mathrm{MLP}(\mathbf{F}). \quad (17)$$

We utilize the cross-entropy loss as our first training object:

$$\mathcal{L}_p = -\frac{1}{N}\sum_{i=1}^{N}\left(y \log \hat{y} + (1-y)\log(1-\hat{y})\right), \quad (18)$$

where N represents the number of training samples, and y denotes the true label of each sample.

Contrastive Semantic Alignment. Since news images and textual semantics often exhibit partial misalignment or loose correlation, we introduce a multiscale contrastive learning module to explicitly bridge the semantic gap. Specifically, we adopt a refined NT-Xent loss function (Chen et al., 2020) that incorporates a Top-K soft-alignment strategy. Unlike standard contrastive frameworks that enforce a strict one-to-one identity mapping, our approach projects textual features $\mathbf{t}_i$ and visual features $\mathbf{i}_j$ into a shared d-dimensional manifold and applies L_2 normalization. We then calculate a temperature-scaled cosine similarity matrix $\mathbf{S}$, where $S_{ij} = (\mathbf{t}_i \cdot \mathbf{i}_j)/\tau$. To accommodate potential semantic overlaps in social media data, the Top-K strategy identifies the most relevant cross-modal pairs as potential semantic neighbors rather than relying solely on the diagonal elements. The contrastive loss for the i-th sample is defined as:

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S_{ii})}{\sum_{j=1}^{N}(S_{ij})}, \quad (19)$$

where τ is the temperature coefficient. This objective incentivizes the model to aggregate coreferential cross-modal features while penalizing deceptive inconsistencies, effectively transforming semantic discrepancies into measurable spatial separation.

Finally, the total training objective is formulated by integrating the classification loss $\mathcal{L}_p$ and the contrastive loss $\mathcal{L}_c$:

$$\mathcal{L} = \mathcal{L}p + \lambda\mathcal{L}c, \quad (20)$$

where λ is a learnable weighting coefficient that adaptively balances the two components.

4 Experiments

4.1 Experimental Settings

We evaluate the CoM-ICA module on three widely used fake news detection benchmarks: Pheme (Zubiaga et al., 2017), Weibo (Jin et al., 2017), and CFND (Zhang et al., 2024). The Pheme dataset consists of tweets from Twitter, focusing on five major

Method	Pheme				CFND				Weibo			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
MVAE(Khattar et al., 2019)	0.776	0.735	0.723	0.728	0.812	0.807	0.811	0.806	0.824	0.828	0.822	0.823
SAFE(Zhou et al., 2020)	0.807	0.787	0.789	0.791	0.795	0.789	0.804	0.796	0.851	0.849	0.849	0.849
SpotFake(Singhal et al., 2019)	0.845	0.809	0.836	0.822	0.830	0.825	0.841	0.833	0.873	0.873	0.874	0.873
CAFE(Chen et al., 2022)	0.832	0.796	0.794	0.795	0.826	0.827	0.846	0.837	0.840	0.840	0.841	0.840
MCAN(Wu et al., 2021)	0.861	0.830	0.840	0.835	0.845	0.831	0.784	0.807	0.899	0.899	0.899	0.899
KDIN(Sun et al., 2021)	0.846	0.815	0.804	0.809	0.847	0.813	0.846	0.830	0.893	0.894	0.892	0.893
LIIMR(Singhal et al., 2022)	0.870	0.848	0.831	0.839	0.852	0.817	0.834	0.826	0.900	0.882	0.823	0.847
BMR(Ying et al., 2023)	0.884	0.872	0.840	0.855	0.859	0.834	0.815	0.824	0.918	0.912	0.909	0.910
NLIN(Zhang et al., 2024)	0.903	0.875	0.883	0.879	0.874	0.848	0.841	0.844	0.922	0.917	0.922	0.919
Event-Radar(Ma et al., 2024)	0.901	0.883	0.878	0.880	-	-	-	-	0.919	0.928	0.910	0.919
CoM-ICA	**0.923**	**0.911**	**0.896**	**0.903**	**0.925**	**0.922**	**0.920**	**0.921**	**0.936**	**0.936**	**0.935**	**0.935**

Table 1: The accuracy, precision, recall, and F1-score of the fake news detection model on three datasets are presented. Bold indicates the best performance, while the second-best performance is underlined.

#	Model	Pheme			
		Accuracy	Precision	Recall	F1-Score
1	**CoM-ICA**	**0.923**	**0.911**	0.896	**0.903**
2	- w/o ICA	0.893	0.876	0.851	0.862
3	- w/o Channel Expansion	0.913	0.908	0.869	0.886
4	- w/o Channel Allocation	0.897	0.890	0.846	0.865
5	**CBAM**(Woo et al., 2018)	0.899	0.900	0.843	0.866
6	**ECA-Net**(Wang et al., 2020)	0.904	0.903	0.851	0.872
7	**GCT**(Yang et al., 2020)	0.908	0.902	0.864	0.880
8	**MCA**(Jiang et al., 2024)	0.910	0.895	0.877	0.885
9	- w/o $\mathcal{L}_c$	0.917	0.905	0.884	0.894
10	**Concat** (Wang et al., 2018)	0.853	0.849	0.770	0.796
11	**MMC** (Han et al., 2023)	0.882	0.856	0.846	0.851
12	**MMT** (Xu et al., 2023)	0.914	0.892	**0.897**	0.894
13	**TextQuery**	0.893	0.871	0.858	0.864

Table 2: The accuracy, precision, recall, and F1-score of the Pheme dataset in the ablation studies.

breaking news events. The Weibo dataset includes real news from authoritative Chinese media such as Xinhua News Agency, while fake news is collected from the Weibo platform and verified through its official fact-checking system. The CFND dataset is compiled from multiple Chinese fact-checking websites and official news sources, covering five distinct domains. In our experiments, we follow the dataset partitioning strategy used in NLIN (Zhang et al., 2024).

We implement CoM-ICA using PyTorch (Paszke et al., 2017), scikit-learn (Pedregosa et al., 2011), and Transformers (Wolf et al., 2020). We use the Adam optimizer and the learning rate is set to 5e-4, 1e-4 and 1e-4 for Pheme, CFND and Weibo dataset respectively. The batch size is set to 30, the temperature for the contrastive learning loss is set to 0.5, and the value of λ is initialized to 0.1. All the experiments are trained and evaluated using 1 Tesla V100 GPU. For the selection of other hyperparameters, please refer to Appendix B.

4.2 Main Results

We compare CoM-ICA with ten representative baselines on three fake news detection benchmarks. Table 1 presents the results and demonstrates that:

Firstly, the proposed CoM-ICA model significantly outperforms all the baseline models. Compared to the best-performing baseline model, CoM-ICA achieves improvements of 2%, 5.1% and 1.4% on the Pheme, CFND and Weibo datasets, respectively. This underscores the effectiveness of CoM-ICA in enhancing the performance of MFND task.

Secondly, CoM-ICA outperforms the event consistency model Event-Radar(Ma et al., 2024) and the external knowledge-based NLIN(Zhang et al., 2024). By focusing on key regions in images, CoM-ICA mines fine-grained correlations between images and text, fully leveraging the complementarity of multimodal information. This not only enhances feature representation but also improves overall performance in fake news detection, demonstrating the advantages of deep multimodal fusion.

Thirdly, CoM-ICA consistently outperforms other MFND models that actively explore multimodal fusion frameworks, such as CAFE (Chen et al., 2022) and MVAE (Khattar et al., 2019), underscoring the efficacy of our proposed multimodal fusion method.

5 Analysis

5.1 Effects of Image Channel Allocation

To investigate the impact of the Image Channel Allocation module, we conduct ablation studies and the results are presented in Table 2. When removing the Image Channel Allocation module (#2), we observe a 3% decrease in accuracy, indicating that ICA is able to reduce interference from irrelevant information and enhance the performance of fake news detection. When removing the Channel Expansion stage (#3), we also observe a 1% drop in performance, validating that this stage is able to minimize the information loss and capture complex

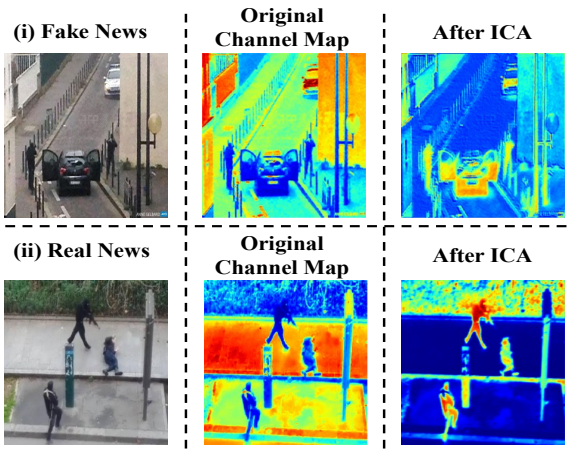

Figure 3: Use heatmaps to visualize the channel maps to demonstrate the effectiveness of the ICA method.

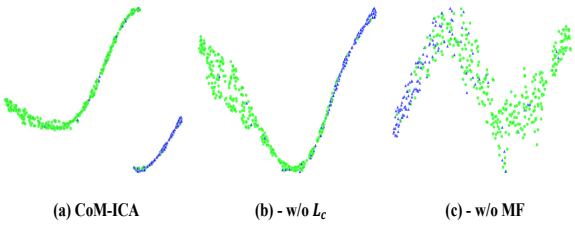

Figure 4: T-SNE visualization of learned representations. MF denotes Multimodal Fusion.

visual features. When removing the Channel Allocation stage (#4), the obvious result drop shows the necessity of our proposed Importance Evaluation stage and the improvement in model performance is attributed to the proper allocation of channels based on channel scores, rather than to the increase in parameters resulting from the Channel Expansion stage.

To further demonstrate the effectiveness of the ICA module intuitively, we compared the visualizations of the original channel maps and the channel maps processed by the ICA module, as shown in Figure 3. The results indicate that the feature maps guided by the ICA module focus more on key regions closely related to the news semantics, such as people and vehicles. Overall, through both quantitative and qualitative analyses, we have thoroughly validated the crucial role of the channel allocation mechanism in the ICA module in optimizing visual attention areas. This effectively enhances the discriminative ability of visual features, providing a more reliable foundation for subsequent fusion and decision-making.

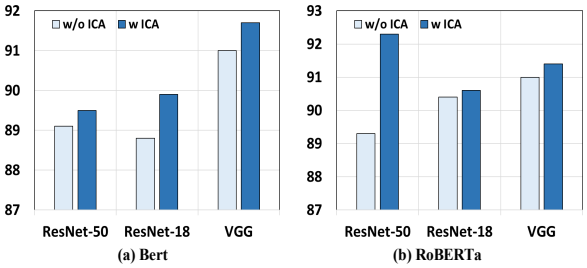

Figure 5: Accuracy on the Pheme dataset under various text and image encoders.

5.2 Effects of the Multimodal Fusion Method

Quantitative Analysis To verify the effectiveness of our proposed multimodal fusion method, we conduct ablation studies, with results shown in Table 2. Removing the contrastive learning loss (#9) led to a performance decline, highlighting its crucial role in modality alignment. Replacing our fusion architecture with **CONCAT** (#10), which directly concatenates text and visual features, a multimodal Transformer(**MMT**,#12), and **MMC** (#11), which fuses features using a cross-attention mechanism, all resulted in inferior performance compared to our method. This indicates that our approach is better suited for dynamic and selective multimodal fusion. Using text as the query vector (**TEXTQUERY**, #13) also caused performance degradation, further demonstrating that visual information plays a dominant role in fusion and emphasizing the importance of optimizing visual feature focus. Overall, the ablation studies thoroughly validate the effectiveness and rationality of our method.

Qualitative Analysis To further verify the effectiveness of the contrastive learning and fusion modules, we conducted t-SNE visualization on the Pheme test set, illustrated in Figure 4. The "-w/o MF" variant shows that the fusion module helps distinguish multimodal rumor features, but there is still overlap between labels. In contrast, CoM-ICA with contrastive learning exhibits clearer boundaries and significantly reduces feature overlap.

5.3 Compare Our Method to Channel Attention Methods

We replace the ICA module with other channel attention mechanisms for comparison experiments. As shown in Table 2, the model performance drops by at least 1.3%. Specifically, **CBAM** (#5) enhances feature representation by combining spatial and channel attention mechanisms; **ECA-Net** (#6) employs local cross-channel interactions to avoid

Model	Feature	English→German		
		Test2016	Test2017	MSCOCO
SA -w/o ICA	ResNet	40.69	32.66	29.10
SA -w ICA	ResNet	**40.83**	**33.24**	**29.32**

Table 3: BLEU scores of our proposed model on Multi30K Test2016, Test2017 and MSCOCO test sets. **SA** denotes Selective Attention model (Li et al., 2022).

the computational complexity of fully connected layers; **GCT** (#7) adjusts feature distribution between channels using a gating mechanism; and **MCA** (#8) characterizes channel distributions with higher-order statistics such as mean, standard deviation, and skewness. In contrast, CoM-ICA models channel correlations based on the covariance matrix and more accurately captures fine-grained variations in channel features. This method not only excels at enhancing informative channel representation and suppressing noise but also offers good interpretability by intuitively reflecting the importance of each channel. It is particularly well suited for fine-grained feature modeling and optimization in multimodal scenarios.

5.4 Apply Our Method to Other Encoders

To comprehensively verify the generalization ability of ICA, we conduct experiments using CoM-ICA with various text and image encoders (see Figure 5). The results show that different encoders affect performance, with RoBERTa as the text encoder and ResNet-50 as the image encoder achieving the best results. Meanwhile, the consistent performance improvements across different encoders demonstrate that by optimizing the channel map, ICA effectively guides visual features to focus on important information. This enhancement does not depend on a specific encoder, indicating good generalizability and application potential.

5.5 Generalization to Multimodal Machine Translation

We further investigate the applicability of ICA in multimodal machine translation (MMT), where visual information is integrated into neural machine translation to enhance language understanding through visual context. As shown in Table 3, we conduct experiments on the Multi30K benchmark (Elliott et al., 2016), with the training and validation sets comprising 29,000 and 1,014 instances, respectively. The results indicate that incorporating ICA into the Selective Attention framework (Li et al., 2022) improves BLEU scores compared with the baseline without ICA. This improvement is attributed to ICA's effectiveness in optimizing the focus of visual features, which is a key challenge in MMT. The demonstrated effectiveness of ICA in MMT shows its ability to filter out image noise at the visual feature level, suggesting potential applicability to other multimodal tasks.

Model	Top-1	Top-5
ResNet-50	0.862	0.993
ECA + ResNet-50	0.877	0.994
ICA + ResNet-50 (Ours)	**0.884**	**0.995**

Table 4: Classification Results on the CIFAR-10 Dataset.

Model	Top-1	Top-5
ResNet-50	0.623	0.858
ECA + ResNet-50	0.643	0.862
ICA + ResNet-50 (Ours)	**0.650**	**0.879**

Table 5: Classification Results on the CIFAR-100 Dataset.

5.6 Generalization to Image Classification

To further evaluate the capability of ICA in pure visual tasks, we extend its application to the Image Classification task. A fundamental challenge in classification lies in effectively distinguishing discriminative object patterns from complex backgrounds or redundant channel information. We conducted experiments on the CIFAR-10 and CIFAR-100 benchmarks (Krizhevsky et al., 2009) using ResNet-50 as the backbone, comparing ICA with the baseline ResNet-50 and the parametric attention-based ECA-Net (Wang et al., 2020).

As summarized in Table 4 and Table 5, the integration of ICA consistently outperforms both the vanilla ResNet-50 and its ECA-enhanced version across all metrics. Specifically, on the CIFAR-10 dataset, ICA + ResNet-50 achieves a Top-1 accuracy of 88.4% and a Top-5 accuracy of 99.5%, surpassing ECA by 0.7% and 0.1%, respectively. A more significant improvement is observed on the more complex CIFAR-100 dataset, where ICA reaches 65.0% (Top-1) and 87.9% (Top-5), representing an increase of 2.7% and 2.1% over the baseline model. These results demonstrate that even in single-modality vision tasks, ICA's variance-based selection mechanism can effectively prioritize informative structural features. This confirms its potential as a robust, parameter-free, and plug-and-

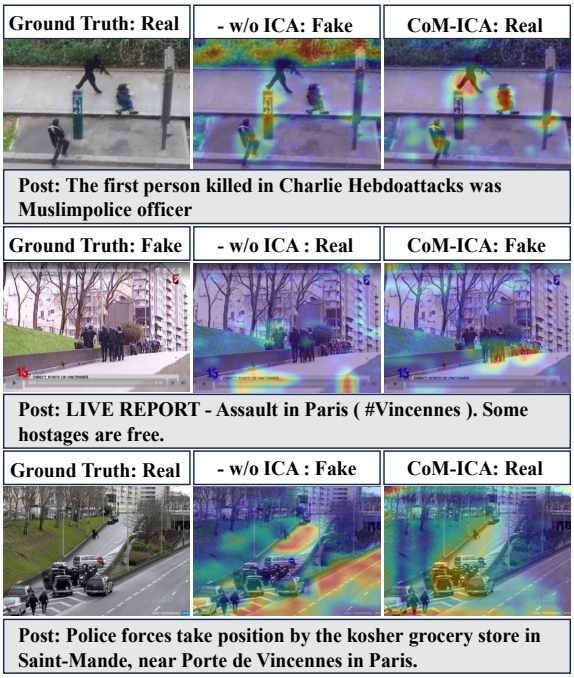

Figure 6: Three detection cases from the MFND task. The left side shows the original images, while the middle and right figures present the visualizations of the visual features used by the model.

play module for a wide range of computer vision applications.

5.7 Case study

Figure 6 presents three detection cases analyzed using two different systems. The "**-w/o ICA**" denotes the baseline model without the ICA module. We employ Grad-CAM (Selvaraju et al., 2017) to generate heatmaps that visualize the attention regions of the visual features. The results indicate that the baseline model tends to focus on background information in the images, such as shrubbery, streets, and text, which may be caused by improper channel map allocation. In contrast, CoM-ICA is more inclined to attend to salient content within the images—for example, "The first person" in the first case, "Some hostages" in the second case, and "Police forces take position" in the third case. These cases demonstrate that the proposed ICA module can effectively guide visual attention through image channel allocation. By refining attention allocation, it enhances the model's ability to perceive key information within the image, thereby further validating the performance improvements brought by the CoM-ICA architecture in the multimodal fake news detection (MFND) task. In addition, more illustrative visual examples are provided in Appendix A to further support our analysis.

6 Conclusion

This paper addresses the issue of image noise caused by inaccurate visual feature focus in the multimodal fake news detection (MFND) task. We propose a covariance matrix-driven image channel allocation module to evaluate channel importance and effectively guide visual features to focus on salient information. Then, a multilayer perceptron transformer is designed to dynamically fuse multimodal features, and finally, contrastive learning is employed to reduce the semantic gap between modalities. Extensive experiments demonstrate that the proposed method can effectively improve the performance of the MFND task.

7 Limitations

Although the proposed CoM-ICA method has achieved considerable results, it has not been deeply integrated with current multimodal large language models. In terms of the collaboration between visual and textual features, our approach mainly focus on the visual modality and the multimodal fusion mechanism. The detailed textual characteristics, such as emotions, stances, and other information are not fully exploited. We identify these directions as key challenges to be addressed in our future research.

Acknowledgements

We thank all the anonymous reviewers for their insightful and valuable comments on this paper. This work is supported by The National Key R&D Program of China under grant 2022YFB3104701.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Tun Lu, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2897–2905. ACM.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 627–638. Association for Computational Linguistics.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807. IEEE Computer Society.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5687–5698. Association for Computational Linguistics.

Yi R. Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen R. McKeown, Mohit Bansal, and Avi Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1683–1698. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5394–5413. Association for Computational Linguistics.

Huawei Han, Jianlei Yang, and Wushour Slamu. 2023. Cascading modular multimodal cross-attention network for rumor detection. In *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*, pages 974–980.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Yangbo Jiang, Zhiwei Jiang, Le Han, Zenan Huang, and Nenggan Zheng. 2024. MCA: moment channel attention networks. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 2579–2588. AAAI Press.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 795–816. ACM.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2915–2921. ACM.

Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. Learning multiple layers of features from tiny images.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2022. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6327–6337. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5809–5821. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zach DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1212–1220. ACM.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 153–162. ACM.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor.*, 19(1):22–36.

Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 726–734. ACM.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *Fifth IEEE International Conference on Multimedia Big Data, BigMM 2019, Singapore, September 11-13, 2019*, pages 39–47. IEEE.

Mengzhu Sun, Xi Zhang, Jianqiang Ma, and Yazheng Liu. 2021. Inconsistency matters: A knowledge-guided dual-inconsistency network for multi-modal rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1412–1423. Association for Computational Linguistics.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. 2020. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11531–11539. Computer Vision Foundation / IEEE.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Zhuang Wang and Jie Sui. 2021. Multilevel attention residual neural network for multimodal online social network rumor detection. *Frontiers in Physics*, 9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: convolutional block attention module. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 3–19. Springer.

Ke Wu, Song Yang, and Kenny Q. Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 651–662. IEEE Computer Society.

Yang Wu, Pengwei Zhan, Yunjian Zhang, LiMing Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2560–2569. Association for Computational Linguistics.

Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):12113–12132.

Zongxin Yang, Linchao Zhu, Yu Wu, and Yi Yang. 2020. Gated channel transformation for visual recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11791–11800. Computer Vision Foundation / IEEE.

Junjie Ye, Junjun Guo, Yan Xiang, Kaiwen Tan, and Zhengtao Yu. 2022. Noise-robust cross-modal interactive learning with text2image mask for multi-modal neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of*

Korea, October 12-17, 2022, pages 5098–5108. International Committee on Computational Linguistics.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 5384–5392. AAAI Press.

Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025. IMOL: Incomplete-modality-tolerant learning for multi-domain fake news video detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30921–30933, Vienna, Austria. Association for Computational Linguistics.

Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zheng-Jun Zha. 2023. Hierarchical semantic enhancement network for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3424–3433, New York, NY, USA. Association for Computing Machinery.

Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zheng-Jun Zha. 2024. Natural language-centered inference network for multi-modal fake news detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 2542–2550. ijcai.org.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II*, volume 12085 of *Lecture Notes in Computer Science*, pages 354–367. Springer.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I*, volume 10539 of *Lecture Notes in Computer Science*, pages 109–123. Springer.

A Additional Cases

Figure 7 presents six additional supplementary cases. The first five cases further validate the effectiveness of the CoM-ICA model in multimodal fake news detection. The sixth case is a failure example that reveals a potential failure mode of ICA. Specifically, ICA hesitates between textual information and key visual entities, causing the visual features to lose focus on critical regions and resulting in an incorrect prediction. This indicates that the complex mixture of text and key visual entities in news images may lead to ICA failure.

B Hyperparameter Selection

Para	Pheme	
	CEM	**CL**
1	0.913	0.897
2	0.906	0.910
3	**0.923**	0.905
4	0.902	**0.923**
5	0.910	0.902

Table 6: Accuracy on the Pheme dataset under different hyperparameters, **CEM** stands for Channel Expansion Multiplier and **CL** denotes Co-attention Layers.

To investigate the impact of different parameters on experimental results, we conducted a grid search on two key parameters, with the results shown in Table 6. First, we examined the channel expansion multiplier, selecting the best parameter from candidates ranging from 1 to 5. Ultimately, the expansion multipliers were set to 3, 3, and 4 for the Pheme, CFND, and Weibo datasets, respectively. Second, we studied the effect of the number of cross-attention layers in the fusion method, also determining the optimal number of layers through a grid search within the range of 1 to 5. As the number of attention layers increased, model performance peaked at a certain point; excessive layers not only significantly increased computational complexity but also failed to effectively enhance multimodal feature fusion. Based on these results, we set the number of Co-Attention layers to 4 for Pheme, and 3 for both CFND and Weibo.

C Computational Efficiency

We analyze the computational cost of CoM-ICA, with results shown in Table 7. All models have the same number of parameters, but CoM-ICA's training time is 410.92 seconds, significantly longer

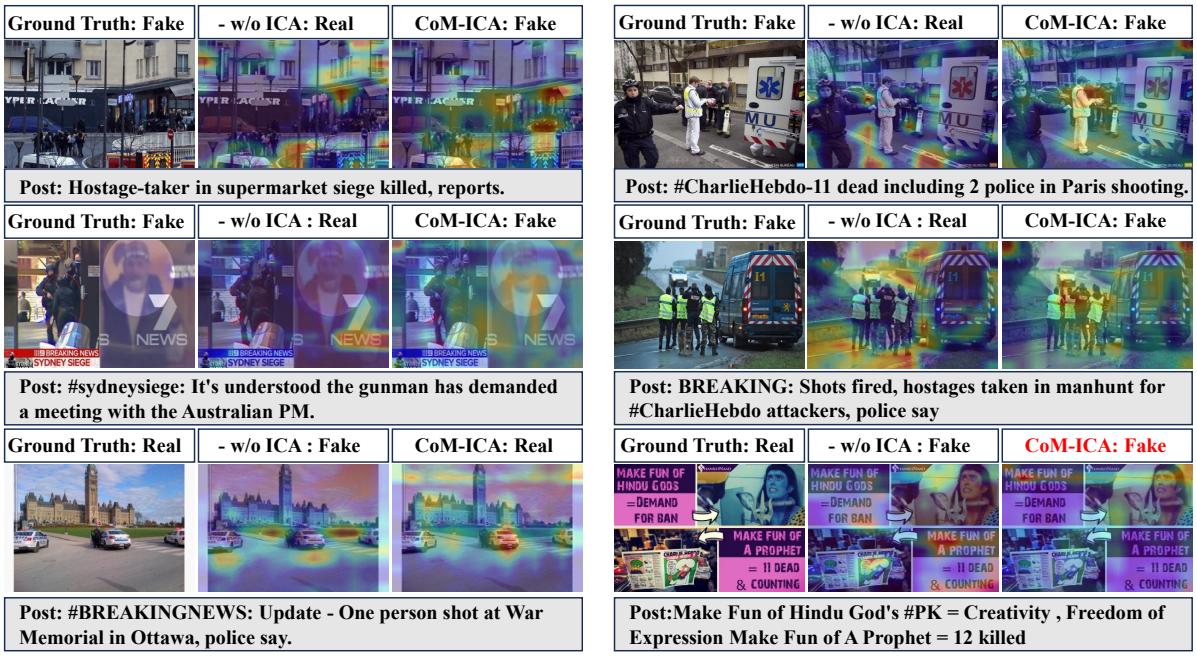

Figure 7: Three additional supplementary cases. The last one marked in red represents a failure case.

Model	Params (M)	Training Time (s)	FLOPs (MMAC)	Latency (ms/sample)	Acc (%)
ICA-Fadar	123.14	410.92	9893.7508	1.3234	**92.3**
-w/o ICA	123.14	303.15	9884.2675	1.2062	89.3
-w/o Channel Expansion	123.14	406.89	9884.2675	1.2287	91.3
-w/o Channel Allocation	123.14	401.01	9893.7508	1.2825	89.7
CBAM (Woo et al., 2018)	123.14	454.49	9889.1848	1.2641	89.9
ECA-Net (Wang et al., 2020)	123.14	397.71	9884.4181	1.2580	90.4
GCT (Yang et al., 2020)	123.14	388.71	9884.2675	1.2222	90.8
MCA (Jiang et al., 2024)	123.14	379.60	9884.2675	1.2696	91.0

Table 7: Computational Efficiency and Performance Comparison of different components and methods.

than 303.15 seconds without the ICA module. This is mainly due to the introduction of channel expansion and enhancement mechanisms, which improve expressive capacity but increase computational complexity.

In terms of computational complexity, CoM-ICA has 9893.75 MMAC FLOPs, only about 0.1% higher than the model without the ICA module, and similarly less than 0.1% increase compared to other models. This minimal overhead brings significant improvements in feature extraction, achieving a good balance between performance and efficiency.

During inference, CoM-ICA's latency is 1.3234 milliseconds, slightly higher than other methods, but it improves accuracy by at least 1.3%. The increase in latency mainly comes from the dynamic channel selection mechanism, yet overall remains within the efficiency requirements of most visual tasks.

In summary, CoM-ICA mainly increases training time and slightly raises inference latency, while significantly boosting accuracy from 89.3% to 92.3%. It outperforms other mainstream attention mechanisms, validating its effectiveness in enhancing the model's discriminative ability and feature perception, with strong practical value.

D Justification of ICA

To further illustrate the rationale for using channel variance as a measure of channel importance in the ICA module, we provide a theoretical explanation based on convolutional mapping and variance analysis below.

After convolutional mapping and channel expansion, let each channel vector be denoted as Y_i . We can regard it as the sum of a "signal" component and a "noise" component,

$$Y_i = S_i + N_i, \tag{21}$$

where S_i represents the signal containing seman-

tic or structural information, and N_i denotes zero-mean random noise. Taking the variance of Y_i and leveraging the additivity of variance (assuming signal and noise are approximately uncorrelated) yields,

$$Var\left(Y_i\right) = Var\left(S_i\right) + Var\left(N_i\right). \tag{22}$$

Through training, the convolutional mapping (implemented via $DWConv_{3\times3}$ followed by 1×1 Conv and nonlinearity $\phi\left(\cdot\right)$) amplifies and aggregates task-relevant spatial patterns in S_i resulting in larger $Var\left(S_i\right)$ for channels carrying useful semantic information. In contrast, random noise typically exhibits relatively small variance and lacks structured correlation across channels. Therefore, for "discriminative" channels, it usually holds that,

$$Var\left(S_i\right) \gg Var\left(N_i\right), \tag{23}$$

which implies,

$$Var\left(Y_i\right) \approx Var\left(S_i\right). \tag{24}$$

For redundant channels, $Var\left(S_i\right)$ is close to zero, leading to a small $Var\left(Y_i\right)$. Normalizing the channel variances $v_i = Var\left(Y_i\right)$ via the Softmax function,

$$\alpha_i = \frac{e^{v_i}}{\sum_j e^{v_j}}, \tag{25}$$

which assigns higher weights to channels with significantly larger $Var\left(S_i\right)$, while the Softmax normalization suppresses the dominance of any single channel with abnormally high variance (if extreme noisy channels exist, their relative weights are still constrained by the overall distribution). In summary, under the assumption that convolutional mappings have learned to denoise and emphasize semantic patterns, using channel variance as a measure of importance combined with Softmax normalization statistically distinguishes channels containing richer semantic information from those dominated by noise, thereby enabling effective channel selection and weighting.

E ICA vs Channel-Attention Methods

To further justify the design of the ICA module, we elucidate its conceptual distinctions from prevalent parametric channel attention mechanisms such as CBAM and ECA-Net. While parametric methods typically treat channel importance as a latent variable to be "predicted" through auxiliary learnable mappings (e.g., MLPs or 1D convolutions) based on training-set priors, our ICA module performs a deterministic feature selection driven by the variance additivity principle. Conceptually, parametric attention focuses on feature recalibration through learned inductive biases, whereas ICA prioritizes feature purification by treating channels as independent signal sources and employing variance as a proxy for information entropy. As derived in Eq. (21)-(24), task-relevant semantic signals S_i inherently exhibit structured variations and higher variance compared to stochastic, low-variance noise components N_i. By utilizing a parameter-free variance-based scoring instead of fixed learned weights, the ICA module adaptively suppresses high-activation noise that often misleads parametric models. This strategic shift from "prior-based prediction" to "evidence-based selection" not only enhances the model's robustness against domain-specific noise but also ensures superior instance-level adaptation and generalization across diverse datasets like Weibo and Pheme without introducing additional computational parameters.

F Off-Diagonal Covariance Analysis

Samples with sparse off-diagonal correlations account for less than 1% of the dataset. This indicates that most samples exhibit dense off-diagonal correlation structures. Therefore, the following analysis mainly focuses on dense-correlation scenarios.

Based on this observation, we remove samples with sparse correlations, retain samples with dense correlations, and conduct ablation experiments to compare different channel importance estimation strategies. We consider two representative metrics that explicitly utilize off-diagonal covariance information.

Row Absolute Sum Variance (RAS Variance) measures the overall correlation strength of each channel by summing the absolute values of off-diagonal entries row-wise:

$$s_i = Var\left(Y_i\right) + \sum_{j \neq i} \left|Cov\left(Y_i, Y_j\right)\right|, \tag{26}$$

where $Var(Y_i)$ denotes the variance of channel i, and $Cov(Y_i, Y_j)$ denotes the covariance between channels i and j.

Top-k Off-Diagonal Augmented Variance (Top-k OAV) augments channel variance with the Top-k most relevant off-diagonal terms:

$$s_i = Var\left(Y_i\right) + \sum_{j \in Top_k(|Cov(Y_i,Y_j)|)} \left|Cov\left(Y_i, Y_j\right)\right|, \tag{27}$$

where $Top_k(\cdot)$ returns the indices of the Top-k largest absolute covariance values associated with channel i.

When the covariance matrix exhibits dense correlations, off-diagonal elements may provide supplementary information; however, they may also introduce noise. As shown in Table 8, directly utilizing all off-diagonal information (RAS Variance) increases the amount of information but does not improve performance. This suggests that noisy correlations may offset the potential benefits. Selecting only the Top-k most relevant off-diagonal terms (Top-k OAV) achieves moderate improvement under dense-correlation scenarios, but still underperforms CoM-ICA.

Notably, ICA relies solely on the diagonal elements of the covariance matrix as importance scores. Under convolutional mappings, semantic signals are amplified, enabling channel variance to effectively distinguish informative channels from noisy channels. Meanwhile, the softmax normalization further suppresses excessive dominance of any single channel. Therefore, relying only on diagonal terms provides a stable and efficient way to estimate channel importance while avoiding the noise introduced by off-diagonal correlations.

Model	Accuracy	Precision	Recall	F1-score
CoM-ICA	**0.907**	**0.894**	**0.872**	**0.882**
w/o ICA	0.890	0.872	0.851	0.860
Top-k OAV	0.896	0.875	0.864	0.869
RAS Variance	0.890	0.877	0.845	0.858

Table 8: Performance comparison under dense off-diagonal correlation conditions on the Pheme dataset.

G Model Robustness Under Distribution Drift

We evaluate the stability of CoM-ICA under distribution drift scenarios. Specifically, Gaussian noise is added to input images to simulate distribution shifts, and the performances of CoM-ICA and the variant without ICA are tested under different noise standard deviations.

As shown in Table 9, the model equipped with ICA demonstrates higher robustness across all noise levels and consistently outperforms the variant without ICA. As the noise intensity increases, the performances of both models decline. However, the model with ICA exhibits a smaller performance drop, highlighting its superior resistance to noise perturbations.

These results indicate that ICA helps suppress noise-sensitive channels while emphasizing more stable semantic channels, thereby improving robustness under distribution shift conditions.

Perturbation Setting	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.2$
Accuracy (ICA)	0.917	0.910	0.899
Accuracy (w/o ICA)	0.915	0.904	0.878

Table 9: Performance on the Pheme dataset under distribution shift simulated by Gaussian noise, where σ denotes the standard deviation of the Gaussian noise.