Multi-Scale Spectral Selection and Entropy-Guided Uncertainty Fusion for Multimodal Rumor Detection

Zongliang Han[1,2], Wenyu Guo[1], Guoqing Jin[1]*, Yang Liu[2]*,
Fan Li[2], Dong Yu[1], Yan Song[3], Fengzhen Zhang[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, Zhangfengzhen}@people.cn, clksong@gmail.com
liuyang22@ucas.ac.cn, {hanzongliang22, lifan221}@mails.ucas.ac.cn

Abstract

Multimodal content combining textual and visual information poses significant challenges for rumor detection on social media. Compared to traditional spatial domain features, frequency domain features have attracted increasing attention due to their stronger discriminative capabilities. However, existing methods still fall short in capturing cross-modal semantic inconsistencies and often overlook inherent noise in multimodal features, which limits overall detection performance. To address these issues, we propose a novel multimodal rumor detection method based on multi-scale spectral selection and entropy-guided uncertainty fusion. Specifically, we first apply the Discrete Cosine Transform (DCT) to image and text features to convert them into the frequency domain. Then, multi-scale convolutional filters are employed to extract fine-grained information across different frequency scales. Next, modality separation is performed to capture both shared and modality-specific features, enabling more effective cross-modal representation learning. Finally, entropy is used to estimate the uncertainty of each prediction branch, calculate confidence scores, and perform adaptive weighted fusion accordingly. Experimental results on multiple benchmark datasets demonstrate that our method outperforms existing state-of-the-art approaches in multimodal rumor detection, demonstrating stronger detection capability and robustness.

1 Introduction

With the rapid development of social media, online platforms such as Twitter and Weibo have become vital sources of information for the public, and a large amount of content combining images and text has emerged. Along with the spread of multimodal media, a more complex and concerning issue has also arisen: multimodal rumors. Multimodal rumors refer to the dissemination of misleading or false information through social media platforms, which integrates various forms of communication, such as text and images. Compared to single-text rumors, multimodal rumors are more intuitive and can elicit stronger emotional responses from users (Shu et al., 2017; Jin et al., 2017b), making them more likely to spread widely. If left unchecked, they may pose a severe challenge to the credibility of news media platforms (Abdelnabi et al., 2022), potentially leading to public panic and social instability. Therefore, detecting and mitigating multimodal rumors is critically important.

Figure 1: Two examples illustrating the issues of feature noise (a) and cross-modal semantic inconsistency (b) in rumor detection.

Intra-modal Noise Problem The noise issue in rumor detection mainly stems from the large amount of irrelevant content mixed in tweets. For example, as shown in Figure 1(a), the text contains non-essential elements such as emojis and links, while the images include background elements or non-key visual regions unrelated to the event. These noisy elements are often encoded by the model during feature extraction, which interferes with semantic alignment between modalities and may cause the model to focus on incorrect information areas, thereby weakening the overall discriminative ability. Therefore, effectively suppressing redundant and irrelevant features has become one of the key challenges in enhancing the

* Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 1090–1100

performance of multimodal rumor detection.

Cross-modal Semantic Inconsistency Semantic differences between modalities are primarily reflected in both their content and presentation styles. As shown in example (b) of Figure 1, the word "surf" in the text is semantically associated with the visual depiction of waves in the image. However, the specific expressions and emotional vocabulary in the text are not directly represented in the image. Meanwhile, elements such as cars appear in the image but are not mentioned in the text. This indicates that although different modalities may share some similar semantic information, each modality still retains its own unique content and form of expression. Addressing such inconsistency is also of great significance for detection tasks.

To address the above issues, we propose a multimodal rumor detection method based on Multi-Scale Spectral Selection and Entropy-Guided Uncertainty Fusion. This method extracts key information from different frequency bands in a fine-grained manner through multi-scale spectral selection. Subsequently, it leverages modality decomposition to obtain both shared features and modality-specific features, and estimates the uncertainty of each prediction branch using entropy to adjust the contribution of private and shared information to the classification results. The main contributions of this work are as follows:

- A multi-scale spectral selection method is proposed to capture key features across different frequency bands, aiming to address the issue of feature noise.
- We propose a frequency feature separation method that extracts shared and private features through modality decomposition, combined with an entropy-based uncertainty fusion that performs dynamic weighted fusion across branches, thereby enabling more effective capture of cross-modal inconsistencies.
- Extensive experiments are conducted on three publicly available datasets, and the results validate the effectiveness of the proposed method.

2 Related Work

2.1 Multimodal Rumor Detection

Previous multimodal rumor detection methods typically fuse image and text features by simply concatenating them (Jin et al., 2017a; Wang et al., 2018; Cui et al., 2019), which integrates multimodal information only in the spatial dimension, neglecting the deeper interactions between different modalities. To learn shared representations of multimodal information, MVAE (Khattar et al., 2019) proposes a multimodal variational autoencoder that reconstructs multimodal representations from a learned probabilistic latent model. CAFE (Chen et al., 2022) further introduces cross-modal alignment and ambiguity learning mechanisms, enhancing multimodal feature fusion and enriching modal representations by leveraging contextual information from hidden states. BMR (Ying et al., 2023) proposes an improved Multi-gate Mixture-of-Expert network, which combines single-view prediction and cross-modal consistency learning strategies to jointly model both unimodal and multimodal features, further improving the model's generalization ability and robustness.

2.2 Multimodal Spectrum Rumor Detection

The Discrete Cosine Transform (DCT) effectively transforms signals from the time or spatial domain into the frequency domain and has been widely applied in deep learning(Liu et al., 2021; Yu et al., 2025; Shen et al., 2021). In multimodal rumor detection tasks, frequency-domain features are often used as supplementary clues to provide additional evidence. For example, MCAN(Wu et al., 2021) treats frequency-domain information as physical characteristics of images to identify recompression artifacts, achieving certain effectiveness. However, such methods still primarily rely on spatial-domain features for rumor detection. In contrast, FSRU(Lao et al., 2024) pioneers a multimodal spectral rumor detection framework that transforms original features into the frequency domain and performs a series of complex-valued operations within this domain for detection. Nonetheless, existing methods often overlook the key attributes of rumor features—namely, the inter-modal disparities and noise caused by redundant information. To address this issue, we propose a multi-scale spectral selection approach for fine-grained filtering of noise information, and design an entropy-driven uncertainty fusion mechanism to effectively capture cross-modal differences, thereby significantly improving rumor detection performance.

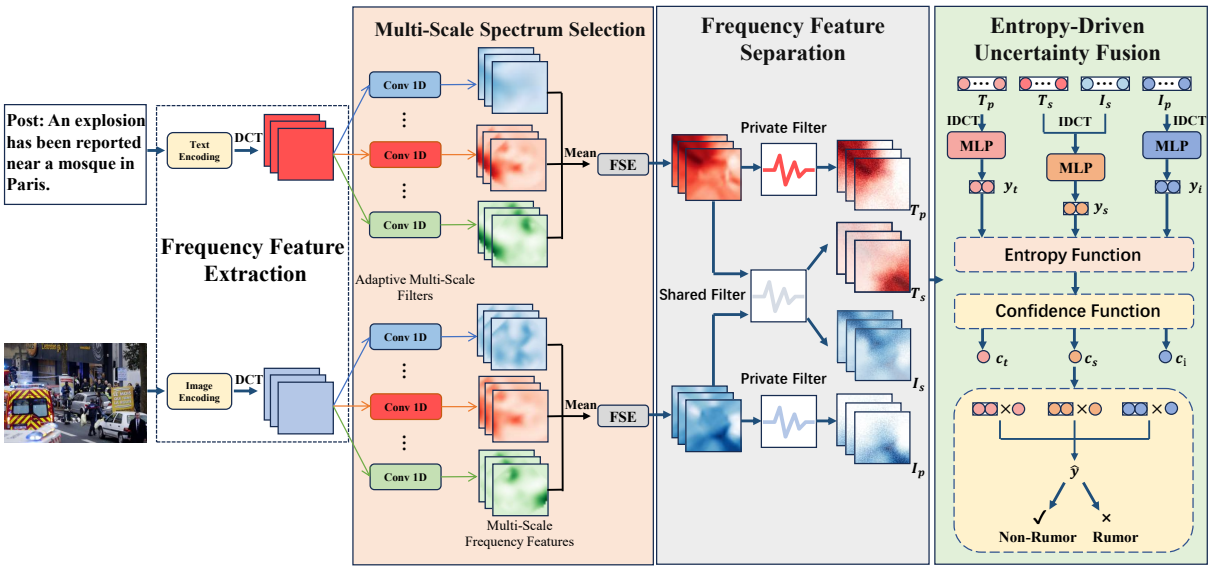

Figure 2: The overall architecture of **MS-SFM**. Where **FSE** stands for the Frequency Squeeze-and-Excitation module; $\{\mathbf{T}_p, \mathbf{T}_s, \mathbf{I}_p, \mathbf{I}_s\}$ represent the private and shared features of the text and image modalities, respectively; $\{\mathbf{y}_t, \mathbf{y}_s, \mathbf{y}_i\}$ denote the three prediction results, and $\{c_t, c_s, c_i\}$ correspond to the confidence scores of these predictions.

3 Methodology

Figure 2 illustrates the overall architecture of MS-SFM, which consists of four key components. First, the frequency-domain feature extraction module transforms spatial domain features into their frequency-domain representations. The multi-scale spectral selection module explores and filters information across different frequency ranges to extract meaningful signals. Next, the frequency-domain feature separation strategy is applied to capture both shared and private information from image and text modalities. Finally, the entropy function is used to compute the confidence of each information branch, and a weighted fusion is performed to produce the final prediction.

3.1 Frequency Feature Extraction

Given an input image, intermediate feature representations are first extracted using a ResNet-50(He et al., 2016) network, denoted as $\mathbf{I} \in \mathbb{R}^{B\times D}$, where B represents the batch size and D is the feature dimension. Similarly, for the text modality, the input text is first encoded by a pre-trained BERT(Devlin et al., 2019) encoder, resulting in the contextual representation tensor $\mathbf{T} \in \mathbb{R}^{B\times L\times D}$, where L is the sequence length. For the image, a one-dimensional Discrete Cosine Transform (DCT)(Ahmed et al., 1974) is applied along the feature dimension to obtain the frequency-domain representation $\mathbf{I}_f \in \mathbb{R}^{B\times D}$. For the text, a one-dimensional DCT is applied to each token feature vector, resulting in the frequency-domain representation $\mathbf{T}_f \in \mathbb{R}^{B\times L\times D}$.

$$\mathbf{I}_f^{(b)}[x] = \alpha_x \sum_{n=0}^{D-1} \mathbf{I}^{(b)}[n] \cdot \cos\left[\frac{\pi}{D}\left(n+\frac{1}{2}\right)x\right], \tag{1}$$

$$\mathbf{T}_f^{(b,l)}[x] = \alpha_x \sum_{n=0}^{D-1} \mathbf{T}^{(b,l)}[n] \cdot \cos\left[\frac{\pi}{D}\left(n+\frac{1}{2}\right)x\right], \tag{2}$$

where α_x is the normalization coefficient.

3.2 Multi-Scale Spectral Selection(MSS)

In many rumors, irrelevant information such as extraneous emojis, special symbols, and links in texts, as well as unrelated visual elements and backgrounds in images, often introduces noise during the detection process, interfering with the model's accurate capture of key information. Given the strong separability of spectral features, we propose a multi-scale spectral selection strategy to perform fine-grained filtering on spectral features, thereby effectively suppressing noisy content while extracting key information. This strategy consists of two sub-modules: one is the adaptive multi-scale filter, which is used to extract fine-grained information from different frequency bands; the other is the frequency squeeze-and-excitation mechanism, which

is used to dynamically select the most salient features from each frequency band.

Adaptive Multi-Scale Filters(AMF) For the input frequency-domain features $\mathbf{F}_x = \mathbf{T}_f \in \mathbb{R}^{B\times L\times D}$ or $\mathbf{F}_x = \mathbf{I}_f \in \mathbb{R}^{B\times D}$, We introduce multi-scale convolutional filters along the feature dimension to model contextual information across different frequency bands. The convolutional output at the s-th scale is denoted as:

$$\mathbf{F}^{(s)} = \text{Conv1D}^{(s)}\left(\mathbf{F}_x\right), \tag{3}$$

the kernel size of each convolution is k. For different scales, the dilation rate is set correspondingly as $d_s = s$, and symmetric padding is used to maintain the feature dimension. Different dilation rates correspond to receptive fields of varying sizes, enabling the capture of fine-grained information from different frequency bands.

Symmetric padding is applied to maintain the feature dimension, and the results from multiple scale convolutions are concatenated and then averaged for fusion:

$$\mathbf{F}_{\text{ms}} = \frac{1}{S}\sum_{s=1}^{S}\mathbf{F}^{(s)}, \tag{4}$$

where S is the total number of scales, and the dimension of $\mathbf{F}_{\text{ms}}$ is consistent with that of $\mathbf{F}_x$.

Frequency Squeeze-and-Excitation(FSE) After obtaining the multi-scale spectral features, inspired by the SE-Net(Hu et al., 2018) architecture, we design a frequency squeeze-and-excitation mechanism to assign different weights to each frequency dimension, thereby enabling fine-grained and adaptive selection of spectral features:

$$\mathbf{F}^{'} = \mathbf{F}_{\text{ms}} \odot \sigma\left(\mathbf{W}\mathbf{F}_{\text{ms}}\right), \tag{5}$$

here, $\mathbf{W} \in \mathbb{R}^{D\times D}$ is a learnable linear transformation matrix, $\sigma(\cdot)$ denotes the sigmoid activation function, $\odot$ represents element-wise multiplication, and $\mathbf{F}' \in \mathbb{R}^{B\times D}$ or $\mathbb{R}^{B\times L\times D}$ is the weighted frequency-domain feature.

3.3 Frequency Feature Separation and Fusion

Compared to features in the spatial domain that may have ambiguous boundaries, frequency-domain representations exhibit clearer energy distributions and stronger structural separability(Xu et al., 2023; Yang et al., 2024; Xu et al., 2024). Based on this, we explore a similarity-based frequency-domain feature separation strategy. By jointly modeling shared and private features, this approach effectively captures the differences between modalities.

Frequency-Domain Feature Separation Specifically, in order to eliminate the scale differences between different modalities and ensure the stability of the decoupling process, we first perform L2 normalization on the image feature $\mathbf{F}_{\text{I}}$ and text feature $\mathbf{F}_{\text{T}}$. This process can be formalized as follows:

$$\hat{\mathbf{F}}_{\text{I}} = \frac{\mathbf{F}_{\text{I}}}{\|\mathbf{F}_{\text{I}}\|_2 + \epsilon}, \tag{6}$$

$$\hat{\mathbf{F}}_{\text{T}} = \frac{\mathbf{F}_{\text{T}}}{\|\mathbf{F}_{\text{T}}\|_2 + \epsilon}, \tag{7}$$

the L2 norms of the image and text features are denoted as $\|\mathbf{F}_{\text{I}}\|_2$ and $\|\mathbf{F}_{\text{T}}\|_2$, respectively, and ϵ is a small constant to prevent division by zero.

Next, the mean is taken along the sequence dimension of the text features, and the similarity between image and text features is calculated as a trigger for modality conflict.

$$\hat{\mathbf{F}}'_{\text{T}} = \frac{1}{L}\sum_{l=1}^{L}\hat{\mathbf{F}}_{\text{T}}\left[:, l, :\right], \tag{8}$$

$$trigge = \alpha = \hat{\mathbf{F}}_{\text{I}} \cdot \hat{\mathbf{F}}'_{\text{T}}, \tag{9}$$

where α represents the similarity between image and text features, and $\cdot$ denotes the dot product operation. The shared filter is generated based on the similarity between image and text features, while the private filters are activated when the trigger is engaged, performing a fine-grained separation by comparing the amplitude differences in the frequency domain.

$$\text{filter}_s = \Pi\left(\alpha > \theta\right), \tag{10}$$

$$\text{filter}_{pi} = \Pi\left(\hat{\mathbf{F}}_{\text{I}} > \hat{\mathbf{F}}'_{\text{T}} \wedge \alpha < \theta\right), \tag{11}$$

$$\text{filter}_{pt} = \Pi\left(\hat{\mathbf{F}}'_{\text{T}} > \hat{\mathbf{F}}_{\text{I}} \wedge \alpha < \theta\right), \tag{12}$$

where $\Pi(\cdot)$ denotes the indicator function, which equals 1 if the condition holds, and 0 otherwise; θ is a threshold.

Through the above filters, the features are decoupled, modeling the shared and private information of the text and image features separately, and then transformed back to spatial and sequential features through Inverse Discrete Cosine Transform(IDCT).

$$\mathbf{I}_p = IDCT\left(\hat{\mathbf{F}}_{\mathrm{I}} \odot \text{filter}_{pi}\right), \quad (13)$$

$$\mathbf{T}_p = IDCT\left(\hat{\mathbf{F}}'_{\mathrm{T}} \odot \text{filter}_{pt}\right), \quad (14)$$

$$\mathbf{I}_s = IDCT\left(\hat{\mathbf{F}}_{\mathrm{I}} \odot \text{filter}_{s}\right), \quad (15)$$

$$\mathbf{T}_s = IDCT\left(\hat{\mathbf{F}}'_{\mathrm{T}} \odot \text{filter}_{s}\right). \quad (16)$$

Entropy-Driven Uncertainty Fusion The contributions of the decoupled shared and private features to the final prediction vary across different scenarios, meaning that the model's confidence in each type of feature is inconsistent. To capture this variability, we measure the prediction uncertainty of each feature type using the entropy of their individual outputs. Based on these uncertainties, the model adaptively assigns fusion weights, thereby dynamically adjusting the influence of each feature type on the final decision.

Specifically, for the three branches of private image features, private text features, and shared features, we calculate the prediction results through independent multi-layer perceptrons (MLPs):

$$\mathbf{y}_s = MLP_s\left([\mathbf{I}_s; \mathbf{T}_s]\right), \quad (17)$$

$$\mathbf{y}_t = MLP_t\left(\mathbf{T}_p\right), \quad (18)$$

$$\mathbf{y}_i = MLP_i\left(\mathbf{I}_p\right), \quad (19)$$

$$\mathbf{p}_k = \text{softmax}\left(\frac{\mathbf{y}_k}{\tau}\right), \quad k \in \{s, t, i\}, \quad (20)$$

where $\mathbf{p}_k \in \mathbb{R}^2$ represents the probability prediction results, and $k \in \{s, t, i\}$ denotes the concatenated shared features, private text features, and private image features, respectively. The parameter τ is the temperature coefficient used to control the smoothness of the softmax output.

Subsequently, we quantify the uncertainty of the predictions by calculating their entropy,

$$H\left(\mathbf{p}_k\right) = -\sum_{j=1}^{2} p_k^j \log\left(p_k^j + \varepsilon\right), \quad (21)$$

where p_k^j represents the predicted probability for class j, ε is a small constant to avoid numerical instability, and $H(\mathbf{p}_k)$ is the prediction entropy for branch k, measuring the average uncertainty of the prediction. Lower entropy indicates higher confidence.

We compute the confidence of each branch based on its entropy, then apply softmax to get the fusion weights, and finally obtain the weighted prediction result:

$$c_k = 1 - \frac{H(\mathbf{p}_k)}{\log(2)}, \quad (22)$$

$$\boldsymbol{\omega} = \text{softmax}\left([c_s, c_t, c_i]\right), \quad (23)$$

$$\hat{\mathbf{y}} = \sum_{k \in \{s,t,i\}} \omega_k \cdot \mathbf{p}_k, \quad (24)$$

where c_k denotes the confidence score for each branch, $\boldsymbol{\omega}$ are the fusion weights, and $\hat{\mathbf{y}}$ is the final prediction result.

3.4 Training Objects

We use the cross-entropy loss as our training objective:

$$\mathcal{L}_p = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)\right), \quad (25)$$

where N is the number of training samples, $y_i \in \{0, 1\}$ is the ground truth label for sample i, and $\hat{y}_i$ is the predicted probability of the positive class for the i-th sample.

4 Experiments

4.1 Experimental Settings

We evaluate the performance of the MS-SFM model on three widely used rumor detection benchmark datasets: the Weibo dataset (Jin et al., 2017a), the CFND dataset (Zhang et al., 2024), and the Pheme dataset (Zubiaga et al., 2017). The Weibo dataset includes real news from authoritative Chinese media such as Xinhua News Agency, as well as fake news verified by the official fact-checking system of the Weibo platform. This dataset contains a total of 9,528 news instances. The CFND dataset is compiled from multiple Chinese fact-checking websites and official news sources, covering five different domains, and includes a total of 26,665 news instances. The Pheme dataset consists of tweets from Twitter, focusing on five major breaking news events. We follow the data processing strategy used in NLIN (Zhang et al., 2024) in our experiments.

We implement the MS-SFM model using PyTorch (Paszke et al., 2017) and optimize the loss function with the Adam optimizer. The input image size is set to 224×224. The maximum sequence length for textual inputs is set to 32 for the Pheme dataset and 64 for both the Weibo and CFND datasets. The embedding dimension for

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
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
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
FSRU(Lao et al., 2024)	0.830	0.815	0.758	0.776	0.879	0.872	0.876	0.874	0.901	0.901	0.903	0.902
Event-Radar(Ma et al., 2024)	0.901	**0.883**	0.878	0.880	-	-	-	-	0.919	**0.928**	0.910	0.919
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
MS-SFM	**0.910**	0.882	**0.884**	**0.883**	**0.915**	**0.915**	**0.906**	**0.910**	**0.924**	0.925	**0.924**	**0.924**

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.

Category	Ablation Settings	Pheme			
		Accuracy	Precision	Recall	F1-Score
Full Model	MS-SFM	**0.910**	**0.882**	0.884	**0.883**
Intra-modal Disparity	- w/o AMF	0.878	0.833	0.867	0.847
	- w/o FSE	0.875	0.831	0.857	0.843
	- w/o MSS	0.893	0.860	0.858	0.859
Inter-modal Disparity	- w/o EUF	0.884	0.843	0.863	0.852
	- w/o FFS	0.880	0.840	0.849	0.844
	- w/o trigge	0.899	0.869	0.865	0.867
	- w/o amplitude comparison	0.893	0.856	0.869	0.862
	Use Max	0.814	0.762	0.721	0.736
	Use Mean	0.891	0.848	**0.890**	0.865

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

Ablation Settings	Pheme			
	Accuracy	Precision	Recall	F1-Score
MS-SFM	**0.910**	**0.882**	**0.884**	**0.883**
- w/o Private Image	0.882	0.841	0.859	0.849
- w/o Private Text	0.853	0.805	0.831	0.816
- w/o Shared Feature	0.873	0.833	0.836	0.834

Table 3: Ablation study on the contribution of private and shared features.

both text and image features is 768. The hyperparameter θ is set to 0.5 across all three datasets, the number of scales S is set to 3, and the temperature coefficient τ is set to 0.7. The model is trained for 30 epochs on each dataset with a learning rate of 1e-3 and a batch size of 30. All experiments are conducted on a single Tesla A100 GPU.

4.2 Main Results

We compare the MS-SFM model with nine representative baseline methods on three multimodal rumor detection benchmark datasets. The detailed experimental results are presented in Table 1, and the analysis is as follows:

Firstly, across all three datasets, our proposed MS-SFM method consistently outperforms other comparison methods on nearly all evaluation metrics, fully demonstrating the effectiveness of MS-SFM in enhancing the performance of multimodal rumor detection tasks.

Secondly, among the baseline models, MVAE and SAFE performed relatively poorly, which may be attributed to their use of Text-CNN and Bi-LSTM, resulting in suboptimal text representations. In contrast, methods such as NLIN and Event-Radar utilize the CLIP text encoder, enabling them to extract more effective text representations and achieve better performance. This highlights the crucial role of text representation in multimodal rumor detection tasks.

Finally, MS-SFM consistently outperforms all baseline methods across all evaluation datasets, demonstrating a clear performance advantage, especially when compared to FSRU, which was the first to introduce a multimodal frequency-domain framework into rumor detection. This improvement indicates that simply transforming modal data into the frequency domain and processing it there is insufficient to fully unleash the potential of frequency-domain features. In contrast, MS-SFM effectively alleviates the problem of feature noise by conducting fine-grained extraction of key information, thereby obtaining more efficient frequency-domain representations. Moreover, the model explicitly models cross-modal heterogeneity, enabling a deeper understanding and integration of multimodal information, which significantly enhances its discriminative ability in rumor detection tasks.

Fusion Methods	Pheme			
	Accuracy	Precision	Recall	F1-Score
Uncertainty Fusion	**0.910**	**0.882**	0.884	**0.883**
Late Fusion(Average)	0.889	0.860	**0.893**	0.874
Late Fusion(Max)	0.878	0.845	0.825	0.834
Learned Attention	0.886	0.852	0.848	0.850

Table 4: Performance of different fusion methods on the Pheme dataset.

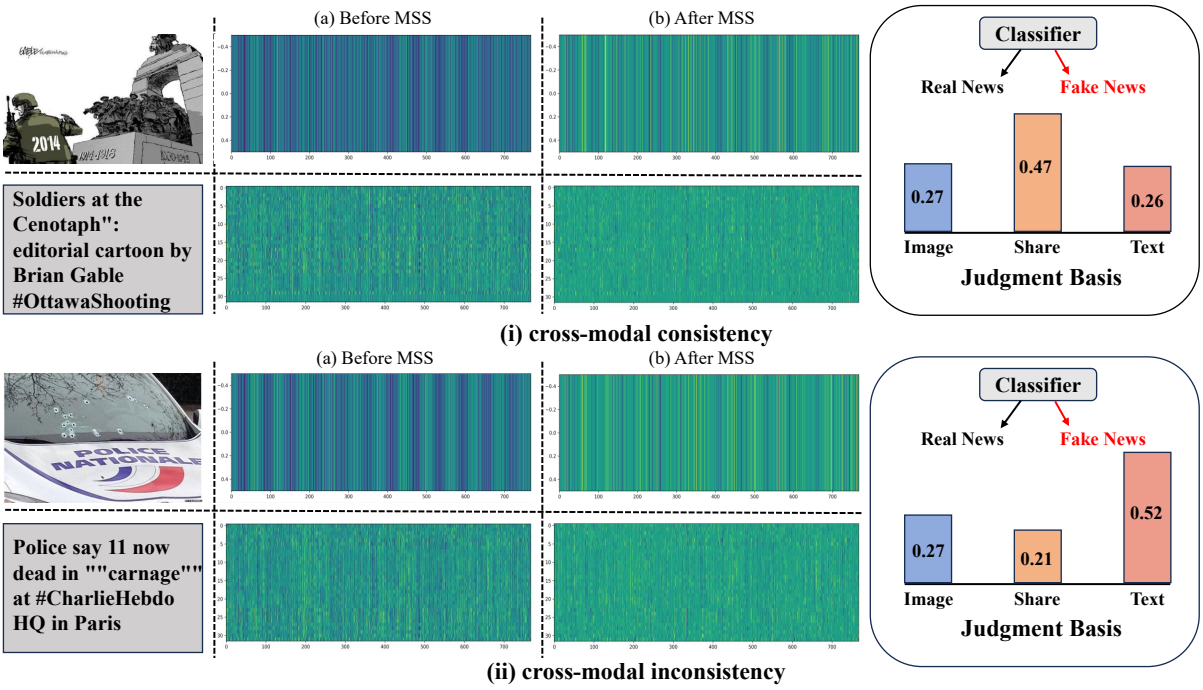

Figure 3: Two visualization examples of the energy distribution of the frequency features and the uncertainty fusion weight during the classification process, including both matched and mismatched cases of images and texts. The dispersion of the energy distribution reflects the noise level in the frequency features, while the fusion weight is the basis for classification used by the classifier after being adjusted by our method.

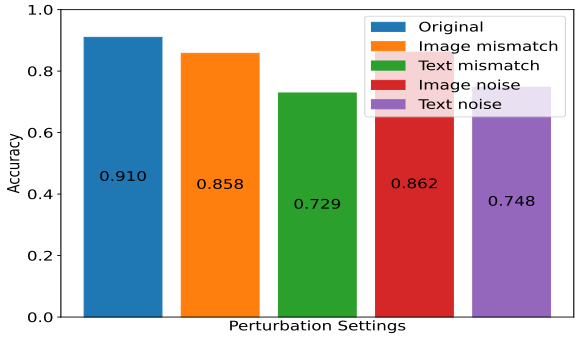

Figure 4: Comparison of model performance under different perturbation settings on the Pheme dataset. Here, Image mismatch and Text mismatch denote randomly shuffling images or texts to simulate cross-modal inconsistency scenarios, while Image noise and Text noise represent the injection of Gaussian noise into image and text features, respectively, to simulate misleading frequency-domain characteristics.

5 Analysis

5.1 Effect of MSS

To investigate the effect of intra-modal discrepancies, we conducted ablation studies, with results shown in Table 2.

When removing the Adaptive Multi-scale Filter ($-w/o$ AMF), the performance dropped by about 3.2%, indicating that multi-scale modeling can more effectively capture fine-grained frequency-domain features, thereby improving rumor detection performance.

When removing the Frequency Squeeze-and-Excitation module ($-w/o$ FSE), the performance significantly decreased, highlighting the importance of this frequency-domain feature selection mechanism, which assigns higher weights to important frequency features during training and thus improves detection accuracy.

When the Multi-scale Spectrum Selection module ($-w/o$ MSS) was removed, the accuracy dropped by 1.7%, indicating that effective feature selection across multiple frequency scales is crucial for capturing key information. This further validates the crucial role of mitigating noise issues in improving detection performance.

5.2 Effect of Uncertainty Fusion

We remove the entropy-driven uncertainty fusion module ($-w/o$ EUF) and instead concatenate modal features directly, followed by a multi-layer perceptron for prediction, resulting in a 2.6% performance drop. This indicates that shared and private information across modalities contribute differently. Replacing weighted fusion with average fu-

sion and max fusion causes performance decreases of 1.9% and 9.6%, respectively, demonstrating that our strategy effectively leverages information differences to assign weights, enhancing decision fusion and task accuracy.

Furthermore, removing the frequency-domain feature separation mechanism ($-w/o$ FFS) and applying a unified weighting strategy to fuse text and image features leads to a 3.0% performance drop, though still better than the case without EUF. This shows that text and image modalities contribute differently to the task. Our method models fine-grained feature differences to achieve more effective feature fusion, improving overall performance. Additionally, removing the trigger operation and frequency amplitude comparison also reduces performance, further validating the effectiveness of the frequency-domain separation strategy.

5.3 Feature Contribution Analysis

To analyze the impact of private and shared features on model performance, we conduct ablation experiments, with results shown in Table 3. Removing private image features, private text features, or shared features all lead to performance declines, demonstrating their importance for rumor detection. Notably, removing private text or shared features causes more significant drops, indicating their crucial roles in capturing modality-specific and cross-modal information. In contrast, removing private image features resulted in a smaller decrease, suggesting a relatively lower contribution. These results show that private and shared features carry different levels of information, and their collaborative integration is a key factor in enhancing model performance.

5.4 Comparison of Different Fusion Methods

To further validate the effectiveness of the proposed uncertainty fusion mechanism, we conduct comparative experiments, including late fusion (max and average fusion) and learnable attention mechanisms, all without entropy calculation or uncertainty evaluation. As shown in Table 4, max fusion performed the worst, as it only selects the prediction branch with the highest probability, resulting in information loss. Although average fusion equally considers information from all branches, it ignores the differences in confidence and reliability across modalities, thereby limiting the improvement of the fusion performance. In contrast, the uncertainty fusion mechanism adaptively adjusts modality weights based on prediction entropy, more accurately capturing inter-modal discrepancies and effectively suppressing the influence of low-confidence modalities on the final decision, thus achieving optimal performance.

5.5 Robustness Analysis

To evaluate the impact of cross-modal inconsistency and misleading frequency features, we conduct perturbation experiments. As shown in Figure 4, substituting irrelevant images yields 85.8% accuracy, indicating robustness to visual perturbations, whereas replacing irrelevant text causes a sharp decline, underscoring the dominant role of textual information. Injecting noise into image features achieves 86.2% accuracy, slightly higher than image replacement, suggesting that frequency-domain noise can be effectively suppressed. By contrast, noise injection into text features reduces accuracy to 74.8%, reflecting the higher vulnerability of textual representations, where local perturbations can disproportionately disrupt semantic integrity.

5.6 Case Study

To more intuitively demonstrate the evolution and fusion process of frequency representations in MS-SFM, we visualize the energy distribution of the frequency-domain features before and after multi-scale spectral selection (MSS), as well as the confidence weights of each information branch in the entropy-driven uncertainty fusion mechanism (see Figure 3). The results show that before selection, the energy distribution of frequency-domain features is dense, with some potential noise present across the frequency spectrum. After the MSS, the energy distribution becomes sparser, indicating that the model can effectively filter out noise while retaining key frequency bands. Additionally, when modalities are consistent, the model increases the weight of shared features; when they are inconsistent, it tends to rely more on the text modality—aligning with the common understanding that text features are generally the dominant modality. This also indicates that a fixed fusion strategy is not applicable, as a certain branch might be overly suppressed, leading to misclassification. This further highlights the crucial role of the uncertainty fusion mechanism.

6 Conclusion

MS-SFM addresses challenges in multimodal frequency-domain rumor detection, including intra-

modal noise and cross-modal content inconsistency. First, a multi-scale spectral selection module filters noise and extracts key frequency features. Second, a modality separation mechanism distinguishes shared and modality-specific image and text features. Third, an entropy-based uncertainty fusion strategy dynamically integrates information across modalities. Experimental results demonstrate that the proposed method effectively improves rumor detection performance.

7 Limitations

Experimental results indicate that weighting different information branches based on feature separation helps capture inter-modal differences. However, the current feature separation strategy has a strong dependence on the hyperparameter θ, which relies on manual tuning. We plan to conduct more in-depth research on this issue in the future, aiming to propose a better solution.

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

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14920–14929. IEEE.

Nasir Ahmed, T. Raj Natarajan, and K. R. Rao. 1974. Discrete cosine transform. *IEEE Trans. Computers*, 23(1):90–93.

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.

Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. SAME: sentiment-aware multi-modal embedding for detecting fake news. In *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019*, pages 41–48. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7132–7141. Computer Vision Foundation / IEEE Computer Society.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017a. 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.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017b. Novel visual and statistical image features for microblogs news verification. *IEEE Trans. Multim.*, 19(3):598–608.

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.

An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. 2024. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector. 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 18426–18434. AAAI Press.

Liangchen Liu, Xi Yang, Nannan Wang, and Xinbo Gao. 2021. Viewing from frequency domain: A dct-based information enhancement network for video person re-identification. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 227–235. ACM.

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.

Xing Shen, Jirui Yang, Chunbo Wei, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Xiaoliang Cheng, and Kewei Liang. 2021. Dct-mask: Discrete cosine transform mask representation for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8720–8729. Computer Vision Foundation / IEEE.

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.

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.

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.

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.

Chengpei Xu, Hao Fu, Long Ma, Wenjing Jia, Chengqi Zhang, Feng Xia, Xiaoyu Ai, Binghao Li, and Wenjie Zhang. 2024. Seeing text in the dark: Algorithm and benchmark. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 2870–2878. ACM.

Chengpei Xu, Wenjing Jia, Ruomei Wang, Xiaonan Luo, and Xiangjian He. 2023. Morphtext: Deep morphology regularized accurate arbitrary-shape scene text detection. *IEEE Trans. Multim.*, 25:4199–4212.

Yunsong Yang, Genji Yuan, and Jinjiang Li. 2024. Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. *IEEE Trans. Geosci. Remote. Sens.*, 62:1–17.

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.

Kairong Yu, Tianqing Zhang, Hongwei Wang, and Qi Xu. 2025. FSTA-SNN: frequency-based spatial-temporal attention module for spiking neural networks. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 22227–22235. AAAI Press.

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 Computational Cost

To objectively present the computational cost of MS-SFM, we use the baseline model (ResNet50+BERT) with all extended components removed as a reference, and the comparison data is shown in Table 5. Experimental results demonstrate that even with the integration of innovative modules such as frequency band selection and multi-scale fusion, MS-SFM still maintains a lightweight architectural feature: the model has only 112.9 million parameters and a computational complexity of 206.3 GFLOPs. In terms of key performance indicators, its accuracy is 8.5 percentage points higher than that of the baseline model (91.0% vs. 82.5%), while the inference latency per sample remains at an extremely low level of 0.0012 seconds. This synergistic advantage of high performance and high efficiency makes MS-SFM show significant adaptability in resource-constrained environments and latency-sensitive scenarios, especially suitable for real-time content detection tasks on social media platforms such as Twitter and Weibo.

Model	Parameters (M)	FLOPs (G)	Latency (s/sample)	Inference Time (s/sample)	Accuracy
MS-SFM	112.896	206.262	0.0012	0.0012	91.0%
ResNet50+BERT	111.516	205.650	0.0006	0.0008	82.5%

Table 5: Comparison of model efficiency and accuracy on the PHEME dataset.