Interpretable Short Video Rumor Detection based on Modality Tampering

Kaixuan Wu[1], Yanghao Lin[1], Donglin Cao[1*], Dazhen Lin[1]
[1] School of Informatics, Xiamen University, China
{kaixuanw,linyanghao}@stu.xmu.edu.cn
{another, dzlin}@xmu.edu.cn

Abstract

With the rapid development of social media and short video applications in recent years, browsing short videos has become the norm. Due to its large user base and unique appeal, spreading rumors via short videos has become a severe social problem. Many methods simply fuse multimodal features for rumor detection, which lack interpretability. For short video rumors, rumor makers create rumors by modifying and/or splicing different modal information, so we should consider how to detect rumors from the perspective of modality tampering. Inspired by cross-modal contrastive learning, we propose a novel short video rumor detection framework by designing two pretraining tasks: modality tampering detection and inter-modal matching, imbuing the model with the ability to detect modality tampering and employing it for downstream rumor detection tasks. In addition, we design an interpretability mechanism to make the rumor detection results more reasonable by backtracking the model's decision-making process. The experimental results show that the method on the short video rumor dataset has an improvement of about 4.6%-12% in macro-F1 compared with other models and can explain whether the short video is a rumor or not through the perspective of modality tampering.

Keywords: Short Video Rumor Detection, Modality Tampering, Contrastive Learning

1. Introduction

The proliferation of the Internet and mobile devices has streamlined information exchange, yet discerning the veracity of such information remains challenging for many, leading to rampant rumor dissemination. Traditionally, rumors surfaced as texts or images on social media, but the emerging trend of short video platforms has seen rumors evolve into video formats, which captivate viewers more effectively than plain text. While major video platforms have instituted manual review mechanisms to curb rumor spread, this approach is labor-intensive and time-consuming. Short video rumor detection models can automate the initial screening of potential rumor content, reserving detailed manual review for flagged content only. By integrating these models into the review processes of video platforms, we can bolster rumor detection efficiency, foster a cleaner online space, and make a tangible impact in the real world.

To solve the multimodal rumor problem, researchers have proposed various methods, and the previous methods can be roughly classified into two categories: pattern-based and evidence-based (Sheng et al., 2021; Hu et al., 2022). Since most rumor contents are often removed after they are officially identified as rumors, obtaining the original metadata and dissemination information is not easy. Therefore, we only use the static data of the rumors. Existing multimodal rumor detection follows the following paradigm: 1) Features are extracted from each modality using a heterogeneous feature extractor. 2) Subsequently, the features from each modality are fused and fed into a classification network. We also follow this paradigm.

Nevertheless, most of the existing multimodal rumor detection work focuses on the fusion of different modal features, ignoring some characteristics of short rumor videos: 1) By observing the data, we find that rumor videos not only contain inconsistent information among different modalities but also suffer from serious information tampering, such as manipulating textual content and splicing irrelevant image and audio information, which inspires us to consider detection from the perspective of modality tampering. 2) The current short video rumor datasets merely consist of rumor data and labels and thus cannot be used for auxiliary task learning and lacks external knowledge. Specifically, we believe prior methods have failed to consider the problems of multimodal information tampering and mismatch among multimodal information. To address this, we design two pre-training tasks for our model, namely modality tampering detection and inter-modality matching, to enhance the efficacy of rumor detection. By integrating information on modality tampering with rumor features, we propose a novel Short Video Rumor Detection Framework including pre-training and fine-tuning called **S**hort **V**ideo **R**umor **P**re-training **M**odel(SVRPM). Furthermore, by explaining the model's decision-making process, we design an interpretable mechanism to make the rumor detection results more transparent and reasonable. Experimental results

[*] Corresponding authors

LREC-COLING 2024, pages 9180–9189
20-25 May, 2024.

indicate that our approach excels in short video rumor detection and can provide intuitive explanations, which assist users in discerning the veracity of video information. We selected some existing multimodal rumor or fake news detection methods as a comparison method and conducted extensive experiments on the dataset we organized. The experimental results validate the effectiveness of our method.

The main contributions of this work are as follows:

- Aiming at the problem of deliberate tampering in short videos, we propose a short video rumor detection method based on modality tampering. The model is first forced to learn modality tampering detection and modality matching pre-training tasks and then use transfer learning for the downstream rumor detection task.
- We extended a short video rumor dataset and constructed the tampering dataset to support the task of modality tampering detection.
- We use the attention-backtracking mechanism to find local features that may have been tampered with to explain whether the short video is a rumor.
- We conducted extensive ablation experiments to demonstrate the effectiveness of the proposed method. Compared to other methods, there's an improvement of approximately 4.6%-12% in the macro-F1 score on our dataset.

2. Related Work

2.1. Rumor Detection

2.1.1. Unimodal Methods

Unimodal rumor detection generally focuses on a single modality, such as textual modality or visual modality, by establishing various frameworks to adapt the rumor detection task. Since many rumors spread on social media, some endeavors detect rumors by analyzing users' behavior on social media, social network structures, and information dissemination patterns. For example, (Li et al., 2021; Ran et al., 2022) constructed a heterogeneous graph from various user information on social media and achieved the best performance in rumor detection. Additionally, there are studies based entirely on textual modality. (Li et al., 2019) employed user information, attention mechanisms, and multitask learning for rumor detection. (Rao et al., 2021) introduced a novel variant of BERT specifically tailored for text-based rumor detection. Moreover, given that rumors and non-rumors typically have distinct patterns in image distribution, several works focus on image manipulation detection, investigating whether images have been tempered or are inconsistent with their background. (Zhou et al., 2018) proposed a two-stream Faster R-CNN network for image manipulation detection. (Cao et al., 2020) conducted a joint study investigating image forensic, semantic, statistical, and contextual features for fake news detection. Their study showed that visual content helps in rumor detection.

2.1.2. Multimodal Methods

Many recent multimodal-based approaches use cross-modal interaction and/or fusion to obtain better rumor detection performance. (Qi et al., 2021) analyzed the distinct features of named entities in textual and visual modalities, while CAFE (Chen et al., 2022) measured cross-modal ambiguity by evaluating the Kullback-Leibler divergence between unimodal feature distributions. FND-CLIP (Zhou et al., 2022) used BERT (Devlin et al., 2018) and ResNet (He et al., 2016) to extract text and image features, while CLIP (Radford et al., 2021) was used to compute similarity. (Qi et al., 2023) provided a new multimodal detection model named SV-FEND, which exploits the cross-modal correlations to select the most informative features and utilizes the social context information for detection.

However, research on multimodal rumor detection has yet to consider the perspective of modality tampering or inter-modality mismatching. In this work, we design two pre-training tasks, modality tampering detection and inter-modality matching, effectively utilizing features from different modalities and making the decision-making process more interpretable.

2.2. Supervised Contrastive Learning

Contrastive learning is a potent self-supervised representation learning paradigm (Chen et al., 2020; He et al., 2020). Its core concept is to learn representations by reducing the distance between similar (positive) samples and increasing the distance between dissimilar (negative) samples. Recently, there has been much work on cross-modal contrastive learning. CLIP (Radford et al., 2021) proposed a large-scale contrastive language-image pre-training model to address the unified representation of language and images. (Zolfaghari et al., 2021), in order to solve the problem of multimodal video representation, provides a contrastive loss to address intra-modal similarity while considering inter-modal similarity. Inspired by cross-modal contrastive learning, our work applies it to text, image, and audio modalities in short videos to detect modality tampering.

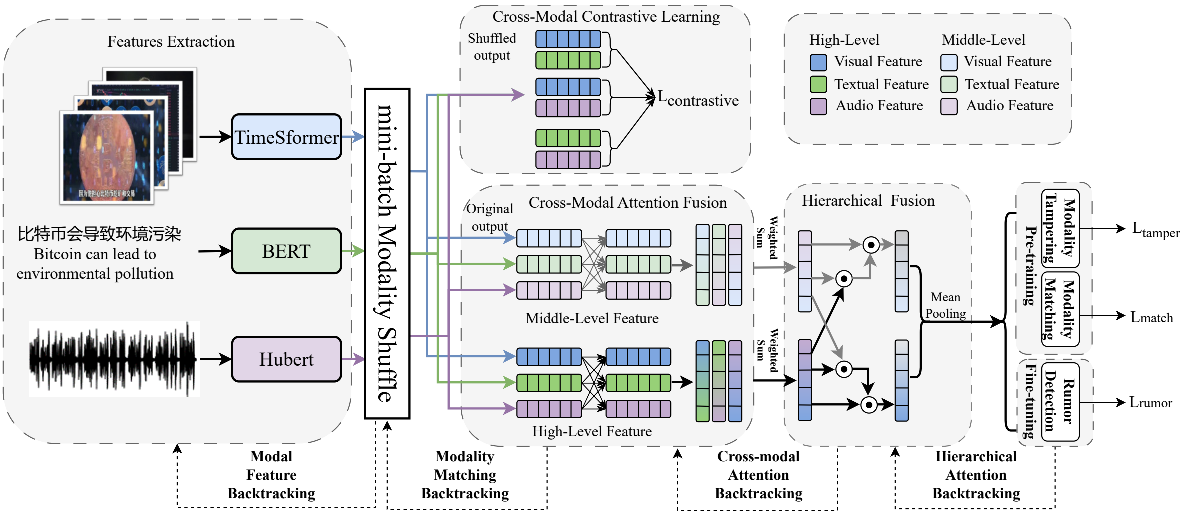


Figure 1: Model Architecture Overview of SVRPM. The model consists of five main modules: (a) Feature extraction: Extract visual, textual, and audio modal features using different encoders, respectively. (b) Mini-batch Modality Shuffle: Randomly shuffle their corresponding modal features for a minibatch. (c) Cross-modal Contrastive Learning: Constructing Positive and Negative Samples using Modality Shuffle Module. (d) Cross-modal Fusion and Hierarchical Fusion. (e) Modality Tampering Backtracking: Using attention backtracking operation to obtain the local features which may be tampered.

3. Proposed Method

In this section, we specifically describe our proposed model, SVRPM. As shown in Figure 1, the SVRPM model is composed of several components: Multimodal Feature Extraction Module, Cross-modal Contrastive Learning, Cross-Modal and Hierarchical Fusion Module, Tampering Selector and Modality Shuffle, and Attention Traceback.

3.1. Multimodal Feature Extraction

3.1.1. Textual Feature Extraction

We utilize a pre-trained BERT model[1] for extracting text features. We input the cleaned short video title and OCR text sequence $(w_1, w_2, ..., w_n)$ into BERT to extract textual features $\boldsymbol{H_T} = \boldsymbol{BERT}(w_1, w_2, ..., w_n)_{[CLS]}$, where $\boldsymbol{H_T} \in \boldsymbol{R}^{1*768}$ is the feature vector of the text sequence, and $[CLS]$ is the special token.

3.1.2. Visual Feature Extraction

After video frame extraction and removal of interfering regions. We opt for the TimerSformer model[2] to obtain a visual feature vector with global information. Like textual feature extraction, we input eight processed video frames into the model, and the output feature is $\boldsymbol{H_V} = \boldsymbol{TimeSformer}(f_1, f_2, ..., f_8)_{[CLS]}$, where f_i denotes the i-th frame, $\boldsymbol{H_V} \in \boldsymbol{R}^{1*768}$ is the feature vector of the video, and $[CLS]$ is the special token.

3.1.3. Audio Feature Extraction

We encode single-channel audio through Hubert (Hsu et al., 2021) model[3] and use MeanPooling for all the tokens to obtain a global audio representation as feature vectors. We input 500000 audio samples into the model, and the output feature is denoted as $\boldsymbol{H_A} : \{\boldsymbol{h_1}, \boldsymbol{h_2}, ..., \boldsymbol{h_n}\} = \boldsymbol{Hubert}(a_1, a_2, ..., a_{500000})$(1)

$$\boldsymbol{H_A} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h_i} \tag{2}$$

where a_i denotes the i-th audio sample, $\boldsymbol{H_A} \in \boldsymbol{R}^{1*768}$ is the feature vector of the audio sequence, and n is the number of audio tokens.

3.2. Multimodal Fusion

In contrast to previous work, we avoid using simple concatenation fusion. Instead, we implement cross-modal fusion to enhance the integration of multiple modalities. (Hsu et al., 2021) find that the representations of the middle layer were the most helpful for the downstream task. Inter-modal complementarity information may be distributed in different layers, leading us to adopt hierarchical fusion.

[1] https://huggingface.co/bert-base-chinese

[2] https://github.com/facebookresearch/TimeSformer

[3] https://huggingface.co/TencentGameMate/chinese-hubert-base

3.2.1. Cross-Modal Fusion

Since short video rumor information may exist in different modalities simultaneously or only in a single modality, there is variability among the information of multiple modalities. Better detection effects can be achieved through information complementation. We utilize the cross-modal attention mechanism for fusion to enable the model to assign importance to different modalities independently. We adopt this fusion approach for middle-level (Layer 6) and high-level (Layer 12) features. As an example, we introduce cross-modal attention with high-level features.

Initially, the multimodal feature matrix $\boldsymbol{M_{TVA}}$ is obtained by stacking the feature vectors of the three modalities.

$$\boldsymbol{M_{TVA}} = [\boldsymbol{H_T}, \boldsymbol{H_V}, \boldsymbol{H_A}]^T \tag{3}$$

where $\boldsymbol{M_{TVA}} \in \boldsymbol{R}^{3*768}$ is the multimodal feature matrix.

Then, the feature vector of each modality is used as the query vector $\boldsymbol{q_i}$ to compute its attention weight on different modalities, yielding a weight vector $\boldsymbol{\alpha} \in \boldsymbol{R}^{1*3}$,

$$\boldsymbol{\alpha} = \boldsymbol{Softmax}(Attention(\boldsymbol{q_i}, \boldsymbol{M_{TVA}})) \tag{4}$$

where $\boldsymbol{q_i}$ is the query vector of the i-th modality.

Finally, the cross-modal feature of this query is obtained by weighting the three modalities, denoted as $\boldsymbol{H_{TVA_H}} \in \boldsymbol{R}^{1*768}$. The same method is applied to obtain the middle-level feature $\boldsymbol{H_{TVA_M}} \in \boldsymbol{R}^{1*768}$.

$$\boldsymbol{H_{TVA_H}} = \boldsymbol{Sum}(\boldsymbol{\alpha} * \boldsymbol{M_{TVA}}) \tag{5}$$

3.2.2. Hierarchical Fusion

The high-level features of the pre-trained model are highly related to the pre-trained tasks, which are almost irrelevant to the rumor detection task. Pre-trained models are often multilayered, and irrelevant features are eliminated as the layer increases. This forgotten information should be considered during the fusion of multiple single-modality models. Thus a hierarchical fusion module is added to the model, and layer six is selected as the middle-level feature for all modalities. The middle-level features $\boldsymbol{H_{TVA_M}}$ and the high-level features $\boldsymbol{H_{TVA_H}}$ are stacked denote as $\boldsymbol{M_{HM}}$ and input to the cross-modal attention module and also used as the query vectors $\boldsymbol{q_M}, \boldsymbol{q_H}$ respectively. Subsequently, through attention-weighted summation, the hierarchical features denoted as $\boldsymbol{\hat{H}_{TVA_H}}$ and $\boldsymbol{\hat{H}_{TVA_M}}$. Finally, after Mean-Pooling, we get the global features with hierarchical fusion.

$$\boldsymbol{M_{HM}} = [\boldsymbol{H_{TVA_H}}, \boldsymbol{H_{TVA_M}}]^T \tag{6}$$

$$\boldsymbol{\alpha_H} = \boldsymbol{Softmax}(Attention(\boldsymbol{q_H}, \boldsymbol{M_{HM}})) \tag{7}$$

$$\boldsymbol{\alpha_M} = \boldsymbol{Softmax}(Attention(\boldsymbol{q_M}, \boldsymbol{M_{HM}})) \tag{8}$$

$$\boldsymbol{\hat{H}_{TVA_H}} = \boldsymbol{Sum}(\boldsymbol{\alpha_H} * \boldsymbol{M_{HM}}) \tag{9}$$

$$\boldsymbol{\hat{H}_{TVA_M}} = \boldsymbol{Sum}(\boldsymbol{\alpha_M} * \boldsymbol{M_{HM}}) \tag{10}$$

$$\boldsymbol{H_{TVA_{HF}}} = \boldsymbol{MeanPooling}([\begin{matrix} \boldsymbol{\hat{H}_{TVA_H}} \\ \boldsymbol{\hat{H}_{TVA_M}} \end{matrix}]) \tag{11}$$

where $\boldsymbol{H_{TVA_{HF}}} \in \boldsymbol{R}^{1*768}$ is the hierarchical multimodal feature.

3.3. Tampering Selector and Modality Shuffle

Considering the potential manipulation (such as tampering with text) and modality mismatch (i.e., most of the information between different modalities is dissimilar) in manually created short videos, we designed two pre-training tasks, modality tampering detection, and modality matching.

3.3.1. Tampering Selector

To simulate the effect of tampering, we achieve it by tampering with some words of the video title. The specific way is as follows: 1). Use "HanLp"[4] to perform lexical segmentation, then take all nouns, adjectives, and adverbs as tampered words for each video title. 2). Use "Synonyms"[5] to obtain the four words with the closest semantic similarity to the tampered words as candidate words. 3). For each video title, 1 to 3 words to be tampered with. Before inputting the text into the model, it goes through a tampering selector to choose whether to tamper or not with a certain probability. Using contrastive learning, tampered samples have an increased distance between modalities, whereas untampered samples have a reduced distance between modalities.

3.3.2. Modality Shuffle

Rumors crafted manually often splice disparate titles, images, and audio, resulting in varying modalities presenting incongruent information. To heighten our model's capability in detecting such rumors, we generate negative samples by shuffling multimodal features from different samples during training, as depicted in Figure 2. We begin by copying a mini-batch's modal features, then shuffle this copied data. Modality mismatch samples are formed by consistently shuffling high and mid-level features. These paired matching and mismatching samples are then input into the subsequent module, and processed through a classification head.

[4] https://github.com/hankcs/HanLP

[5] https://github.com/chatopera/Synonyms

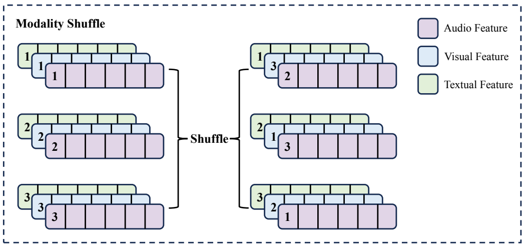


Figure 2: Modality Shuffle Module. Different colors represent different modalities, and different numbers represent different samples in a mini-batch. Compared with modality tampering, modality shuffle belongs to coarse-grained tampering.

During pretraining, we aim to widen the gap between differing modalities in mismatched samples and narrow it in matched ones.

3.4. Classification and Loss Function

3.4.1. Classification

During the entire model pretraining process, two tasks are set: modality tampering detection and modality matching. Classifiers are set up to map the global feature vector $\boldsymbol{H_{TVA_{HF}}}$ to a two-dimensional space through a linear layer.

3.4.2. Loss Function

In the pretraining tasks, cross-modal contrastive learning is added to make the feature similarity of each modality as close as possible. This involves three combinations of text-visual, text-audio, and visual-audio, and measuring their cosine similarity with $\boldsymbol{Sim_{TV}}$, $\boldsymbol{Sim_{TA}}$, and $\boldsymbol{Sim_{VA}}$, respectively.

$$\boldsymbol{Sim_{TV}} = \frac{\boldsymbol{H_T H_V^T}}{\|\boldsymbol{H_T}\|\|\boldsymbol{H_V}\|} \tag{12}$$

$\boldsymbol{Sim_{TA}}$, and $\boldsymbol{Sim_{VA}}$ are obtained in the same way. The contrastive loss function of modality tampering can be written as follows ($\boldsymbol{Sim_{VA}}$ is not applicable in equation 13):

$$\boldsymbol{\mathcal{L}}_{c_t} = \sum_{i=1}^{N} (-1)^{m_i} (\boldsymbol{Sim_{T_iV_i}} + \boldsymbol{Sim_{T_iA_i}}) \tag{13}$$

where m_i is the tampering label of the i-th sample, N is the number of samples in a mini-batch, and $m_i = 0$ indicates that the sample is tampered, otherwise $m_i = 1$.

$$\begin{aligned} \boldsymbol{\mathcal{L}}_{tamper} = -\sum_{i=1}^{N} [m_i log(\boldsymbol{F_t}(\boldsymbol{H_i})) \\ + (1 - m_i) log(\boldsymbol{F_t}(\boldsymbol{H_i}))] \end{aligned} \tag{14}$$

$$\boldsymbol{\mathcal{L}}_t = \boldsymbol{\mathcal{L}}_{c_t} + \boldsymbol{\mathcal{L}}_{tamper} \tag{15}$$

Cross-entropy loss is used in tamper detection along with contrastive loss. The tamper detection loss is denoted as $\boldsymbol{\mathcal{L}_{tamper}}$, while $\boldsymbol{F_t}$ represents the linear classification layer and $\boldsymbol{H_i}$ denotes the final multi-modal features. The contrastive loss in the modality matching task is similar.

$$\boldsymbol{\mathcal{L}}_{c_m} = \sum_{i=1}^{N} (-1)^{n_i} (\boldsymbol{Sim_{T_iV_i}} + \boldsymbol{Sim_{T_iA_i}} + \boldsymbol{Sim_{V_iA_i}}) \tag{16}$$

where n_i is the label of the i-th sample, N is the number of samples in a mini-batch, and $n_i = 0$ indicates that the sample is mismatched, otherwise $n_i = 1$.

The modality matching classification loss $\boldsymbol{\mathcal{L}}_{match}$ and total loss $\boldsymbol{\mathcal{L}}_m$ are as follows:

$$\begin{aligned} \boldsymbol{\mathcal{L}}_{match} = -\sum_{i=1}^{N} [n_i log(\boldsymbol{F_m}(\boldsymbol{H_i})) \\ + (1 - n_i) log(\boldsymbol{F_m}(\boldsymbol{H_i}))] \end{aligned} \tag{17}$$

where $\boldsymbol{F_m}$ denotes the linear classification layer.

$$\boldsymbol{\mathcal{L}}_m = \boldsymbol{\mathcal{L}}_{c_m} + \boldsymbol{\mathcal{L}}_{match} \tag{18}$$

3.5. Modality Tampering Backtracking

We use attention backtracking to detect possible modality tampering in short videos, indicating potential misinformation. See Algorithm 1. Hierarchical attention parameters are denoted as $Attn_{Hr}$. The high-level and middle-level cross-modal attention fusion layers are represented by $Attn_{CMH}$ and $Attn_{CMM}$. We first calculate hierarchical attention scores (Equation 7, 8): high-level ($Score_H$) and middle-level ($Score_M$). Cross-modal attention scores are determined by comparing these levels. If $Score_H > Score_M$, we use $Attn_{CMH}$; otherwise, we use $Attn_{CMM}$. Next, we compare scores for modalities T, V, and A. The highest scoring modality x denotes the most significant features (Equation 4). The attention parameter is $Attn_{xM\alpha}$ (α just for indicating the high or middle level). By now, we can get the attention $Score$ for each $Token$, and its index is n. Lastly, each $Token$ is a local feature, higher scores suggest more attention and tampering risk. The top k features are visualized.

4. Experiments

4.1. Dataset

1) We use the dataset from (You et al., 2022). The data used in this paper is crawled from Tiktok and manually labeled. There are 584 rumor

Algorithm 1 Modality Tampering Backtracking

```
Input: $H_{TVA_M}, H_{TVA_H}, H_T, H_V, H_A, Token$
Output: Modality local original features.
Initialize: $Attn_{Hr}, Attn_{CMH}, Attn_{CMM}$
            $Attn_{xMH}, Attn_{xMM} \leftarrow \boldsymbol{x}$ modal high-level and middle-level attention, $\boldsymbol{x} \in \{T, V, A\}$
$[Score_H, Score_M]^T = Attn_{Hr}(H_{TVA_H}, [H_{TVA_H}, H_{TVA_M}]^T) + Attn_{Hr}(H_{TVA_M}, [H_{TVA_H}, H_{TVA_M}]^T)$, (Equation 7, 8)
if $Score_H > Score_M$ then
    $[Score_T, Score_V, Score_A]^T = Sum\left(Attn_{CMH}\left([H_T, H_V, H_A]^T, [H_T, H_V, H_A]^T\right)\right)$, (Equation 4)
    modality $\boldsymbol{x} = Max_{modal}\left([Score_T, Score_V, Score_A]^T\right)$, the values of $\boldsymbol{x}$ are $T, V, A$
    $[Score_{Token_1}, \cdots, Score_{Token_n}]^T = Attn_{xMH}\left(Token_{cls}, [Token_1, \cdots, Token_n]^T\right)$, Token from modality $\boldsymbol{x}$
else
    $[Score_T, Score_V, Score_A]^T = Sum\left(Attn_{CMM}\left([H_T, H_V, H_A]^T, [H_T, H_V, H_A]^T\right)\right)$, (Equation 4)
    modality $\boldsymbol{x} = Max_{modal}\left([Score_T, Score_V, Score_A]^T\right)$, the values of $\boldsymbol{x}$ are $T, V, A$
    $[Score_{Token_1}, \cdots, Score_{Token_n}]^T = Atten_{xMM}\left(Token_{cls}, [Token_1, \cdots, Token_n]^T\right)$, Token from modality $\boldsymbol{x}$
end if
Local feature index: $Index = Max_{index}\left([Score_{Token_1}, \cdots, Score_{Token_n}]^T, k\right)$
return Local feature: $Feature_x\,(index)$ =0
```

short videos and 625 non-rumor videos. To ensure that the data aligns with the actual distribution, we extend the dataset by crawling a lot of the non-rumor data from TikTok in the same way. We merge these and divide the dataset for training and testing, respectively. 2) We use FakeSV (Qi et al., 2023) which is the largest Chinese short video dataset about fake news. The details of the division of the dataset are shown in Table 1.

Table 1: The statistics of two datasets.

Dataset	Split	Rumor	non-Rumor	Total
Ours	train	467	4795	5262
	test	117	1204	1321
FakeSV	train	1233	1303	2536
	test	304	238	542

4.2. Implementation Details

We use the Adam optimizer with a learning rate of 2e-5. The maximum textual sequence length is 256. The number of frames input to TimeSformer is 8. The number of audio samples input to Hubert is 500000. For category rumor, we use 10x oversampling.

Pre-training task setup. For the modality tampering task, we set the tampering probability as 0.5, the maximum of tampering words as 3, and the tampering candidates as 4 with no modality shuffle. We use the data in Section 4.1 and produce tampered data using the method in Section 3.3.1, totaling 5,155,350 pieces of data. The ratio of the training set to the test set is 8:2. For the modality matching task, we set the tampering probability as 0 with modality shuffle.

Fine-tuning task setup. For the downstream rumor detection task, due to the category imbalance, we use a category-balanced focal loss with a tampering probability of 0 without shuffling the modality and contrastive learning.

Evaluation Metrics. we use Accuracy, Precision, Recall, and F1 as evaluation metrics.

4.3. Baselines

SAFE (Zhou et al., 2020) selects Text-CNN as the textual feature extractor, and the image2sentence pre-trained model is selected to extract visual features. The overall information of the rumor is obtained by concatenating different modal features and semantic similarity features. ViLT (Kim et al., 2021) is a pre-trained visual-language model on an English dataset that merges text embeddings into a visual transformer (ViT). To maintain fairness, the textual content of the short video rumor data is translated into English. VideoMae (Tong et al., 2022) extract more spatiotemporal features of the video and use them for the short video rumor classification task. MEA (Wei et al., 2022) extracts separate features for different modalities, fuses them with a linear layer, and uses them for rumor detection tasks. CHEF (Hu et al., 2022)is a tool for fact-checking that retrieves evidence from relevant documents to predict the accuracy of a claim. We used a search engine to gather title-related data as evidence.

4.4. Experimental Results

4.4.1. Results of pre-training tasks

In Table 5, it can be found that the accuracy is higher than 80% on both tasks, which indicates that the pre-trained model can better identify whether

Table 2: Results(%) of different methods on our short video rumor dataset and FakeSV dataset. "*" denotes that the text content is in English. The subscript "0" represents "Rumor as Positive" and "1" denotes "non-Rumor as Positive" in computing the precision, and recall. "F1" denotes macro-F1 values. "–" means that corresponding experiments were not carried out due to lack of partial data. The best performance is highlighted in boldface.

Method	Ours						FakeSV					
	Acc	F1	P_0	P_1	R_0	R_1	Acc	F1	P_0	P_1	R_0	R_1
SAFE(2020)	96.90	90.98	78.36	98.99	89.74	97.59	77.49	77.43	84.21	71.01	73.68	82.35
ViLT*(2021)	81.53	65.46	29.04	97.15	75.21	82.14	–	–	–	–	–	–
VideoMae(2022)	98.03	93.55	94.17	98.36	82.91	99.50	73.80	72.99	74.40	72.86	81.25	64.29
MEA(2022)	94.70	86.19	64.07	99.13	91.45	95.02	70.66	70.51	**86.43**	61.52	56.58	**88.66**
CHEF(2022)	97.58	92.67	84.55	98.91	88.89	98.42	–	–	–	–	–	–
SVRPM(ours)	**99.39**	**98.15**	**95.04**	**99.83**	**98.29**	**99.50**	**79.34**	**78.55**	78.07	**81.50**	**87.83**	68.49

Table 3: Comparison of results(%) on Chinese-to-English and Chinese. "-En" denotes the translation of Chinese into English, and "-Ch" denotes the experimental result of directly adopting Chinese.

Method	Acc	P	P_0	P_1	R	R_0	R_1	$F1$	$F1_0$	$F1_1$
ViLT-En	81.53	63.10	29.04	97.15	78.68	75.21	82.14	65.46	41.90	89.02
ViLT-Ch	78.34	59.44	23.25	95.63	71.19	62.39	79.98	60.49	33.87	87.11
SVRPM-En	98.71	96.52	93.86	99.17	95.44	91.45	99.42	95.97	92.64	99.29
SVRPM-Ch	**99.39**	**97.44**	**95.04**	**99.83**	**98.90**	**98.29**	**99.50**	**98.15**	**96.64**	**99.67**

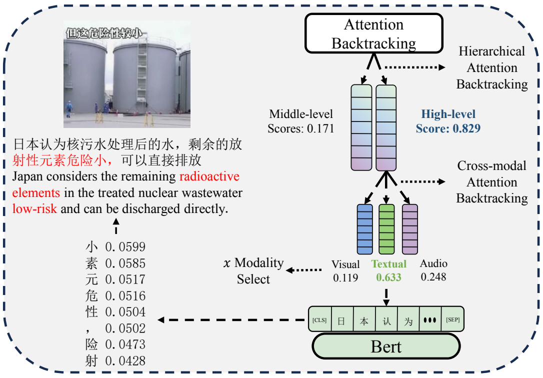

但这危险性较小
日本认为核污水处理后的水，剩余的放射性元素危险小，可以直接排放
Japan considers the remaining radioactive elements in the treated nuclear wastewater low-risk and can be discharged directly.
小 0. 0599
素 0. 0585
元 0. 0517
危 0. 0516
性 0. 0504
， 0. 0502
险 0. 0473
射 0. 0428
Attention Backtracking
Hierarchical Attention Backtracking
Middle-level Scores: 0.171
High-level Score: 0.829
Cross-modal Attention Backtracking
x Modality Select
Visual 0.119
Textual 0.633
Audio 0.248
[CLS]
日
本
认
为
[SEP]
Bert

Figure 3: Modality Tampering Backtracking Visualization. Backtracking hierarchical and cross-modal attention scores and the corresponding modality is identified. In this case, the critical local features are in the text. The words in red are of the highest attention.

the modality matches or not and whether the modality has been tampered or not, and it can also assist the model in performing the task of rumor detection in short videos and explaining the reasons.

4.4.2. Compare with Baselines

As shown in Table 2, our proposed model "SVRPM" outperforms other models in short video rumor detection. Except for ViLT, SVRPM has about 1.3%-4.7% improvement in accuracy and about 4.6%-12% enhancement in macro-F1 value compared to other models on our dataset. Notably, SAFE, MEA, CHEF, and SVRPM exhibit superior recall values in rumor detection. Conversely, VideoMae registers commendable performances across several metrics but is found wanting in recall, attributable to an information deficit in its unimodal pre-training. ViLT lags behind all counterparts, including even the unimodal pre-trained model. This could be ascribed to potential losses of critical rumor characteristics during textual translation, especially if the resultant style diverges from its native form. Moreover, the features of the multimodal model are fused from the low level, so the feature extraction is easily affected by other modalities, and the translated text appears to be mismatched with the visual modality, so there is a serious modality tampering problem in the translated text. Table 3 compares the results of Vilt(pre-trained on the English dataset) on Chinese-to-English and Chinese titles. It can be noticed that our model is much less affected by language than ViLT and outperforms direct fine-tuning with modality tampering tasks.

4.4.3. Ablation Study

We conduct an evaluation of our model to decipher the impact of each component. Table 4 presents the results of our ablation experiments, from which we draw several insights and highlight the most salient observations below.

Cross-Modal Fusion: The data in Table 4 reveals that cross-modal fusion outperforms concatenated fusion, exhibiting a notable 3%-4% boost in rumor recall. This enhancement facilitates more accurate rumor detection. Moreover, there's a significant 6%-10% improvement in the macro-F1 score. Such results underscore the distributed nature of rumor information across different modalities. The

Table 4: Performance(%) of ablation experiments. For simplicity, modalities are abbreviated("T": Textual modality, "V": Visual modality, "A": Audio modality). "CMF" denotes cross-modal fusion, "HF" denotes hierarchical fusion, and Tamper, Match stand for pre-training tasks, respectively.

Method	Modality	CMF	HF	Tamper	Match	Acc	P	R	F1
Concat	T, V	✗	✗	✗	✗	93.19	78.16	92.02	83.18
Concat	T, V, A	✗	✗	✗	✗	96.37	85.82	96.46	90.23
SVRPM	T, V	✓	✗	✗	✗	98.18	95.33	93.22	94.24
SVRPM	T, V	✗	✓	✗	✗	98.18	94.37	94.37	94.37
SVRPM	T, V, A	✓	✗	✗	✗	98.86	97.00	95.90	96.44
SVRPM	T, V, A	✗	✓	✗	✗	99.17	**97.60**	97.23	97.41
SVRPM	T, V, A	✓	✓	✓	✗	99.24	97.31	98.04	97.67
SVRPM	T, V, A	✓	✓	✗	✓	99.24	96.36	**99.20**	97.73
SVRPM	T, V, A	✓	✓	✓	✓	**99.39**	97.44	98.90	**98.15**

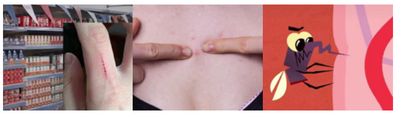
Title	Audio (speech to text)	Key Frames	Prediction
Chinese: 注意！这样艾滋就会找上你 English: Pay attention! This is how AIDS finds you.	Chinese: 超市里面划破手就会得艾滋；胸前长痘痘就是艾滋；小心蚊子，因为蚊子也会传染艾滋。 English: You can get AIDS if you cut your hand inside a supermarket; a pimple on your chest is AIDS; and watch out for mosquitoes because they can also transmit AIDS.		Rumor
Chinese:日本认为核污水处理后的水，剩余的放射性元素危险小，可以直接排放 English: Japan considers that the remaining radioactive elements in the treated nuclear wastewater are of low risk and can be discharged directly.	Chinese:日本政府认为这些水经过技术处理后，剩下的放射性元素主要是氚，它虽然含量相对较高，但这危险性较小，衰变产生的低能量β粒子对生命组织造成的损伤也比较小，因此可以直接排放。 English: The Government of Japan believes that after the technical treatment of this water, the remaining radioactive element is mainly tritium, which has a relatively high content, but this is less dangerous, and the low-energy beta particles produced by decay cause less damage to living tissues, so it can be discharged directly.		Rumor
Chinese:日本人经常吃生鱼片，是怎么预防寄生虫的 English: Japanese people often eat sashimi, how do they prevent parasites.	Chinese: 日本人经常吃生鱼片难道不怕长寄生虫吗？原来在日本一般会选择深海鱼来做生鱼片，因为深海鱼的寄生虫相对较少，而且会采用低温冷冻来运输。 English: Japanese people often eat sashimi is not afraid of growing parasites? It turns out that in Japan usually choose deep-sea fish to make sashimi, because deep-sea fish has relatively fewer parasites, and will use low-temperature refrigeration to transport.		Non-Rumor

Figure 4: Qualitative examples of short videos are provided. Examples with a green background indicate non-rumors, while those with a red background signify rumors. Only a subset of the audio (converted from speech to text) is shown for display purposes.

Table 5: Experimental results(%) for the modality matching task and the modality tampering task. Acc, P, R, and F1 stand for Accuracy, Precision, Recall, and macro-F1, respectively.

Task	Acc	P	R	F1
Modality Matching	82.56	84.96	79.05	81.90
Modality Tampering	86.41	88.82	83.32	85.98

cross-modal attention effectively allocates attention weights for various rumor samples, prioritizing modalities pivotal for rumor detection.

Hierarchical Fusion: We incorporate hierarchical fusion across various modality combinations to elucidate the significance of this fusion approach. Empirical results indicate a universal enhancement across all metrics when tested with three modalities. Notably, there's a 2.5% uptick in the rumor recall, bolstering rumor detection efficacy. This proves that rumor information may be distributed in different feature layers, and the interaction of information between the layers can help the model detect rumors better.

Pre-training Tasks: We evaluate models enhanced with various pre-training tasks to ascertain their impact on rumor detection. While the inclusion of the modality matching and tampering tasks slightly diminishes Precision, there's a notable rise in Recall for rumors. This trend implies a greater tendency of the model to categorize ambiguous non-rumors as rumors. Even though modality mismatches and tampering can occur in both rumors and non-rumors, adopting stringent criteria for rumor detection is justified given the potential harm of rumors. Intriguingly, our experiments reveal that jointly utilizing the two pre-training tasks fosters mutual reinforcement, culminating in an enhanced macro-F1 score. This underscores the efficacy of modality matching and tampering tasks in refining multimodal rumor detection.

4.4.4. Modality Tampering Backtracking

We provide a visualization of modality tampering backtracking for a specific instance in Figure 3 and highlight successful predictions in Figure 4. Through this backtracking analysis, peak attention aligns with the item most pertinent to the query.

5. Conclusion

We design an interpretable short video rumor detection model based on modality Tampering, which performs the pre-training task of modality tampering recognition on the modality tampering dataset, and performs the pre-training task of modality matching by shuffling the modalities. For the downstream rumor detection task, we use the transfer learning approach. The model can focus on whether there is a modality tampering between multimodal local features and visualize the local tampering features using the modality tampering backtracking to improve the interpretability.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62076210, No. 81973752), the Natural Science Foundation of Xiamen city (No. 3502Z20227188) and the Open Project Program of The Key Laboratory of Cognitive Computing and Intelligent Information Processing of Fujian Education Institutions, Wuyi University(No. KLCCIIP2020203).

7. References

Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 141–161.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Xuming Hu, Zhijiang Guo, Guanyu Wu, Aiwei Liu, Lijie Wen, and Philip S Yu. 2022. Chef: A pilot chinese dataset for evidence-based fact-checking. *arXiv preprint arXiv:2206.11863*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Chen Li, Hao Peng, Jianxin Li, Lichao Sun, Lingjuan Lyu, Lihong Wang, S Yu Philip, and Lifang He. 2021. Joint stance and rumor detection in hierarchical heterogeneous graph. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2530–2542.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1173–1179.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14444–14452.

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

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

Hongyan Ran, Caiyan Jia, Pengfei Zhang, and Xuanya Li. 2022. Mgat-esm: Multi-channel graph attention neural network with event-sharing module for rumor detection. *Information Sciences*, 592:402–416.

Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. Stanker: Stacking network based on level-grained attention-masked bert for rumor detection on social media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3347–3363.

Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating pattern-and fact-based fake news detection via model preference learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1640–1650.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

Pengfei Wei, Fei Wu, Ying Sun, Hong Zhou, and Xiao-Yuan Jing. 2022. Modality and event adversarial networks for multi-modal fake news detection. *IEEE Signal Processing Letters*, 29:1382–1386.

Jinpeng You, Yanghao Lin, Dazhen Lin, and Donglin Cao. 2022. Video rumor classification based on multi-modal theme and keyframe fusion. In *CCF Conference on Computer Supported Cooperative Work and Social Computing*, pages 58–72. Springer.

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061.

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

Yangming Zhou, Qichao Ying, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. 2022. Multimodal fake news detection via clip-guided learning. *arXiv preprint arXiv:2205.14304*.

Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. 2021. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459.