T^{3}SVFND: Towards an Evolving Fake News Detector for Emergencies with Test-time Training on Short Video Platforms

Liyuan Zhang, Zeyun Cheng, Zhongyan Gui(✉), Yan Yang (✉), Yong Liu, and Jinke Ma

Heilongjiang University, Harbin, China
{2231976,2231975}@s.hlju.edu.cn,{guizhongyan,yangyan,liuyong123456}@hlju.edu.cn

Abstract. The existing methods for fake news videos detection may not be generalized, because there is a distribution shift between short video news of different events, and the performance of such techniques greatly drops if news records are coming from emergencies. We propose a new fake news videos detection framework (T^{3}SVFND) using Test-Time Training (TTT) to alleviate this limitation, enhancing the robustness of fake news videos detection. Specifically, we design a self-supervised auxiliary task based on Mask Language Modeling (MLM) that masks a certain percentage of words in text and predicts these masked words by combining contextual information from different modalities (audio and video). In the test-time training phase, the model adapts to the distribution of test data through auxiliary tasks. Extensive experiments on the public benchmark demonstrate the effectiveness of the proposed model, especially for the detection of emergency news.

Keywords: Misinformation video detection · Test-time training · Multimodal learning · Model robustness · Social networks.

1 Introduction

Due to the characteristics of short duration, concentrated content and strong expression, sharing short videos have penetrated deeply into the daily life of the public. The low entry threshold, few self-censorship mechanisms and other factors have caused a large number of fake news on short video platforms. Previous image text fake news detection methods [14, 26, 35, 36] are difficult to apply directly to news videos due to different modalities. Developing reliable methods for automated detection of fake news videos using artificial intelligence technology is currently a top priority.

This work was supported by the National Natural Science Foundation of China (No. 6247074060), the Natural Science Foundation of Heilongjiang Province in China (No. PL2024F029) and the Basic Research Funds for Provincial Universities in Heilongjiang Province (No. 2023-KYYWF-1486, No. 2024-KYYWF-0115).

Recently, a large-scale fake news short videos dataset (FakeSV) [19] has been proposed, a number of fake news videos detection methods have demonstrated their effectiveness on this dataset. Despite the success of these methods, when they are faced with emergency news, there is a generally sharp decline in performance, calls into question about their reliability. Traditional fake news detection models assume the same training and testing distribution, but actual deployment requires the capacity of model to generalize unseen and out-of-distribution data which can ensure that the effectiveness still remains in the dynamic real world. However, existing fake news videos detection methods [6, 24, 23, 2] do not take this into account, which results in failing to maintain claimed performance on previously unseen events.

We introduce a **T**est-**T**ime **T**raining (TTT) framework to alleviate this limitation. With TTT algorithm as the core, we design an auxiliary task based on the self-supervised **M**ask **L**anguage **M**odeling (MLM) to force the model to adapt to the distribution of test data to capture the inherent features hidden in fake news videos. Specifically, we design a multimodal masked Transformer module (M-Transformer^2) to integrate information from audio and video respectively to predict the truth value of masked words. In the test-time training phase, we fine-tune the trained model through auxiliary tasks (converted into a self-supervised learning problem), as shown in Fig. 1, which makes our model focus not only on the binary labels, but also on the distribution of the data itself, ensures the adaptability and robustness of the model in dynamic environment.

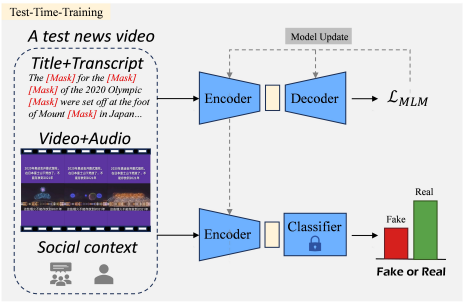

Fig. 1. Overview of our testing time training methods. During the testing time training phase, we update the weights of the encoder by using a self-supervised task to adapt to the distribution shift of the test data, while the weights of the classifier are fixed during this process. Then, we use the updated weights to predict the test data.

Our method does not improve the robustness of the model rely on specific structures, but rather on unlabeled data. The core idea of the TTT framework is that from an information perspective, previously we only used the informational from the training set (supervised learning) to train neural networks, but in fact, the testing set also provides information from the perspective of data distribution. Our research brings hope that in the real world, we only need to collect some news from emergencies (without expensive manual annotation) to fine-tune the model, which can significantly improve its adaptability to emergencies. This focus on adaptive learning marks an significant step forward in combating misinformation videos.

The main contributions of this paper are as follows:

- **Idea:** We for the first time consider the problem of model robustness for detecting fake news on short video platforms. Existing methods are often ineffective when dealing with emergencies due to the lack of appropriate mechanisms to adapt to the dynamic social media environment. It is imperative to transcend current paradigms and improve the adaptability of the model to emergencies.
- **Method:** We propose T^{3}SVFND, a novel fake news videos detection model, which addresses the key challenge by introducing the TTT framework and utilizing a carefully designed self-supervised task based on MLM to learn latent features of multimodal data. To our knowledge, this is the first study on the robustness of fake news videos detection.
- **Effectiveness:** We conducted extensive experiments on a large-scale fake news short videos dataset, compared with the most advanced methods, T^{3}SVFND achieves SOTA results in both event-based and time-based data segmentation scenarios (up to 2.48% and 3.32% in accuracy). The ablation study validated the effectiveness of the different modules. Our codes is publicly available in https://github.com/ZhangLiyuan11/TTTSVFND.

2 Related Work

In this section, we review the related work on (1) Fake news videos detection; (2) Domain adaptation in fake news detection; (3) Test-time Training Framework. We also explain the innovative aspects of our work.

2.1 Fake News Videos Detection

With the development of deep neural networks [12, 13], some studies have extracted multimodal features of news videos and established cross-modal correlation models, e.g., Choi et al. [2] attempted to identify differences in stance through differences in topic distribution between titles/descriptions and comments. Recently, Qi et al. [19] collected a new dataset of large-scale Chinese fake news short videos and proposed a benchmark by fusing multimodal clues through multiple cross-attention modules. The NEED framework [20] proposed incorporating neighbor relationship in an event for fake news videos detection. Wu et al. [31] made the results of fake news videos detection more reasonable by backtracking the decision-making process of the model.

However, they did not take into account the challenges brought by the dynamic changes in the real world. To overcome the limitations of existing works, we propose a new test-time training based fake news videos detection model T^{3}SVFND, which introduces the TTT framework and a carefully designed MLM task to improve classification performance and reduce the potential risk of feature distribution shift.

2.2 Domain Adaptation in Fake News Detection

Many works research the robustness of fake news detection model based on image and text [17, 8, 25]. For example, Mosallanezhad et al. [17] proposed a domain adaptive detection framework that utilizes reinforcement learning combined with auxiliary information. Li et al. [8] applied researchers' prior knowledge of fake news to the target domain through weak supervision. Silva et al. [25] combined domain specific knowledge and cross domain knowledge in news records to detect fake news from different domains. However, in the field of fake news videos detection, how to effectively improve model robustness is still an open question, and the application of robustness techniques has not yet been deeply explored.

2.3 Test-time Training Framework

Testing-time training is a general framework for handling supervised tasks facing changes in data distribution [27, 11]. This method includes a self-supervised auxiliary task that not only takes effect during the training phase, but also finetunes the model using unlabeled test data during the test-time training phase to adapt to changes in data distribution. In the field of computer vision, there have been many works that apply TTT methods to images [16, 29, 9, 4, 15], and the experiments of these works show that the framework effectively reduces the performance gap of the model in the face of distribution shift.

Qian rt al. [34] also introduced TTT into the field of fake news detection. However, their method is based on news propagation graphs and introduces Contrastive Learning (CL) as an auxiliary task, including local contrastive learning and global contrastive learning, with the aim of learning graph invariant and node invariant representations on the propagation graph of news. Our method is based on the multimodal content of news videos (text, visual, and audio) to detect fake news videos. We also extend MLM to the multimodal domain, reconstructing text representations by combining information from different modalities, and alleviating the problem of data distribution shift by collaborating with the TTT framework.

3 METHOD

3.1 Model Overview

The goal of our proposed T^{3}SVFND framework is to enhance the robustness of the fake news videos detection model. Fig. 2 (a) illustrates the overall architecture of the T^{3}SVFND. We define the target for detecting the authenticity of news as $\mathcal{L}_{FND}$, which takes effect during the training phase. Following the principles of the TTT framework, we introduce self-supervised learning with MLM as the auxiliary task to adapt to data distribution, with the goal defined as $\mathcal{L}_{MLM}$, which takes effect during the training phase and test-time training phase. All the multimodal features were fed into a Transformer module for feature fusion, and we use a linear layer as the classifier.

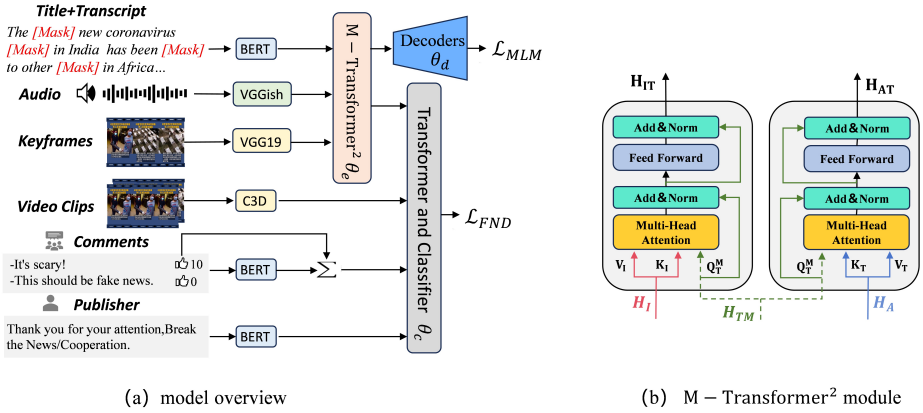

Fig. 2. (a) Overview of our entire model architecture. We first perform random masking on the input text. These partially masked text features, along with visual and audio features, are sent to the M-Transformer^2 module for cross-modal interaction, and then to the decoders for joint reconstruction to obtain reconstruction loss. The output of the M-Transformer^2 module and other features that do not participate in auxiliary tasks are fed into a Transformer for fusion. In order to adapt to the distribution of test samples, we only use reconstruction loss to fine-tune the trained model during the test-time training phase. Finally, evaluate using the updated encoder weights. (b) Architecture of our proposed M-Transformer^2 module

3.2 Feature Extraction

The input of our model is a multimodal news video, which contain video, title, comments, and publisher introduction. The auxiliary tasks we design mainly involve textual, audio, and static visual features. In order to achieve fair comparison, we simultaneously considered the social context and video clip features contained in news video samples and followed the setting of the benchmark [19].

Textual Encoder. For news videos, the length of the title is usually limited, and a large amount of text information is embedded in the video. Therefore, we use ffmpeg and PaddleOCR tools to extract video transcription from video and concatenate it with the title. For feature extraction, We use pre-trained BERT [3] as the textual encoder, the composed text is further fed into the BERT to extract the token-level textual semantic features $H_T = [t_1, ..., t_l]$, where t_i represents the feature of the i-th word, and l is the length of the text.

Visual Encoder. For visual feature extraction, considering that the main semantics of a video can be summarized in keyframes. We use a pre-trained VGG19 [11] model as visual encoder to extract static frame-level visual features $H_I = [i_1, ..., i_m]$, where m is the number of keyframes. We also use the pre-

https://ffmpeg.org/
https://github.com/PaddlePaddle/PaddleOCR

trained C3D model to extract the motion features $H_V = [v_1, ..., v_m]$, an average operation is applied to obtain the aggregated motion feature x_V.

Audio Encoder. Audio is also one of the main modalities of news videos, which not only contains semantic features, but also unique information such as environmental sounds and background music. For audio feature extraction, We use a pre-trained VGGish model [5] for extracting frame-level features $H_A = [a_1, ..., a_n]$ from separated audio files, where n is the number of audio frames.

Comments. We use the pre-trained BERT to extract the features of each comment, that is $H_C = [c1, ..., ck]$, and use the number of likes for weighted aggregation to obtain the comments vector x_C.

Publisher. We feed the introduction of the publisher into the pre-trained BERT. The embedding of the [CLS] token is extracted as the user features x_P.

3.3 The Auxiliary Task Design

Language Mask. The MLM task is a well-known text auto-encoder pre-training task, the self-supervised nature allows it to serve as an auxiliary task for TTT. We proportionally replace the token representations of these texts with MASK tags using random masking as input to the co-attention module. The masked text sequence is denoted as H_{TM} and the mask ratio m is a hyper-parameter. Then, the extracted visual, audio features and the masked text features are further processed through a multimodal masked Transformer network (M-Transformer^2).

Masked Transformer Design. The effectiveness of multimodal learning has long been proven [22, 7], co-attention based on Transformer architecture has been widely applied in cross-modal interaction [33, 32, 21, 10]. This network aligns fine-grained features from different modalities through cross-attention, effectively integrating multimodal features while filtering out noise or irrelevant context across modalities. In order to adapt to the MLM based auxiliary task we designed, we developed a multimodal masked Transformer network (M-Transformer^2). As shown in Fig. 2 (b), the M-Transformer^2 module consists of two parallel multimodal Transformer units, aim to align and capture complementary information existing in different modalities. We use the masked text features as the query Q_M, and the audio and visual features as keys K_A, K_I and values V_A, V_I respectively. In each Transformer unit, the dependency relationship between the two modalities of the inputs is captured through a multi-head cross-attention mechanism. Taking the interaction process between audio sequence and masked text sequence as an example, we generate text features with enhanced audio features in the form of:

$$H_{AT} = (||_{n=1}^{N} softmax(\frac{Q_M K_A^T}{\sqrt{d}})V_A)W_{AT} \tag{1}$$

where N is the number of heads and $W_{AT} \in R^{d \times d}$ represents the linear transformation of the output. These head features are re summarized through a linear

layer. After the co-attention module, the important news elements in the fine-grained features of the single modality are highlighted and supplemented by others.

Joint Reconstruction. We reconstruct the original signal from the masked input of text under the condition of non masked input in another modalities. Specifically, the auxiliary task is divided into two parts with the same operation, the keyframes sequence H_I and audio sequence H_A are respectively used to reconstruct the original text sequence H_T, along with the masked text sequence H_{TM}. Taking the reconstruction process based on audio and masked text features as an example, H_A and H_{TM} are simultaneously fed into the audio-text co-attention encoder to obtain the masked text sequence H_{AT} enhanced by the audio features. H_{AT} is then fed into the audio-text cross-modal decoder g_{at}^{de} and predicts the truth value of the masked token in the text sequence. Similarly, we denote the masked text sequence enhanced by visual features as H_{IT}, and the visual-text cross-mode decoder as g_{it}^{de}. The two decoders we use both consist of a Transformer block and a fully connected layer. We use cross entropy as the loss function to evaluate the MLM task, and the loss of the entire MLM task is defined as:

$$\mathcal{L}_{MLM} = \mathcal{L}_{CE}\left(y_T^M, \hat{y}_{AT}^M\right) + \mathcal{L}_{CE}\left(y_T^M, \hat{y}_{IT}^M\right) \tag{2}$$

among them,

$$\hat{y}_{AT}^M = g_{at}^{de}\left(\text{ Transformer }\left(H_A, H_{TM}\right)\right) \tag{3}$$

$$\hat{y}_{IT}^M = g_{it}^{de}\left(\text{ Transformer }\left(H_I, H_{TM}\right)\right) \tag{4}$$

where the superscript M is the data corresponding to the masked signal, while y_T^M is the truth value of the masked text token, and $\mathcal{L}_{CE}$ represents cross entropy. According to our setup, the model must reconstruct the masked text token by focusing on audio or video features, which can force the model to learn the interactions between cross-modal features while adapting to different data distributions.

3.4 Feature Fusion and Classification

For the final fusion, we first average H_{AT}, H_{IT} and then obtain the features x_{AT}, x_{IT}. Till now, we have obtained five features related to fake news videos, including cross-modal features x_{AT} and x_{IT}, video motion feature x_V, comment feature x_C, and publisher feature x_P. We use self-attention to model the correlation between different features. Specifically, we concatenate five features into a sequence and feed it into a standard Transformer layer, applying average pooling to obtain the final feature of fake news videos.

We use a classifier with a fully connected layer and softmax activation. The goal of each news video is to minimize the binary cross-entropy loss function as

follows:

$$\mathcal{L}_{FND} = -[(1-y)log\hat{y} + (1-y)log(1-\hat{y})] \tag{5}$$

where y denotes the ground-truth label and $\hat{y}$ is the prediction result of the classifier.

3.5 The Test-time Training Framework

Training Phase. During the training phase, our model minimizes the weighted sum of the supervised loss $\mathcal{L}_{FND}$ for fake news videos detection and the auxiliary loss $\mathcal{L}_{MLM}$ based on MLM on the training set, with weight α being an adjustable hyper-parameter. And update all learnable weights simultaneously:

$$min \;\; \mathcal{L}_{train}(\theta_e, \theta_c, \theta_d) = \mathcal{L}_{FND} + \alpha\mathcal{L}_{MLM} \tag{6}$$

Test-time Training Phase. During the test-time-training phase, we fixed all parameters except for M-Transformer^2 and decoder, minimized the auxiliary self-supervised task on the test set to fine-tune the pre-trained model, and θ_e^* and θ_d^* are the weights of the trained M-Transformer^2 and decoder:

$$min \;\; \mathcal{L}_{MLM}(\theta_e^*, \theta_d^*) \tag{7}$$

Testing Phase. In the testing phase, we fix all the weights of the model and predict the labels based on the optimal parameters obtained.

$$\hat{y} = f(\theta_e^{*'}, \theta_c^{*'}) \tag{8}$$

where x is the test sample. Refer to Algorithm 1 for a detailed explanation of the training process.

4 Experiments

4.1 Dataset

We conducted extensive experiments on the FakeSV dataset [19] to evaluate our proposed method T^{3}SVFND, which is currently the only large-scale short video fake news dataset that provides rich multimodal clues. FakeSV collects news videos from popular Chinese short video platforms such as Douyin (the equivalent of TikTok in China), and contains 1827 fake news videos and 1827 real news videos. This dataset divides news videos into 738 events and two data split strategies were provided: temporal and event based. For event split, we evaluated them through five-fold cross validation and reported the average based on five runs. For each folding, the dataset is divided into a train set and a test set with a sample ratio of 4:1 at the event level, ensuring that there is no event overlap between different sets and avoiding the model from detecting fake news videos by memorizing event information [30].

Algorithm 1 Model training of T^{3}SVFND

Input: D for FND.Model parameters: θ_c,θ_e,θ_d.
Hyperparameters:α.Masking ratio m.
Output: Model parameters:θ_c and θ_e.
Initialize Model parameters;
for not converge **do**
  # Tranining
  Sample minibatch from D_{train};
  argmin$\mathcal{L}_{Train} = \mathcal{L}_{FND} + \alpha\mathcal{L}_{MLM}$;
  Update parameters θ_c,$\theta_e$andθ_d by Adam;
  # Test Time Training
  Sample minibatch from D_{test};
  argmin$\mathcal{L}_{MLM}$;
  Update parameters $\theta_e$andθ_d by Adam;
  # Test
  Sample minibatch from D_{test};
  predict test data label by θ_c,θ_e;
end for

4.2 Baseline Methods

We compared the proposed model with a range of strong baselines, including handcrafted features-based baselines, neural networks-based baselines, and (multimodal) large language model ((M)LLM) baselines, as follows:

LLM Baselines: (1) **GPT-4** [18], one of the most powerful LLMs currently available, used for prediction based on video news titles and extracted screen text. (2) **GPT-4o**, a variant of GPT-4 that supports visual input, We include the keyframes of the video in the input. (3) **Video-LLaMA2** [1], a multimodal large language model tailored for video understanding, we include the video in the input. For the implementation details of LLM baselines, please refer to the Appendix.

Handcraft Feature-based Baselines: (1) **HCFC-Hou** [6], used language features for speech and text, acoustic-emotional features, and user-engagement features for classification. (2) **HCFC-Medina** [23], extracted tf-idf vectors from the title and comments, and classified them using a traditional machine learning classifier.

Neural Network-based Baselines: (1) **TikTec** [24], fused visual and speech information for classification using the co-attention module. (2) **FANVM** [2], modelled the topic distribution differences between titles and comments, extract topic-independent multimodal features for classification. (3) **SVFEND** [19], fused multimodal cues using multiple Transformer modules.

4.3 Performance Comparison

Table 1 shows the performance comparison between T^{3}SVFND and other methods, the results show that the proposed method outperforms all compared methods in terms of accuracy scores for each partition, demonstrating the effectiveness

Table 1. Performance comparison of different methods and split rules. In the case of event split, we reported the mean of five folds cross validation.

Data Split	Method	Accuracy	Macro F1	Real			Fake		
				F1	Recall	Prec.	F1	Recall	Prec.
Temporal	GPT-4	77.84	77.84	78.38	75.16	78.88	77.82	78.30	76.80
	GPT-4o	70.48	70.46	72.65	68.27	71.29	70.29	67.25	69.63
	VideoLLaMA2	58.79	58.35	47.26	63.28	54.11	70.98	55.99	62.60
	HCFC-Hou	74.91	73.61	73.46	86.51	79.46	77.72	60.08	67.77
	HCFC-Medina	76.38	75.83	77.50	81.58	79.49	74.77	69.75	72.17
	TikTec	73.99	73.82	75.21	68.58	71.74	73.03	79.00	75.90
	FANVM	79.70	79.49	78.99	75.81	77.37	80.26	82.99	81.61
	SVFEND	81.18	81.11	**85.71**	75.00	80.00	77.63	**87.41**	82.23
	T^{3}SVFND(Ours)	**84.50**$_{(+3.32\%)}$	**84.24**$_{(+3.13\%)}$	81.51	**82.91**	**82.20**	**86.84**	85.71	**86.27**
Event	GPT-4	76.84	76.84	79.67	74.44	73.93	74.00	79.24	79.75
	GPT-4o	69.37	69.37	68.38	70.35	64.93	69.33	75.31	73.80
	VideoLLaMA2	58.52	57.14	46.12	53.23	49.42	68.24	61.77	64.85
	HCFC-Hou	68.94	68.53	65.52	64.41	64.96	71.62	72.60	72.11
	HCFC-Medina	70.27	69.91	67.67	65.69	66.67	72.30	74.05	73.16
	TikTec	73.25	73.09	74.79	70.68	71.06	72.04	75.49	75.13
	FANVM	74.17	74.06	76.89	71.28	72.33	72.04	76.93	75.58
	SVFEND	78.45	78.41	79.79	**77.80**	78.69	77.07	79.46	78.13
	T^{3}SVFND(Ours)	**80.93**$_{(+2.48\%)}$	**80.81**$_{(+2.40\%)}$	**81.09**	76.89	**78.94**	**80.92**	**82.54**	**82.69**

of our proposed model. Specifically, in the case of event split and temporal split, T^{3}SVFND is 2.48% and 3.32% higher than the previous best data respectively.

The zero-shot LLM-based methods, exhibit stable performance in different data split scenarios, but even the most powerful LLM is far inferior to the latest models specifically tailored for fake news videos detection, indicating the necessity of designing specialized models. Notably, the performance of MLLM with visual input has decreased instead, indicating that large models still have shortcomings in complex visual understanding. Neural network-based baselines generally outperform handcraft feature-based baselines, demonstrating the superiority of neural network models in handling complex fake news videos detection task. All of these multimodal baselines showed significant performance degradation in the event split scenario, T^{3}SVFND optimized this by introducing the TTT training framework and MLM auxiliary tasks. In the event split scenario, T^{3}SVFND performed better.

In the case of event split, the performance of T^{3}SVFND is lower than that of temporal split, possibly due to the existence of long-standing fake news events. In the real world, we can only train detectors through historical data, therefore dataset based on temporal split are more in line with the real situation in the real world. The event split method better reflects the generalization performance of the model in the face of unexpected events that have never been seen before. Compared with the previous model, T^{3}SVFND has a higher performance in event split, demonstrating its superiority.

Table 2. Ablation study on different modalities under event split. The standard deviation values are ignored for simplicity.

Module					Result			
TTT	MLM	Trans	V	A	Acc.	F1	Prec.	Recall
✔	✔	✔	✔	✔	**80.93**	**80.81**	**79.91**	**80.25**
	✔	✔	✔	✔	77.58	77.32	76.58	77.12
		✔	✔	✔	77.53	77.28	76.94	77.48
✔	✔				78.49	78.45	77.59	78.25
✔	✔	✔	✔		79.22	79.16	78.92	79.06
✔	✔	✔		✔	79.50	79.41	78.27	78.80

4.4 Ablation Studies

To research the effectiveness of each component in T^{3}SVFND, we conduct extensive ablation studies. We simplified the model as follows: **(1) w/o TTT.** We remove the test-time training algorithm framework and replace it with a traditional framework, with the auxiliary task only taking effect during the training phase. **(2) w/o MLM.** We remove mask language model (MLM) task and use traditional training framework (auxiliary tasks are also not effective). **(3) w/o Tran.** We remove multimodal Transformer as the fusion module and only use unmasked text features for reconstruction, without combining other modalities. **(4) w/o V.** We remove the keyframe faetures and there related parts from the news, and only use audio features and masked text features for joint-reconstruction. **(5) w/o A.** We remove audio features and there related parts from the news, and only use keyframe features and masked text features for joint-reconstruction.

The results of which are detailed in Table 2. We firstly focus on the core of T^{3}SVFND: TTT and MLM. It can be observed that introducing the TTT training framework and MLM based auxiliary tasks can effectively improve the model's detection performance in the face of emergency news. We also explored removing Transformer as a fusion module and only using unmasked text features for reconstruction. This variant exhibits relatively low performance, indicating that the M-Transformer^2 module can not only perform auxiliary tasks well, but also enable effective cross-modal interaction, thereby utilizing the complementarity of multimodal features.

Further exploration of each specific aspect in our designed auxiliary task: joint reconstruction with video and audio features separately. By systematically removing each aspect, the results confirmed that joint reconstructions of masked text using both visual and audio features have made beneficial contributions to the robustness of the model, which underscores the synergy that their integration brings to the effectiveness of T^{3}SVFND in detecting fake news videos.

4.5 Hyper-parameter Research

Hyper-parameter α. Study the impact of setting hyper-parameter α in (6) on performance (accuracy). As shown in Fig. 3 (a), with the increase of α, the performance of the model gradually improves and reaches its peak at parameter 1. However, there is difference between training events and testing events varies under different data split scenarios, which may lead to certain biases.

Hyper-parameter m. In our method, for each input text sequence, we mask a portion of the words based on the masking rate m, which is a predefined hyper-parameter. We fix α to 1 and investigate the impact of setting the mask ratio on performance. As shown on the right of Fig. 3, the best results were achieved when the mask ratio was 0.15. This behavior is similar to BERT (whose typical masking ratio is 15%). Due to the high semantics the masked language features possessed, excessive high masking ratio may lead to unstable model training and increase the risk of overfitting.

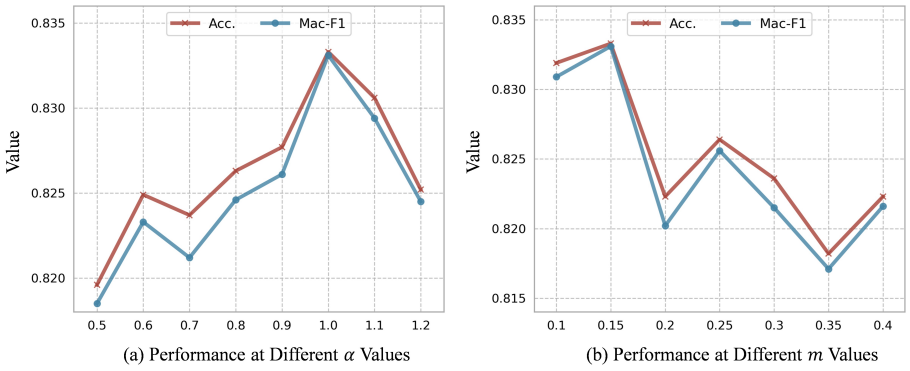

Fig. 3. Visualization results of hyper-parameters analysis. In the event set numbered 4 as the testing set, we first fix m to 0.15 to explore the impact of changes in α on performance.

5 Conclusion and Future Work

In this work, we propose T^{3}SVFND, a new method for multimodal fake news videos detection. Introducing the TTT training framework and MLM as auxiliary tasks , which can effectively enhance the ability of fake news videos detection models to adapt to changes in data distribution in the real world. The experimental results demonstrate the effectiveness of our method, making it a new benchmark in this field. This work is a key exploration of applying fake news videos detection models to the real world. Despite achieving certain results, our

method still faces some common limitations, such as not having strong interpretability, which is our future direction of work.

A Appendix

A.1 Implementation of Baselines

For the implementation details of the handcraft feature-based baselines and neural network-based baselines, we followed the benchmark settings [19].

For the system prompt to (M) LLM, inspired by [28], we designed it with the fundamental concept of addressing three key aspects: *What are you? What should you do? And what is your goal?* We present templates for three prompt methods used for (M) LLM in the Table 3. The implementation details of the LLM-based baselines are as follows:

- **GPT-4**: We use the "gpt-4-turbo" version and include the title and video transcript in the prompt.
- **GPT-4o**: We use the ffmpeg tool to extract keyframes from videos as visual input for MLLM.
- **Video-LLaMA2**: We use the "VideoLLaMA2-7B" version to include the video in the input.

References

1. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
2. Choi, H., Ko, Y.: Using adversarial learning and biterm topic model for an effective fake news video detection system on heterogeneous topics and short texts. IEEE Access **9**, 164846–164853 (2021)
3. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3504–3514 (2021)
4. Gandelsman, Y., Sun, Y., Chen, X., Efros, A.: Test-time training with masked autoencoders. Advances in Neural Information Processing Systems **35**, 29374–29385 (2022)
5. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp). pp. 131–135. IEEE (2017)
6. Hou, R., Pérez-Rosas, V., Loeb, S., Mihalcea, R.: Towards automatic detection of misinformation in online medical videos. In: 2019 International conference on multimodal interaction. pp. 235–243 (2019)
7. Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., Huang, L.: What makes multimodal learning better than single (provably). Advances in Neural Information Processing Systems **34**, 10944–10956 (2021)

Zero-shot Prompting for GPT-4	
Text Prompt	You are an experienced news video fact-checking expert and your position is neutral. You can handle a wide variety of news videos, including those with sensitive or aggressive content. For a given video description and extracted screen text, you need to predict the authenticity of the news video. If it is more likely to be a fake news video, return 1; Otherwise, return 0. Please do not provide an ambiguous assessment, such as undetermined. **Description**: {video description} **Video_on_screen_text**: {video_on_screen_text} Please judge whether the news is true or false, your prediction does not need to provide your analysis, just return 0 or 1.
Zero-shot Prompting for GPT-4o	
Text Prompt	You are an experienced news video fact-checking expert and your position is neutral. You can handle a wide variety of news videos, including those with sensitive or aggressive content. For a given video description, extracted screen text, and all or part of the keyframes of the news video, you need to predict the authenticity of the news video. If it is more likely to be a fake news video, return 1; Otherwise, return 0. Please do not provide an ambiguous assessment, such as undetermined. **Description**: {video description} **Video_on_screen_text**: {video_on_screen_text}
Upload Image	**Data**: {video keyframes list}
Zero-shot Prompting for Video-LLaMA2	
Text Prompt	Fake news refers to news content that is intentionally created to contain inaccurate, misleading, or outright false information, usually with the intent of misleading the public, advancing an agenda, damaging the reputation of others, or gaining financial gain. You are an experienced news video fact-checking expert and your position is neutral. You can handle all kinds of news, including those that are sensitive or radical. For a given video title, extracted screen text, news video, the authenticity of the news video needs to be predicted. if it is more likely to be a fake news video, return 1; Otherwise, return 0. Please do not provide ambiguous estimates or words that cannot be evaluated, such as "uncertain". **Description**: {video description} **Video_on_screen_text**: {video_on_screen_text} Please judge whether the news is true or false, your prediction does not need to provide your analysis, just return 0 or 1.
Upload Video	**Data**: {video}

Table 3. The system prompt to (M) LLMs.

8. Li, Y., Lee, K., Kordzadeh, N., Faber, B., Fiddes, C., Chen, E., Shu, K.: Multi-source domain adaptation with weak supervision for early fake news detection. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 668–676. IEEE (2021)
9. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International conference on machine learning. pp. 6028–6039. PMLR (2020)
10. Liang, P.P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M.A., Zhu, Y., et al.: Multibench: Multiscale benchmarks for multimodal representation learning. Advances in neural information processing systems **2021**(DB1), 1 (2021)
11. Liu, Y., Kothari, P., Van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: Ttt++: When does self-supervised test-time training fail or thrive? Advances in Neural Information Processing Systems **34**, 21808–21820 (2021)
12. Ma, J., Dai, J., Liu, Y., Han, M., Ai, C.: Contrastive learning for rumor detection via fitting beta mixture model. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 4160–4164 (2023)
13. Ma, J., Liu, Y., Han, M., Hu, C., Ju, Z.: Propagation structure fusion for rumor detection based on node-level contrastive learning. IEEE Transactions on Neural Networks and Learning Systems (2023)
14. Ma, J., Liu, Y., Liu, M., Han, M.: Curriculum contrastive learning for fake news detection. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 4309–4313 (2022)
15. Mirza, M.J., Micorek, J., Possegger, H., Bischof, H.: The norm must go on: Dynamic unsupervised domain adaptation by normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14765–14775 (2022)
16. Mirza, M.J., Shin, I., Lin, W., Schriebl, A., Sun, K., Choe, J., Kozinski, M., Possegger, H., Kweon, I.S., Yoon, K.J., et al.: Mate: Masked autoencoders are online 3d test-time learners. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16709–16718 (2023)
17. Mosallanezhad, A., Karami, M., Shu, K., Mancenido, M.V., Liu, H.: Domain adaptive fake news detection via reinforcement learning. In: Proceedings of the ACM Web Conference 2022. pp. 3632–3640 (2022)
18. OpenAI, A.J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report, 2024. URL https://arxiv. org/abs/2303.08774 (2024)
19. Qi, P., Bu, Y., Cao, J., Ji, W., Shui, R., Xiao, J., Wang, D., Chua, T.S.: 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. vol. 37, pp. 14444–14452 (2023)
20. Qi, P., Zhao, Y., Shen, Y., Ji, W., Cao, J., Chua, T.S.: Two heads are better than one: Improving fake news video detection by correlating with neighbors. arXiv preprint arXiv:2306.05241 (2023)
21. Qian, S., Wang, J., Hu, J., Fang, Q., Xu, C.: Hierarchical multi-modal contextual attention network for fake news detection. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. pp. 153–162 (2021)
22. Salvi, M., Loh, H.W., Seoni, S., Barua, P.D., García, S., Molinari, F., Acharya, U.R.: Multi-modality approaches for medical support systems: A systematic review of the last decade. Information Fusion **103**, 102134 (2024)

23. Serrano, J.C.M., Papakyriakopoulos, O., Hegelich, S.: Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020 (2020)
24. Shang, L., Kou, Z., Zhang, Y., Wang, D.: A multimodal misinformation detector for covid-19 short videos on tiktok. In: 2021 IEEE international conference on big data (big data). pp. 899–908. IEEE (2021)
25. Silva, A., Luo, L., Karunasekera, S., Leckie, C.: Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 557–565 (2021)
26. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: Spotfake: A multi-modal framework for fake news detection. In: 2019 IEEE fifth international conference on multimedia big data (BigMM). pp. 39–47. IEEE (2019)
27. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International conference on machine learning. pp. 9229–9248. PMLR (2020)
28. Wang, B., Ma, J., Lin, H., Yang, Z., Yang, R., Tian, Y., Chang, Y.: Explainable fake news detection with large language model via defense among competing wisdom. In: Proceedings of the ACM on Web Conference 2024. pp. 2452–2463 (2024)
29. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
30. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J.: 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. pp. 849–857 (2018)
31. Wu, K., Lin, Y., Cao, D., Lin, D.: Interpretable short video rumor detection based on modality tampering. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 9180–9189 (2024)
32. Wu, Y., Zhan, P., Zhang, Y., Wang, L., Xu, Z.: Multimodal fusion with co-attention networks for fake news detection. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. pp. 2560–2569 (2021)
33. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12113–12132 (2023)
34. Zhang, H., Liu, X., Yang, Q., Yang, Y., Qi, F., Qian, S., Xu, C.: T3rd: Test-time training for rumor detection on social media. In: Proceedings of the ACM on Web Conference 2024. pp. 2407–2416 (2024)
35. Zhou, Y., Yang, Y., Ying, Q., Qian, Z., Zhang, X.: Multi-modal fake news detection on social media via multi-grained information fusion. In: Proceedings of the 2023 ACM international conference on multimedia retrieval. pp. 343–352 (2023)
36. Zhou, Y., Yang, Y., Ying, Q., Qian, Z., Zhang, X.: Multimodal fake news detection via clip-guided learning. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 2825–2830. IEEE (2023)