TriSPrompt: A Hierarchical Soft Prompt Model for Multimodal Rumor Detection with Incomplete Modalities

Jiajun Chen[1], **Yangyang Wu**[2], **Xiaoye Miao**[1], **Mengying Zhu**[2], **Meng Xi**[2]
[1] Center for Data Science, Zhejiang University, Hangzhou, China
[2] School of Software Technology, Zhejiang University, Ningbo, China
{chenjjcccc, zjuwuyy, miaoxy, mengyingzhu, ximeng} @zju.edu.cn

Abstract

The widespread presence of incomplete modalities in multimodal data poses a significant challenge to achieving accurate rumor detection. Existing multimodal rumor detection methods primarily focus on learning joint modality representations from *complete* multimodal training data, rendering them ineffective in addressing the common occurrence of *missing modalities* in real-world scenarios. In this paper, we propose a hierarchical soft prompt model TriSPrompt, which integrates three types of prompts, *i.e.*, *modality-aware* (MA) prompt, *modality-missing* (MM) prompt, and *mutual-views* (MV) prompt, to effectively detect rumors in incomplete multimodal data. The MA prompt captures both heterogeneous information from specific modalities and homogeneous features from available data, aiding in modality recovery. The MM prompt models missing states in incomplete data, enhancing the model's adaptability to missing information. The MV prompt learns relationships between subjective (*i.e.*, text and image) and objective (*i.e.*, comments) perspectives, effectively detecting rumors. Extensive experiments on three real-world benchmarks demonstrate that TriSPrompt achieves an accuracy gain of over 13% compared to state-of-the-art methods. The codes and datasets are available at https://anonymous.4open.science/r/code-3E88.

1 Introduction

Social media has emerged as a primary platform for the rapid dissemination of information, where multimodal communication—integrating various forms of media—has surpassed traditional unimodal formats to become the dominant mode (Wang et al., 2024a). Compared to unimodal content, multimodal information conveys richer and more nuanced messages (Tandoc Jr et al., 2018), enhancing engagement but also introducing greater complexity, which poses significant challenges for effective rumor detection (Zou et al., 2024).

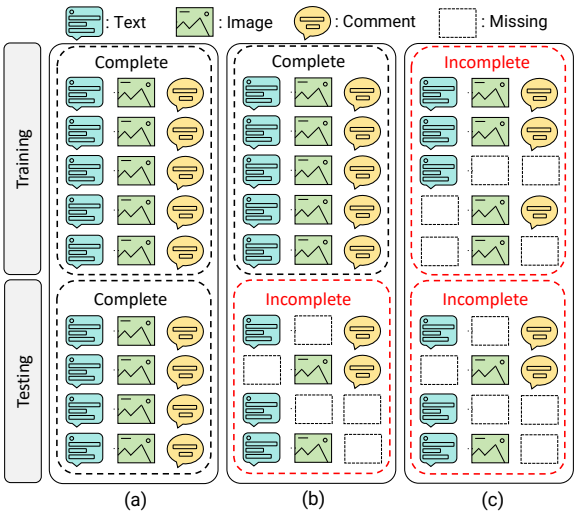


Figure 1: Multimodal rumor detection configurations. (a) *Cross-phase completeness alignment*: Full modalities in both phases (*ideal data integrity scenario*). (b) *Cross-phase completeness discrepancy*: Train with full modalities; test with incomplete modalities. (c) *Cross-phase incompleteness consistency*: Incomplete modalities in both phases (*real-world modality loss scenarios*).

Earlier research on rumor detection predominantly centered on text-based (unimodal) analysis (Rubin et al., 2015; Przybyla, 2020), leveraging features such as writing style and dissemination patterns to identify rumors. However, as multimodal content has become the dominant form of communication, the focus has shifted towards multimodal rumor detection. This approach integrates information from different modalities, such as text and images, to enhance detection accuracy. For instance, some studies employ feature fusion techniques to combine text and image data, aiming to improve performance. Yet, simple fusion methods often fall short of capturing the intricate relationships between modalities. To overcome this limitation, Wu et al. (2021) introduced a cross-attention mechanism to more effectively integrate features from each modality. Building on this, Zheng et al. (2022) utilized graph neural networks to incorporate con-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21687–21699
November 4-9, 2025

textual information from comments. Dhawan et al. (2024) proposed a fine-grained fusion approach that directly operates on the original modality information using graph-based methods, representing a significant advancement in multimodal rumor detection research. However, existing multimodal rumor detection methods often fail to address two critical practical challenges.

Firstly, *incomplete modalities (**CH1**).* In real-world scenarios, rumors spread rapidly, often outpacing the ability to collect complete multimodal data. For instance, during the COVID-19 pandemic in 2020, a deluge of false information and rumors circulated rapidly, leading to inevitable modality loss during data collection due to factors such as network instability, device limitations, or other uncontrollable circumstances. Despite this reality, most existing methods assume the availability of complete modalities. While CLKD-IMRD (Xu et al., 2023) partially addresses missing modalities using a distillation model, its reliance on complete training data significantly limits its generalizability. When faced with more common scenarios where both training and testing data are incomplete, as illustrated in Figure 1, the performance of such methods degrades considerably. Moreover, existing *Multimodal Large Language Models* (MLLMs), excelling in understanding multimodal content, struggle with rumor detection due to their inherent assumption that input information is "reality" and lack a mechanism for verifying content accuracy.

Secondly, *the nature of rumors (**CH2**).* In a post comprising text, images, and comments, each modality plays a distinct role. Text and images typically convey the subjective perspective of the publisher, while comments often reflect the objective viewpoints of reviewers or readers. Existing methods, however, tend to process and fuse all three modalities indiscriminately, overlooking the nuanced relationships between subjective and objective perspectives. This lack of differentiation not only prevents the model from effectively distinguishing between the two types of information but also introduces noise through direct fusion, which hinders judgment and degrades performance.

To address these challenges, we propose a novel *hierarchical soft prompt* model, TriSPrompt, consisting of three types of soft prompts: *modality-aware (MA)* prompt, *modality-missing (MM)* prompt, and *mutual-views (MV)* prompt, that effectively detects rumors in incomplete multimodal data. Specifically, *for addressing* **CH1**, in cases of missing modalities, the MA prompt learns both heterogeneous information from specific modalities and homogeneous features from available modality data, enabling the recovery of missing modalities. Moreover, the MM prompt integrates the missing states in incomplete multimodal data, enabling the model to better understand and adapt to the nature of modality missing information. *For addressing* **CH2**, the MV prompt reveals the potential relationships between the subjective perspectives (*i.e.*, text and image) provided by the publisher and the objective perspectives (*i.e.*, comment) contributed by other reviewers in the post, to effectively detect rumors. The main contributions of this paper can be summarized as follows:

- We propose a novel *hierarchical soft prompt* model, TriSPrompt, designed to effectively detect rumors in incomplete multimodal data. To the best of our knowledge, this is the first approach to tackle the general modality-missing problem with incomplete data in both training and testing phases in the field of multimodal rumor detection.

- In TriSPrompt, the MA prompt extracts features from available modalities to restore missing ones, while the MM prompt models missing states to enhance adaptability to incomplete information, effectively addressing the issue of information loss caused by modality absence, thus solving the problem of information loss caused by incomplete modalities.

- The MV prompt of TriSPrompt uncovers latent relationships between the subjective perspectives (*i.e.*, text and image) provided by the publisher and the objective perspectives (*i.e.*, comments) from other reviewers, facilitating efficient rumor detection.

- Extensive experiments on three real-world benchmark datasets demonstrate the effectiveness of TriSPrompt in distinguishing rumors, outperforming existing methods.

2 Related Work

2.1 Unimodal Rumor Detection

Research on unimodal rumor detection has primarily focused on image-based ones and text-based ones. Image-based rumor detection methods (Qi et al., 2019; Jin et al., 2016; Li et al., 2015; Gupta

et al., 2013) focus on evaluating the logical consistency of image content and detecting signs of modifications, such as artificial splicing. These methods aim to detect whether image content is logically coherent and free from manipulations that suggest tampering or forgery. Text-based methods (Ma et al., 2019; De Sarkar et al., 2018; Chen et al., 2018) focus on detecting rumors by analyzing linguistic features, emotions, and writing styles (Rubin et al., 2015; Przybyla, 2020), aiming to classify content as rumor or non-rumor. With the evolution of social media, metadata has become an important feature in rumor detection. Metadata, such as the identity of the poster, dissemination pathways, and engagement metrics like likes and shares, has significantly enhanced the ability of models to detect rumors (Ma et al., 2017; Lao et al., 2021). This approach exploits the wealth of contextual data available on social platforms to enhance the performance of rumor detection models. Nan et al. (2021) leverages a mixture-of-experts approach to extract modality-specific representations and integrates them using domain-specific gates for identifying fake news. In parallel, Zhu et al. (2022) introduces a domain adapter combined with a memory bank to effectively address the challenges posed by domain shifts and incomplete domain labeling, enhancing the robustness of fake news detection models across varying contexts.

2.2 Multimodal Rumor Detection

Multimodal information dissemination has become the dominant form of communication, resulting in increased research interest in multimodal rumor detection. Early studies often used a simplistic approach, concatenating text and image features (Singhal et al., 2019), without fully leveraging the potential of multimodal data. Qi et al. (2021) introduces a similarity-based model for fake news detection, which computes the similarity between multimodal and cross-modal features. In addition, Wu et al. (2021) advances feature integration across modalities by employing cross-attention mechanisms, improving the fusion of multimodal data. Chen et al. (2022) solves the cross-modal ambiguity learning problem by quantifying the ambiguity between different unimodal features using the distribution divergence. Zheng et al. (2022) presents a GAT-based model that integrates text, image content, and comment graphs for rumor detection. Recently, Dhawan et al. (2024) proposed GAME-ON, a novel end-to-end trainable GNN-based framework that allows for granular interaction modalities to fuse them early in the framework. Despite these advances, existing methods often overlook the unique properties of rumors; Text and images typically convey the subjective perspective of the publisher, while comments often reflect the objective viewpoints of reviewers or readers. Current approaches treat these three modalities as a single perspective, neglecting the intrinsic relationships between perspectives and potentially leading to misleading model judgments.

Given the rapid spread of rumors and the common occurrence of partial modality loss during data collection, Xu et al. (2023) address the issue of incomplete modalities with a novel framework that leverages contrastive learning and knowledge distillation, filling a gap in the field. However, the reliance on a teacher-student distillation structure is limited due to its dependency on complete training data and structural inflexibility. To overcome these challenges, we propose an end-to-end model with a hierarchical soft prompt architecture that simultaneously accounts for both the dual-perspectives nature of rumors and the challenges associated with incomplete training and test data.

3 Methodology

3.1 Problem Formulation

Let's define $P = \{p_1, \cdots, p_n\}$ as a set of posts, where each post p consists of three modalities: $\{T, I, C\}$ where T donates a source text, I represents an image, and C refers to a comment. In complete scenarios, all modalities are observed and easily integrated to support rumor detection. However, in real-world settings, some modalities may be missing, requiring their recovery for effective fusion. Considering the three modalities mentioned, there are seven different missing modality cases, as shown in Appendix A. For simplicity, we introduce an indicator $\epsilon = \{\epsilon_1, \epsilon_2, \epsilon_3\}$ for each post p, where $\epsilon_j = 0, j \in \{1, 2, 3\}$ signifies the missing of the j-th modality in the sample, and a non-zero value signifies its availability. This definition is consistent across both training and test sets.

Definition 1 Rumor Detection. Given a multimodal post p with a missing-modality indicator ϵ, the goal of rumor detection is to classify posts with incomplete modalities as either rumors or non-rumors; that is, the aim is to learn a function f_θ such that $f_\theta(p) \rightarrow \hat{y}$, where p represents a given multimodal post, and $\hat{y}$ is the outcome of the de-

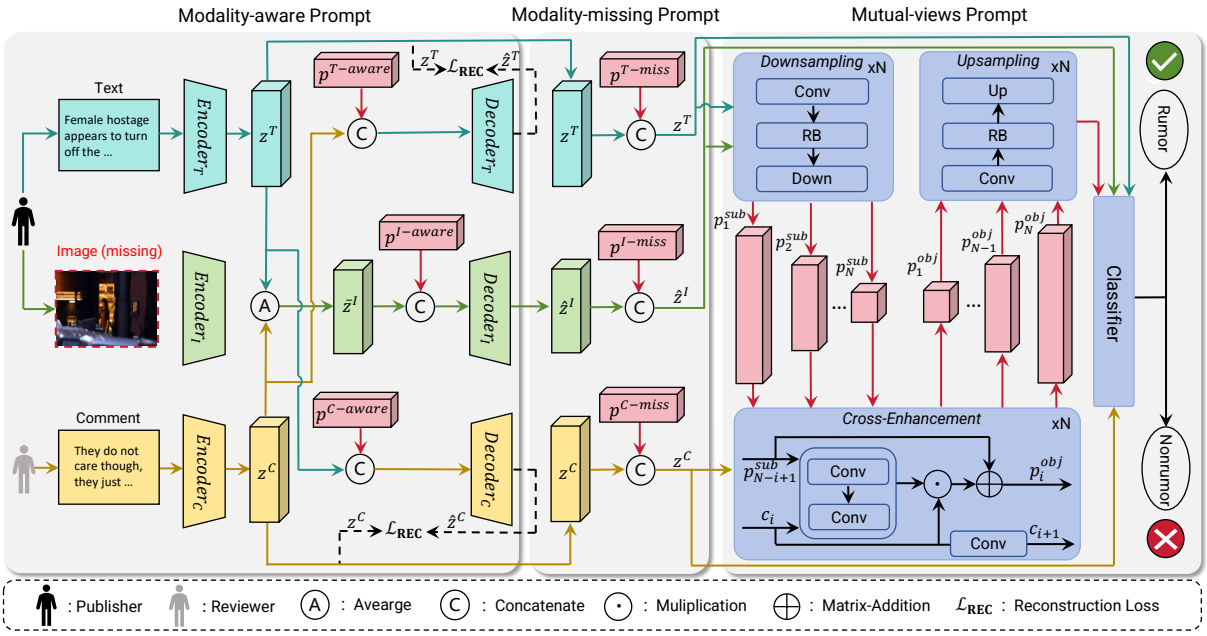


Figure 2: The architecture of TriSPrompt. Given the input incomplete multimodal data (we assume "Image" modality is missing). It consists of modality-aware prompt, modality-missing prompt, mutual-views prompt, Conv (*Convolutional-Layer*), RB (*Residual-Block*), Down (*Downsampling-Layer*), and Up (*Upsampling-Layer*).

tection. Here, $\hat{y} = 1$ denotes a rumor, and $\hat{y} = 0$ denotes a non-rumor.

3.2 Network Overview

Our proposed TriSPrompt framework, as depicted in Figure 2, consists of three types of soft prompts: *modality-aware* (MA) prompt, *modality-missing* (MM) prompt, and *mutual-views* (MV) prompt.

Specifically, in TriSPrompt, all available modalities are first encoded into a latent space, from which features are extracted via sampling. To account for the heterogeneity across modalities, we propose the MA prompt to learn the unique characteristics of each modality. For missing modalities, the MA prompt employs the available modality features to reconstruct ones. This ensures that each post is represented by a complete set $\{T, I, C\}$. To enhance the model's awareness of missing modalities, the MM prompt indicates which modality is missing and has been reconstructed. Subsequently, the MV prompt leverages all available modalities for comprehensive fusion, enhancing the rumor detection classification process. Further details are provided in subsequent sections. For simplicity, but without loss of generality, we assume the absence of the "Image" modality, *i.e.*, $p = \{T, I(missing), C\}$ as illustrated in Figure 2.

3.3 Modality-Aware Prompt

The interplay between modality-specific heterogeneity and homogeneity has long been a key focus in multimodal research. Ideally, models should extract homogenous features across different modalities to reduce redundancy and introduce heterogeneous features for complementarity, thus fully leveraging the advantages of multimodal data. When modalities are missing, restoring the missing modality often relies on the homogeneity of available modalities, potentially overlooking their heterogeneity. In light of these challenges, we propose the *modality-aware* (MA) prompt, which is designed to capture the heterogeneous features of each modality, thereby improving the accuracy of missing modality restoration.

Specifically, in the MA prompt, the available modalities T and C are first encoded into a latent space and obtain a representation z^T and z^C, *i.e.*,

$$z^T \sim N(\mu^T, \sigma^T) = Encoder_T(T), \quad (1)$$

$$z^C \sim N(\mu^C, \sigma^C) = Encoder_C(C), \quad (2)$$

where $Encoder_T$ and $Encoder_C$ refer to a generic encoder architecture, potentially composed of multiple layers of Transformers for each modality. μ^T and μ^C represent the mean values, while σ^T and σ^C represent the variances. We adopt the *variational autoencoder* (VAE) paradigm (Kingma, 2013), constraining the encoding results to follow

a normal distribution. This constraint leads to the *Kullback–Leibler* (KL) divergence loss $\mathcal{L}_{\mathbf{KL}}$, which quantifies the gap between the latent representation and the normal distribution, *i.e.*,

$$\begin{aligned}\mathcal{L}_{\mathbf{KL}} =&[\mathrm{KL}(N(\mu^T,\sigma^T) \parallel N(0,1))+\\ &\mathrm{KL}(N(\mu^C,\sigma^C) \parallel N(0,1))]/2,\end{aligned}$$

For the missing modality I, we utilize the available modality features, namely z^T and z^C, for recovery. First, we obtain the modality homogenous feature $\overline{z}^I$ by averaging. For each modality, there exists a prompt designed to learn heterogeneous features, denoted here as $p^{x-aware} \in \mathbb{R}^{1\times L^{aware}}$ with $x \in \{T, I, C\}$, where L^{aware} donates the length of $p^{x-aware}$. We combine the homogeneous feature $\overline{z}^I$ with $p^{I-aware}$ and decode it to restore the missing modality feature $\hat{z}^I$, *i.e.*,

$$\overline{z}^I = [z^T + z^C]/2, \tag{3}$$

$$\hat{z}^I = Decoder_I(\overline{z}^I, p^{I-aware}), \tag{4}$$

where $Decoder_I$ denotes a generic decoder architecture, similar to $Encoder$. Due to the incomplete nature of the training data, direct supervision of the reconstruction is not feasible. Thus, we reconstruct the available modalities and apply constraints to ensure consistency, *i.e.*,

$$\hat{z}^T = Decoder_T(z^C, p^{T-aware}), \tag{5}$$

$$\hat{z}^C = Decoder_C(z^T, p^{C-aware}), \tag{6}$$

Finally, we introduce a *multimodal reconstruction* loss function $\mathcal{L}_{\mathbf{REC}}$ that uses *mean squared error* (MSE) to quantify the distance between the original features (z^T and z^C) and their reconstructed counterparts ($\hat{z}^T$ and $\hat{z}^C$), *i.e.*,

$$\mathcal{L}_{\mathbf{REC}} = \left[MSE(\hat{z}^T, z^T) + MSE(\hat{z}^C, z^C)\right]/2.$$

3.4 Modality-Missing Prompt

Using the MA prompt, we obtain a complete multimodal representation $z = \{z^T, \hat{z}^I, z^C\}$, where z^T and z^C are the features from available modalities, and $\hat{z}^I$ is the feature from the recovered modality. To enhance the model's ability to discern between original and recovered modalities, we introduce the *modality-missing* (MM) prompt as indicators. The detailed procedure is as follows:

$$z^T = Concat(z^T, p^{T-miss}), \tag{7}$$

$$\hat{z}^I = Concat(\hat{z}^I, p^{I-miss}), \tag{8}$$

$$z^C = Concat(z^C, p^{C-miss}), \tag{9}$$

where $p^{x-miss} \in \mathbb{R}^{1\times L^{miss}}$ with $x \in \{T, I, C\}$ refers to the MM prompt. L^{miss} donates the length of p^{x-miss}. By utilizing this prompt, we enhance the representation of modality features by incorporating the missing states of incomplete multimodal data, enabling the model to more effectively recognize and adapt to modality absence. This improves the model's robustness in handling incomplete data.

3.5 Mutual-Views Prompt

Traditional multimodal representations often overlook the distinct characteristics of rumors by applying overly simplistic or uniform fusion strategies to text, image, and comment features. A post typically comprises three components: text, image, and comment, each performing distinct functions. The text and image typically represent the publisher's subjective view, reflecting the content creator's perspective, while the comment represents the objective views of other reviewers interacting with the post. The latent relationship between these two perspectives is crucial for rumor detection.

To this end, we propose the *Mutual-Views* (MV) prompt to uncover latent relationships between dual perspectives in rumor propagation, where the text and image represent the publisher's subjective view of the post, and comments reflect the objective views of other reviewers.

Architecture. The MV prompt mainly consists of three modules, *i.e.*, a *downsampling* module, a *cross-enhancement* module, and a *upsampling* module. Specifically, the *downsampling* module integrates the subjective view information from both the text and the image. First, it employs a *cross-attention* mechanism (Lu et al., 2019) to fuse these modalities, capturing the interdependencies between textual and visual features. This integration is formulated as:

$$p_1^{\text{sub}} = \text{CrossAttn}\left(z^T, \hat{z}^I\right), \tag{10}$$

where z^T and $\hat{z}^I$ represent the text embedding and image embedding, respectively. Subsequently, each network layer of the N-layer *downsampling* module comprises a convolutional layer, a residual block, and a downsampling layer, which refine features while progressively reducing dimensionality. For the i-th network layer F_i^{down}, the specific calculation process is $p_{i+1}^{sub} = F_i^{down}(p_i^{sub})$, where $i = 1, \cdots, N$. p_i^{sub} denotes the output of the i-th layer. This iterative process effectively distills the fused subjective-view information, enhancing the representation for subsequent analysis.

To generate the *objective-view* prompt, we use an N-layer *cross-enhancement* module that hierarchically integrates the subjective-view prompt with comment embedding, with each layer computed as:

$$w_i = Conv(p^{sub}_{N-i+1}, c_i), \tag{11}$$

$$p^{obj}_i = p^{sub}_{N-i+1} + c_i \cdot w_i, \tag{12}$$

$$c_{i+1} = Conv(c_i), \tag{13}$$

where p^{obj}_i denotes the i-th enhanced objective-view prompt. $Conv$ denotes the convolution operation. c_i denotes the comment embedding. By retaining intermediate results at each layer, we obtain a sequence of enhanced objective-view prompts, *i.e.*, $p^{obj} = \{p^{obj}_1, \cdots, p^{obj}_N\}$, where $c_1 = z^C$. After generating the subjective-view and objective-view prompts, we perform fine-grained fusion to integrate these perspectives, enabling comprehensive modeling of rumor characteristics.

The fused representations are subsequently processed through N *upsampling* modules. Each network layer comprises a convolutional layer, a residual block, and an upsampling layer. For the i-th network layer F^{up}_i,

$$p^{mutual}_{i+1} = F^{up}_i(p^{obj}_i, p^{mutual}_i), \quad i = 1, \cdots, N,$$

initialized with $p^{mutual}_1 = p^{sub}_N$. This refinement cascade produces the integrated prompt: $p^{mutual} \in \mathbb{R}^{1\times L^{mutual}} = p^{mutual}_N$, where L^{mutual} denotes the dimensionality of the final prompt. The classification head processes the concatenated features: $[p^{mutual}, z^T, \hat{z}^I, z^C]$ through fully-connected layers to classify the post as either a rumor ($\hat{y} = 1$) or non-rumor ($\hat{y} = 0$).

Loss Function. In general, we introduce a rumor detection loss function $\mathcal{L}_{\textbf{CLS}}$ that uses *cross-entropy* (CE) loss for binary classification to train the model, which helps optimize the parameters by minimizing the difference between the predicted $\hat{y}$ and the true label y, *i.e.*,

$$\mathcal{L}_{\textbf{CLS}} = CE(\hat{y}, y),$$

where $\hat{y} = Classifier(z^T, \hat{z}^I, z^C, p^{mutual})$. The $Classifier$ is a multimodal rumor detection classifier composed of multi-layer perceptions.

3.6 Objective Function

The overall objective function $\mathcal{L}_{\textbf{Total}}$ of our model TriSPrompt contains three types of losses, including the rumor detection loss $\mathcal{L}_{\textbf{CLS}}$, the KL divergence loss $\mathcal{L}_{\textbf{KL}}$ and the multimodal reconstruction loss $\mathcal{L}_{\textbf{REC}}$, *i.e.*,

$$\mathcal{L}_{\textbf{Total}} = \mathcal{L}_{\textbf{CLS}} + \lambda_1 \cdot \mathcal{L}_{\textbf{KL}} + \lambda_2 \cdot \mathcal{L}_{\textbf{REC}},$$

where λ_1 and λ_2 are trade-off hyperparameters.

4 Experiments

Datasets. In our experiments, we utilize three multimodal rumor detection datasets: the Chinese datasets *Weibo-19* (Song et al., 2019) and *Weibo-17* (Jin et al., 2017), along with the English dataset *Pheme* (Zubiaga et al., 2017). Details are in Appendix B. Each dataset comprises source text and images; however, *Weibo-19* and *Pheme* include comments, whereas *Weibo-17* does not. For TriSPrompt, we directly use p^{sub}_1 obtained from text and image to replace the MV prompt on *Weibo-17* dataset.

Metrics. To evaluate and compare our framework with others, we employ standard metrics for binary classification, including Accuracy, F1 score, and AUC. For feature extraction, we utilize a pre-trained BERT model (Kenton and Toutanova, 2019) to process text and comment information, generating a 768-dimensional hidden state as the feature representation. Similarly, for image data, we use a pre-trained ResNet-34 model (He et al., 2016), which encodes the visual information into a 512-dimensional hidden state representation.

Baselines. We compare TriSPrompt with three state-of-the-art *Multimodal Rumor Detection* (MRD) approaches: CAFE (Chen et al., 2022), CLKD (Xu et al., 2023), and Game-On (Dhawan et al., 2024); two *Incomplete Modality Learning* (IML) methods: Dicmor (Wang et al., 2023) and RedCore (Sun et al., 2024); and two *Multimodal Large Language Models* (MLLMs): Qwen2.5-VL-72B (Bai et al., 2025) and GPT-4o (Hurst et al., 2024).

Implementation details. To simulate real-world modality dropout, we implement a random missing protocol (Wang et al., 2024b), where each complete post is subject to the random absence of one or two modalities. This dropout mimics the incomplete information often present in real-world social media posts, making the model more robust to missing data. We denote the global missing rate as R_m to estimate the prevalence of such absences. The R_m is defined as $R_m = 1 - \frac{1}{L\times M}\sum_{i=1}^{L} a_i$, where a_i denotes the number of available modalities for the i-th sample, L denotes the total number of samples, and M indicates the number of modalities. We

Methods		*Weibo-19*			*Pheme*			*Weibo-17*		
		ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
MRD	CAFE	67.89±0.83	44.89±2.65	0.673±0.018	61.65±0.14	60.16±3.15	0.599±0.022	65.02±1.39	65.51±5.02	0.622±0.036
	CLKD	69.19±1.64	30.68±4.85	0.574±0.048	64.18±0.84	67.31±1.63	0.616±0.016	63.26±0.70	62.10±1.55	0.574±0.028
	GameOn	76.07±0.62	56.98±1.99	0.747±0.062	66.22±0.43	71.65±0.49	0.702±0.011	66.78±1.44	64.81±6.38	0.739±0.008
IML	Dicmor	77.48±0.85	68.73±1.88	0.823±0.021	73.17±0.72	76.10±0.97	0.776±0.008	72.58±0.80	72.48±2.41	0.789±0.006
	RedCore	80.05±0.73	70.82±2.45	0.846±0.026	71.58±0.12	75.65±0.40	0.745±0.002	73.68±0.57	73.62±0.13	0.816±0.011
MLLM	Qwen-VL	61.23±0.67	38.46±2.12	0.597±0.011	50.83±0.25	23.59±0.69	0.516±0.002	50.47±0.23	21.48±2.88	0.504±0.017
	GPT-4o	64.61±0.11	41.13±1.54	0.642±0.001	47.98±0.36	22.29±0.59	0.501±0.004	56.47±0.13	24.21±1.12	0.556±0.007
TriSPrompt		**81.51±0.59**	**73.38±0.81**	**0.867±0.023**	**75.82±0.44**	**77.33±0.69**	**0.788±0.015**	**75.15±0.24**	**74.17±1.67**	**0.836±0.016**

Table 1: Evaluation results on three datasets at $R_m = 0.5$, including three metrics: ACC, F1, and AUC.

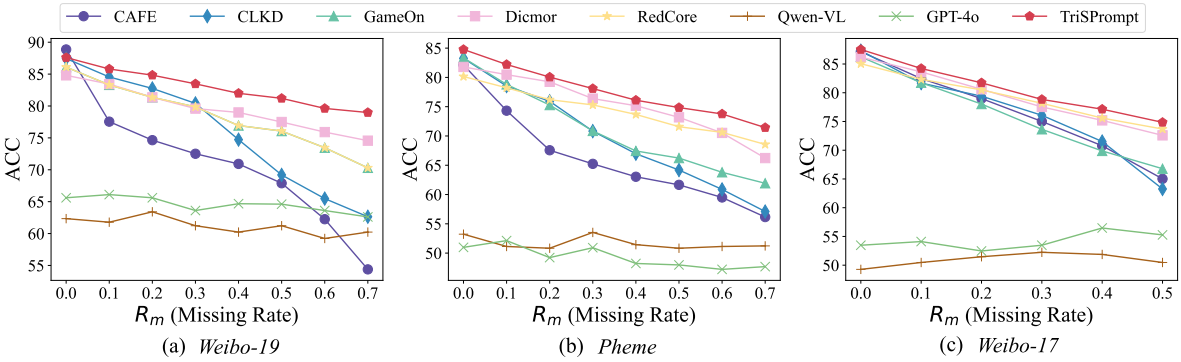


Figure 3: Performance demonstration under different missing rates.

also ensure that at least one modality is available for each sample, so $a_i \geq 1$ and $R_m \leq \frac{M-1}{M}$. The details are shown in Appendix C. For three modalities $\{T, I, C\}$ datasets *Weibo-19* and *Pheme* (text, image, and comment), we choose the R_m from $[0.0, \cdots, 0.7]$, where 0.7 is an approximation of $\frac{M-1}{M}$ with the same meaning. For two modalities $\{T, V\}$ dataset *Weibo-17* (text, image), which only involves the scenario of one modality being missing, we choose the R_m from $[0.0, \ldots, 0.5]$. We use Adam optimizer with $\beta_1 = 0.99$ and $\beta_2 = 0.999$. The learning rate is set to 0.002. The training batch size is set to 64. Number of network layers is set to 3. The hyperparameters $\lambda_1 = 0.1$ and $\lambda_2 = 0.001$. For fairness, all results are averaged over five times.

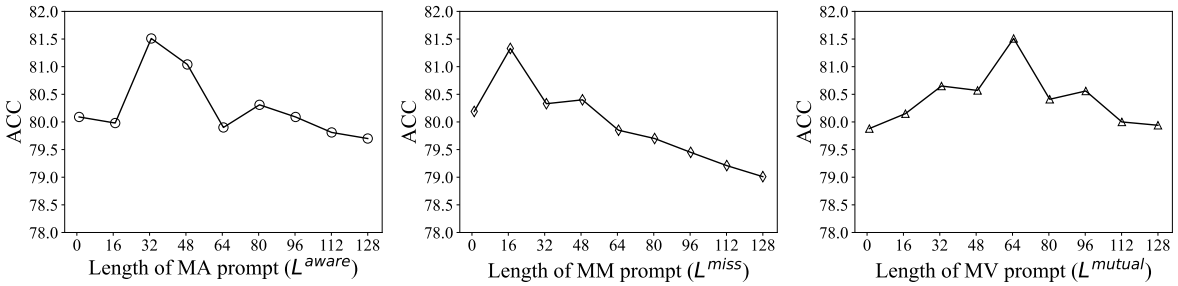
4.1 Comparison Study

Table 1 presents the accuracy of various rumor detection methods on three publicly real-world datasets under a global missing rate of $R_m = 0.5$, indicating that half of the modalities are missing.

One can observe that, TriSPrompt significantly outperforms all of the baselines. In terms of prediction accuracy (*i.e.*, ACC, F1, and AUC), TriSPrompt exceeds the best-performing baseline, RedCore, by an average of 2.99%, and even increases up to 5.92% on the *Pheme* dataset w.r.t. Accuracy. In particular, compared with MRD based methods, TriSPrompt demonstrates significantly superior performance. It outperforms the best MRD based method, GameOn, by an average of 14.08%, and even increases up to 28.78% on the *Weibo-19* dataset w.r.t. F1 score. This is because TriSPrompt fully leverages the available modalities, integrating heterogeneous and homogeneous features to effectively reconstruct missing data. We provide detailed experimental verification in Appendix D. to support this claim. It also perceives and adapts to the modality-missing mechanism while deeply exploring both subjective and objective perspectives in rumors. Although MLLMs exhibit strong capabilities in cross-modal understanding, their performance on rumor detection remains suboptimal compared to traditional methods. This primarily stems from the fact that MLLMs commonly treat input as an intrinsic representation of "reality", without incorporating dedicated mechanisms to evaluate content veracity—a crucial component for robust rumor detection.

4.2 Parameter Evaluation

Effect of missing rate. The experimental results for varying the missing rate R_m from 0% to 70% are shown in Figure 3. We can find that, with the



(a) MA Prompt (b) MM Prompt (c) MV Prompt

Figure 4: Ablation study on different lengths of prompts on the *Weibo-19* dataset.

growth of the missing rate, the prediction accuracy (*i.e.*, ACC) of each algorithm descends consistently. It is attributed to the less data information for detection when the missing rate turns high. Among all the algorithms, TriSPrompt performs the best in each case. Notably, across different missing rates, the average performance degradation ratio of TriSPrompt is just 6.14%, far lower than the best-performing baseline RedCore's 8.32%. In other words, TriSPrompt is more robust with the increasing missing rate than others. The reason behind this is that, TriSPrompt learns the data distribution from all observed multimodal data conditional on the feature missing state, so as to decrease or delay the impact of increased missing rate on the forecasting performance. In real-world rumor outbreaks, severe modality absence often renders existing detection algorithms ineffective, making our TriSPrompt approach crucial for multimodal rumor detection. Despite their flexibility in handling arbitrary modality inputs, including unimodal and multimodal content, MLLMs still fall short in achieving strong performance on rumor detection.

Effect of prompt length. With the prompt lengths (*i.e.*, L^{aware}, L^{miss}, L^{mutual}) corresponding to the MA prompt, MM prompt, and MV prompt, varying from 1 to 128, Figure 4 illustrates the detection accuracy (*i.e.*, ACC) of TriSPrompt on the *Weibo-19* dataset. The experimental results demonstrate that model performance generally improves with increased prompt length, up to a certain threshold. Beyond this point, further extension leads to a decline in accuracy, indicating that excessively long prompt may introduces noise or dilute the importance of input features. This suggests the necessity of selecting an appropriate length. Specifically, TriSPrompt is the best, in terms of the larger Accuracy, when L^{aware} is 32, L^{miss} is 16, and L^{mutual} is 64. It also confirms that, the three prompts of TriSPrompt benefit the detection accuracy.

Datasets	no-MA	no-MM	no-MV	no-MAM	TriSPrompt
Weibo-19	80.14	80.09	79.88	79.17	**81.51**
Pheme	73.84	73.17	72.77	71.85	**75.82**
Weibo-17	73.67	73.01	72.12	70.69	**75.15**

Table 2: Ablation study on the hierarchical soft prompts at R_m = 0.5. Quality metrics: ACC. **Bold** indicates the best performance.

4.3 Ablation Study

We investigate the influence of different elements of TriSPrompt on the rumor detection performance. The corresponding experimental results over the three datasets, including ACC, are presented in Table 2. no-MA is the variant of TriSPrompt without the MA prompt. no-MM is the variant of TriSPrompt without the MM prompt. no-MV is the variant of TriSPrompt without the MV prompt. no-MAM is the variant of TriSPrompt without both the MA and MM prompts.

We observe that each of the three prompts in TriSPrompt — the MA prompt, MM prompt, and MV prompt — positively impacts detection performance. When the three prompts are removed respectively, the detection accuracy (measured by ACC) of TriSPrompt decreases on average by 2.09%, 2.70%, and 3.35%. In addition, when both the MA prompt and MM prompt are removed simultaneously, the model's performance experiences the largest decline, with an average decrease of 4.68%. These findings highlight the indispensability of all three prompts in TriSPrompt, demonstrating their combined significance in achieving effective performance.

4.4 Analysis of Effectiveness and Efficiency

To rigorously assess the effectiveness and efficiency of TriSPrompt, we systematically compared TriSPrompt with representative methods

Methods	Params (M)	Training Time/Epoch (ms)	Inference Time/Batch (ms)	Max GPU Memory (GB)
CAFE	2.95	1996	910	0.98
CLKD	2.81	1411	1260	**0.84**
GameOn	**1.02**	**606**	**271**	1.60
Dicmor	3.46	4511	1833	3.26
RedCore	7.72	1923	822	2.41
Qwen-VL	–	–	1455	–
GPT-4o	–	–	1623	–
TriSPrompt	2.41	2032	931	1.28

Table 3: Resource consumption and computational efficiency of different methods.

from three major domains (MRD, IML, and MLLM) on the *Pheme* dataset, covering model size, training/inference efficiency, memory usage, and actual detection performance presented in Table 3. All experimental results we obtained under the same experimental environment, averaged over multiple independent runs to ensure fairness and reliability of the comparisons. Specifically, in the MRD domain, methods such as CAFE, CLKD, and GameOn mainly focus on multimodal feature fusion but have limited adaptability to missing modalities. On the Pheme dataset, TriSPrompt achieves an ACC of 75.82%, which is 9.6 percentage points higher than GameOn (66.22%). Meanwhile, TriSPrompt's inference time and resource consumption are comparable to these baselines. In the IML domain, methods such as Dicmor and RedCore focus on scenarios with missing modalities. TriSPrompt outperforms RedCore by 4.24% in ACC (75.82% vs. 71.58%), and its performance degrades more slowly under high missing rates (average drop of 6.14%, compared to RedCore's 8.32%), demonstrating stronger practical robustness. In the MLLM domain, models like Qwen-VL and GPT-4o possess strong cross-modal understanding but perform poorly on rumor detection tasks. In contrast, TriSPrompt not only achieves higher detection accuracy but also consumes significantly fewer resources (e.g., inference time is only 1/17 that of GPT-4o).

5 Conclusion

In this paper, we innovatively propose a hierarchical soft prompt framework, namely TriSPrompt, targeting the more prevalent modality absence situation in multimodal rumor detection, where both the training and testing data suffer from modality missing. It consists of the MA prompt, MM prompt, and MV prompt. The MA prompt extracts both heterogeneous and homogeneous features from available modalities to recover missing data, while the MM prompt models missing states to improve adaptability to incomplete information. The MV prompt connects subjective and objective perspectives, enhancing rumor detection. The effectiveness of TriSPrompt has been verified on three real-world datasets. Particularly, TriSPrompt exhibits excellent robustness under the circumstances of severe modality absence.

Limitations

Our proposed method TriSPrompt focuses on the provided real-world multimodal rumor detection datasets, enabling controlled evaluation and comparison with existing approaches. However, it falls short in addressing the complexities and variability of real-world multimodal data. This limitation underscores the need for more robust models capable of interacting effectively with real-world scenarios. Beyond this work, we believe some promising future works with large language models would solve this problem.

Acknowledgments

This work was supported in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01212), and the Ningbo Yongjiang Talent Programme Grant 2024A-158-G. Yangyang Wu is the corresponding author of the work.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection.

In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22*, pages 40–52. Springer.

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.

Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. 2018. Attending sentences to detect satirical fake news. In *Proceedings of the 27th international conference on computational linguistics*, pages 3371–3380.

Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. 2024. Game-on: Graph attention network based multimodal fusion for fake news detection. *Social Network Analysis and Mining*, 14(1):114.

Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608.

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

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

An Lao, Chongyang Shi, and Yayi Yang. 2021. Rumor detection with field of linear and non-linear propagation. In *Proceedings of the Web Conference 2021*, pages 3178–3187.

Yali Li, Shengjin Wang, Qi Tian, and Xiaoqing Ding. 2015. A survey of recent advances in visual feature detection. *Neurocomputing*, 149:736–751.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The world wide web conference*, pages 3049–3055.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 490–497.

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.

Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 518–527. IEEE.

Victoria L Rubin, Niall J Conroy, and Yimin Chen. 2015. Towards news verification: Deception detection methods for news discourse. In *Hawaii international conference on system sciences*, pages 5–8.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Ced: Credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3035–3047.

Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. 2024. Redcore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15173–15182.

Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Boyue Wang, Guangchao Wu, Xiaoyan Li, Junbin Gao, Yongli Hu, and Baocai Yin. 2024a. Modality perception learning-based determinative factor discovery for multimodal fake news detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Yuanzhi Wang, Zhen Cui, and Yong Li. 2023. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22034.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2024b. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36.

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*, pages 2560–2569.

Fan Xu, Pinyun Fu, Qi Huang, Bowei Zou, AiTi Aw, and Mingwen Wang. 2023. Leveraging contrastive learning and knowledge distillation for incomplete modality rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13492–13503.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In *IJCAI*, volume 2022, pages 2413–2419.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7178–7191.

Ting Zou, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2024. Pvcg: Prompt-based vision-aware classification and generation for multi-modal rumor detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11036–11040. IEEE.

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 9*, pages 109–123. Springer.

A Missing Modality Cases

Considering the three modalities $\{T, I, C\}$, where T donates a source text, I represents an image, and C refers to a comment, as shown in Figure 6, there are seven different cases of missing modalities.

B Dataset Description

Datasets	#Ns	#Rs	#Tuples	Modality types
Weibo-19	877	590	1,467	T & I & C
Pheme	1,428	590	2,018	T & I & C
Weibo-17	4,749	4,779	9,528	T & I

Table 4: The information of each dataset or our experiments, *i.e.*, the number of non-rumors (#Ns), rumors (#Rs), samples (#Tuples), and modality types, respectively. T (Text), I (Image), C (Comment).

Table 4 lists the information of each dataset.

- ***Weibo-19*** dataset is collected from Weibo, one of the most popular social platforms in China. It contains 1,467 data samples, among which there are 877 non-rumor samples and 590 rumor samples.
- ***Pheme*** dataset is constituted by tweets on the Twitter platform and based on five breaking news. It contains 2018 data samples, among which there are 1428 non-rumor samples and 590 rumor samples.
- ***Weibo-17*** dataset is collected from the Chinese social media platform, Weibo, containing 4,749 real, 4,779 fake tweets, and 9,528 images. The fake news in the dataset was verified by the debunking system from May 2012 to January 2016.

C Random Missing Protocol

In the experiments, we adopted the Random Missing Protocol mechanism to simulate the modality-missing situations in the real world, where each complete post is subject to the random absence of one or two modalities. For the three-modality datasets *Weibo19* and *Pheme*, we choose the R_m from $[0.0, \cdots, 0.7]$. For the two-modality dataset *Weibo17*, there are only samples with one modality missing and we choose the R_m from $[0.0, \cdots, 0.5]$. Table 5 shows the corresponding relationships between the proportion of samples with one or two modalities missing and the total number of samples at specific missing rates.

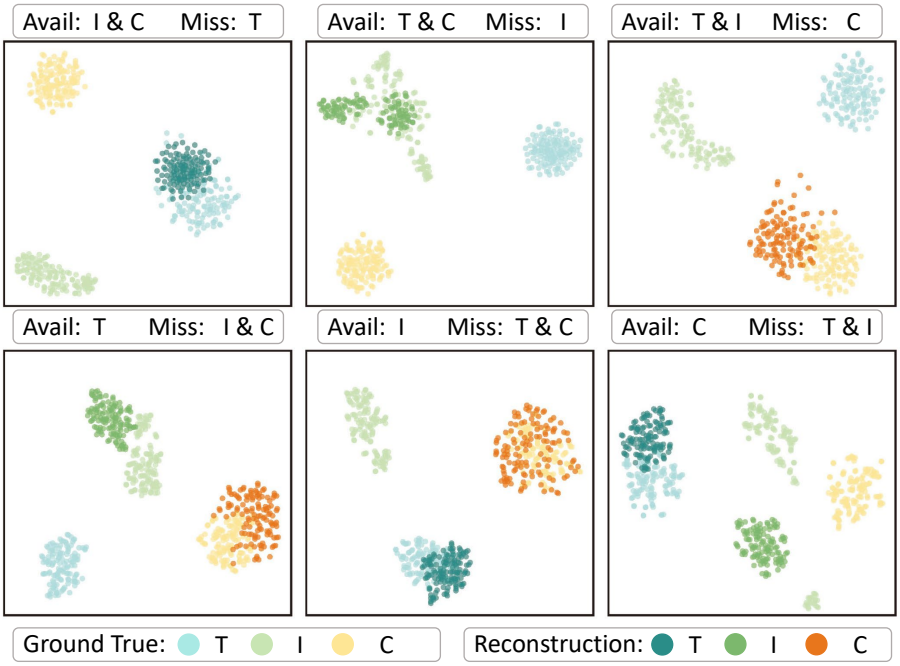


Figure 5: Visualization of reconstructed modality feature embeddings versus ground truth under the different missing modality cases. T (Text), I (Image), C (Comment).

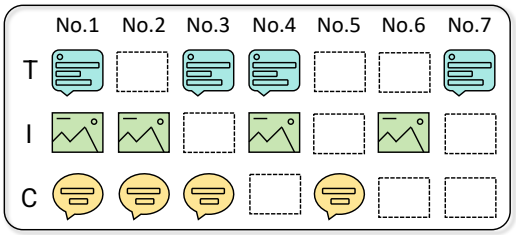


Figure 6: The seven different missing modality cases. Full modality: {No.1}. One modality is missing: {No.2, No.3, and No.4}. Two modalities are missing:{No.5, No.6, and No.7}. T (Text), I (Image), C (Comment).

R_m	*Weibo-19*		*Pheme*		*Weibo-17*
	R_o	R_t	R_o	R_t	R_o
0.0	0.0	0.0	0.0	0.0	0.0
0.1	0.1	0.1	0.1	0.1	0.2
0.2	0.2	0.2	0.2	0.2	0.4
0.3	0.3	0.3	0.3	0.3	0.6
0.4	0.2	0.5	0.2	0.5	0.8
0.5	0.1	0.7	0.1	0.7	1.0
0.6	0.0	0.9	0.0	0.9	–
0.7	0.0	1.0	0.0	1.0	–

Table 5: Modality-missing proportion at different missing rates. *i.e.*, the global missing rate (R_m), one modality missing rate (c), two modality missing rate (R_t).

D Reconstruct Missing Modality

Figure 5 visualizes the distribution of reconstructed modality feature embeddings versus ground truth under different missing modality cases. We randomly sample 128 instances in the testing set from the *Pheme* dataset for this comparison. The features of the selected samples are projected into a 2D space by t-SNE (Van der Maaten and Hinton, 2008). From visualization results, we observe that under varying modality-missing scenarios, TriSPrompt successfully reconstructs missing modality feature embeddings by synergistically leveraging homogeneous features and the heterogeneous features of each modality from modality-aware (MA) prompt. The reconstructed modality feature embeddings demonstrate distributional similarity to ground-truth feature embeddings, validating the efficacy of our reconstruction module. From experimental observations, we find that when both Text and Image modalities are absent, relying solely on the Comment modality fails to reconstruct accurate feature vectors. This phenomenon originates from inherent modality characteristics: Text and images generally represent the content creator's personal, subjective perspective, whereas comments tend to express more objective perspective from external users such as reviewers or readers.

We conducted quantitative experiments to complement our qualitative visualizations and provided a more objective evaluation of the reconstructed modality features showed in Table 6. Specifically, we computed the cosine similarity between the re-

Case (A, M)	Reconstructed	Random
I & C , T	0.5011	0.0038
T & C , I	0.5712	-0.0020
T & I , C	0.4638	-0.0095
T , I & C	0.4233	0.0051
I , T & C	0.4688	-0.0182
C , T & I	0.2765	-0.0056

Table 6: Cosine similarity results. Case (A, M): Available modalities (A), missing modality (M). Recon: Cosine similarity (Reconstructed vs GT). Random: Cosine similarity (Random vs GT). T (Text), I (Image), C (Comment).

constructed features and the ground truth (GT) features under various missing modality scenarios. As shown in the updated Table, the cosine similarity between the reconstructed and GT features ranges from 0.2765 to 0.6134 depending on the available modalities. In contrast, the cosine similarity between random features and GT features remains close to zero in all cases, confirming that our reconstruction is substantially better than random guessing. These quantitative results are consistent with our visualizations. This additional analysis demonstrates that our reconstruction method produces features that are quantitatively and qualitatively similar to the ground truth, and the results are not due to cherry-picking or visual bias.