The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

# Learning Complex Heterogeneous Multimodal Fake News via Social Latent Network Inference

**Mingxin Li, Yuchen Zhang, Haowei Xu, Xianghua Li*, Chao Gao, Zhen Wang**

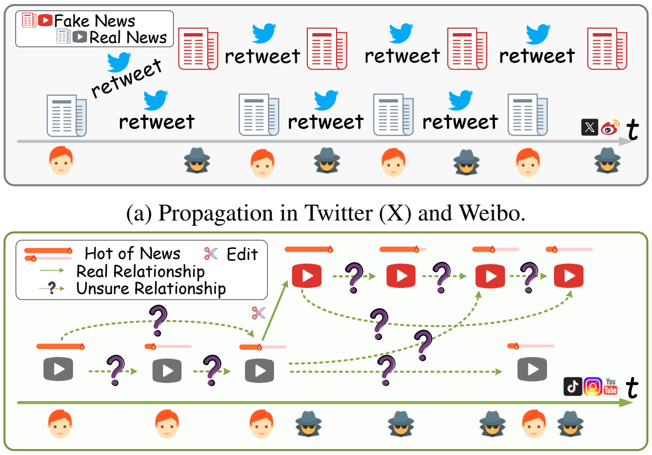
Northwestern Polytechnical University
{mingxinli, yc20011204, hwxu}@mail.nwpu.edu.cn, {li_xianghua, cgao, w-zhen}@nwpu.edu.cn

## Abstract

With the diversification of online social platforms, news dissemination has become increasingly complex, heterogeneous, and multimodal, making the fake news detection task more challenging and crucial. Previous works mainly focus on obtaining social relationships of news via retweets, limiting the accurate detection when real cascades are inaccessible. Given the proven assessment of the spreading influence of events, this paper proposes a method called **HML** (Complex **H**eterogeneous **M**ultimodal Fake News Detection method via **L**atent Network Inference). Specifically, an improved social latent network inference strategy is designed to estimate the maximum likelihood of news influences under the same event. Meanwhile, a novel heterogeneous graph is built based on social attributes for multimodal news under different events. Further, to better aggregate the relationships among heterogeneous multimodal features, this paper proposes a self-supervised-based multimodal content learning strategy, to enhance, align, fuse and compare heterogeneous modal contents. Based above, a personalized heterogeneous graph representation learning is designed to classify fake news. Extensive experiments demonstrate that the proposed method outperforms the SOTA in real social media news datasets.

## Introduction

Nowadays, **complex heterogeneous multimodal news** (e.g. short videos) has spread and proliferated rapidly due to their sharing nature on social media platforms (Gao, Liu, and Gao 2023; Zheng 2023). However, the increasing sophistication of forgery techniques has resulted in a large number of multimodal fake news flooding the Internet. These news may involve picture synthesis, audio and video editing, and the application of artificial intelligence technology, making it difficult for audiences to distinguish the real from the fake, and posing a serious challenge to the dissemination of information and social opinion (Bu et al. 2023; Bhargava et al. 2023). Therefore, it is an urgent and pressing task to study the detection of complex heterogeneous multimodal fake news.



(a) Propagation in Twitter (X) and Weibo.

(b) Unsure relationships in TikTok, Instagram and YouTube.

Figure 1: Differences in news dissemination across social media platforms.

To detect the increasingly complex and varied fake news, some studies have been proposed. To summarize, existing methods are mainly classified into two aspects, social relationship-based (Shu et al. 2020) and semantic feature-based (Wang et al. 2018). In the first stage, the social relationship is mainly represented between different news through graphs. In recent years, the methods have been used primarily for mining the social cascades through commenting and retweeting mechanism (Cheng et al. 2021; Zhang et al. 2023; Yin et al. 2024; Zhu et al. 2024). Graph structure mining is performed through GNN-based methods to accomplish the node classification or graph classification tasks to predict fake news. Further, related works focus on the learning and mining of attribute graphs, fusing graph and modal information for classification tasks (Nguyen et al. 2020; Phan, Nguyen, and Hwang 2023). For the other stage, existing methods mainly rely on extracting, fusing (Zhou, Wu, and Zafarani 2020; Nan et al. 2021), and enhancing different modal data (Zhu et al. 2023; Qi et al. 2021). To avoid simple feature fusion engineering, some researchers perform information integrity through fact checking (Vo and Lee 2018), reading interest (Wu, Liu, and Zhang 2023), external knowledge (Hu et al. 2021), etc. These methods have been confirmed to achieve good results.

*corresponding author

However, new issues have arisen to be addressed in the increasing online social platforms. As shown in Figure 1(a), when a hot social topic emerges, researchers can get almost complete information cascades on social platforms like Twitter, Weibo, etc. Nevertheless, as shown in Figure 1(b), with the emergence of video-based social media platforms such as TikTok, Instagram, and YouTube, **it has become increasingly difficult to directly acquire such relationships due to platform mechanics**. In addition, although there are many methods for heterogeneous multimodal fake news detection (Choi and Ko 2021; Shang et al. 2021; Qi et al. 2023a,b), all of them are simple splicing of heterogeneous multimodal features, which can only **primarily rely on pre-trained models combined with attention or transformer, using embeddings for modality alignment in news, which loses the relationship between some modalities**. For deep and fine-grained features such as editing and factual alteration, traditional analytics methods make it difficult to obtain the differences between them and real semantic features. Even powerful AI models can hardly detect deepfake videos effectively. Therefore, **mining the latent relationships of complex heterogeneous multimodal news and effectively aggregating the features**, becoming a great challenge.

To address the above issues, this paper proposes a novel Complex Heterogeneous Multimodal Fake News Detection method, HML. The method builds news latent cascade relationships through proven effective point time processes and social hot influence and performs latent network inference based on news attributes. Further, for better enhancement, alignment, fusing and comparison of different modal features, this paper proposes a Self-supervised-based Multimodal Content Learning strategy, which can get more effective feature representation. Finally, a personalized heterogeneous graph representation is established to represent the attribute heterogeneous graphs obtained above and to accomplish the fake news detection task. Overall, the application of network inference and self-supervised learning can effectively improve the robustness of the model.

For these reasons, the main contributions of this paper are as following three aspects:

- **News Latent Heterogeneous Graph Inference.** Based on the self-excitation of events over time and the influence of event popularity, latent cascade inference is performed. Combining this with news attributes and inferences a latent heterogeneous news graph.
- **Self-supervised-based Multimodal Content Feature Representation.** Content comprehension and augmentation of unimodal content are achieved using self-supervised learning strategy. Additionally, content contrast enhancement of cross-modal helps obtain effective features with robustness.
- **Personalized Representation and Better Performance.** Personalized heterogeneous graph representation of the above features was performed for the Complex Heterogeneous Multimodal Fake News Detection task. Extensive experiments on benchmark datasets show that the proposed method effectively obtains relationships among news and outperforms the SOTA method. [1].

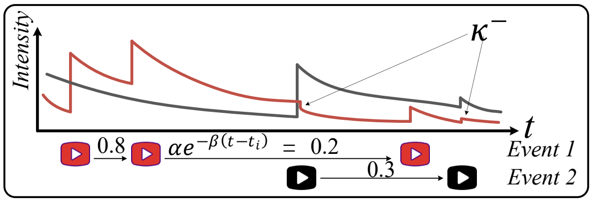


Figure 2: Illustrate of proposed event-based cascade influence. The red line represents the influence intensity of Event 1 over time $t$, which is affected by Event 2 (gray line).

## Preliminary

**Definition 1 (Attribute Heterogeneous News Graph)** *A attribute heterogeneous news graph is defined as a graph $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \mathcal{O}_E, \mathbf{A}, \mathcal{X})$ with news nodes $\mathcal{V}$ and multiple types of edges $\mathcal{E}$. $\mathcal{O}_E$ represents the set of object types of edges and $\mathcal{X}$ is the attribute representation of nodes, respectively. In addition, each node is associated with heterogeneous multimodal news.*

**Task 1 (Heterogeneous Latent Network Inference in Social Media Platforms)** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{O}_E, \mathbf{A})$ represents a latent network where $\mathcal{V}$ is the known set of nodes with $|\mathcal{V}| = n$, $\mathcal{E}$ is the unknown set of edges, and $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the unknown adjacency parameter matrix. Given a set of news attribute relationships $\mathcal{R}$ and latent cascades $\mathcal{C}$, the goal of news latent network inference is to estimate the adjacency parameter matrix $\mathbf{A}$ using $\mathcal{C}$ and $\mathcal{R}$, thereby inferring the underlying edge set $\mathcal{E}$ of the network $\mathcal{G}$.

**Task 2 (Complex Heterogeneous Multimodal Fake News Detection)** The dataset is defined as $V = \{v_1, v_2, \ldots, v_n\}$, where each $v_i$ is a news instance. Each instance $\mathbf{x}_i$ has at least 3 complex heterogeneous modalities, such as text, video frame, image, audio et al. Besides, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ are the feature representations of $V$, We aim to learn a self-supervised function, $g$, defined as:

$$g : V \to Y,$$

where $V$ represents news instances with their latent relationship processes and $Y \in \{Fake, Real\}$ denotes either fake or real news.

## Methodology

This section will introduce the proposed method HML in detail, the illustration of the method is shown in Figure 3.

### Social Latent Network Inference (Stage 1)

In the stages of news dissemination and evolution, there exist latent cascade paths that change over time, and different news items evolving at different times and states influence each other. Through these relationships, the process of news dissemination and evolution can be effectively established. Therefore, this section innovatively proposes a social latent network inference strategy, which builds a latent social network in the complete absence of real social relationships.

[1] https://github.com/cgao-comp/HML.

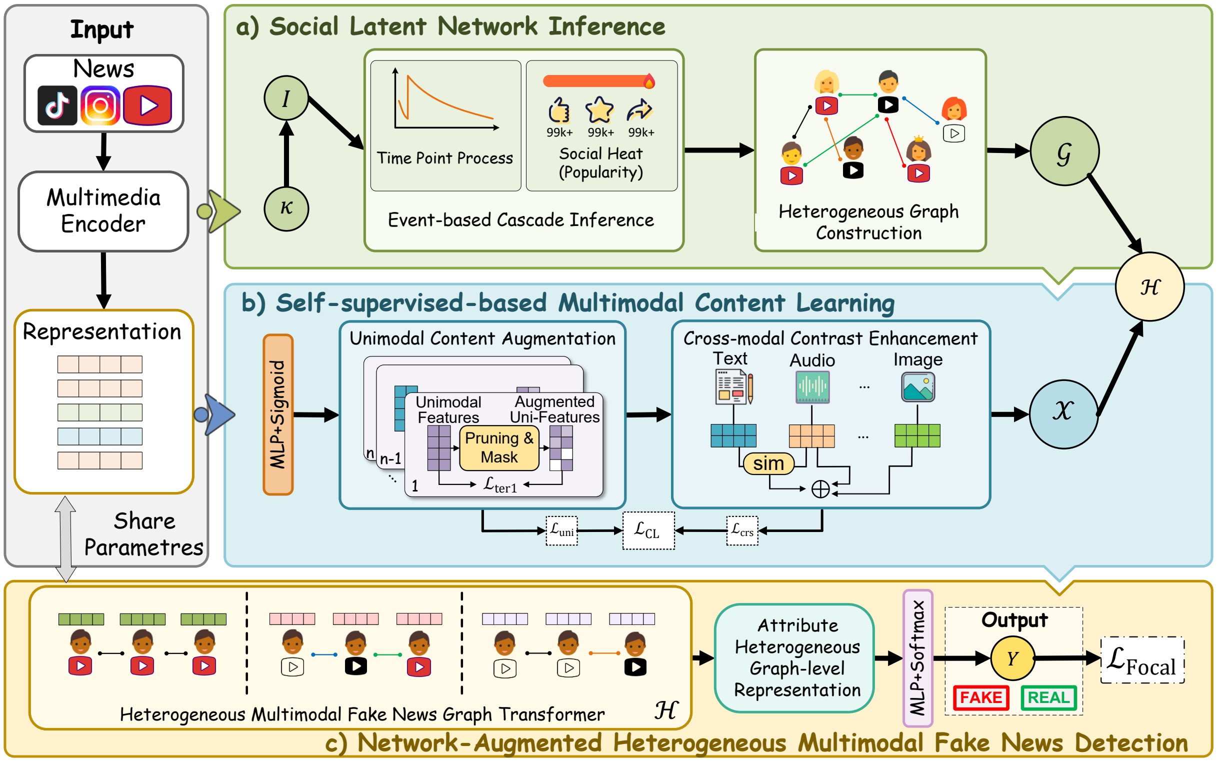

Input
News
Multimedia Encoder
Representation
Share Parametres
a) Social Latent Network Inference
$I$
$\kappa$
Time Point Process
99k+
99k+
99k+
Social Heat (Popularity)
Event-based Cascade Inference
Heterogeneous Graph Construction
$\mathcal{G}$
$\mathcal{H}$
b) Self-supervised-based Multimodal Content Learning
MLP+Sigmoid
Unimodal Content Augmentation
Unimodal Features
Augmented Uni-Features
Pruning & Mask
n
n-1
1
$\mathcal{L}_{\text{ter1}}$
Cross-modal Contrast Enhancement
Text
Audio
Image
...
sim
$\oplus$
$\mathcal{X}$
$\mathcal{L}_{\text{uni}}$
$\mathcal{L}_{\text{CL}}$
$\mathcal{L}_{\text{crs}}$
Heterogeneous Multimodal Fake News Graph Transformer
$\mathcal{H}$
Attribute Heterogeneous Graph-level Representation
MLP+Softmax
Output
$Y$
FAKE
REAL
$\mathcal{L}_{\text{Focal}}$
c) Network-Augmented Heterogeneous Multimodal Fake News Detection

Figure 3: Framework of HML. The whole framework is divided into three modules.

**Event-based Cascade Inference via Hawkes Process** When a hot social topic emerges, a large number of similar news appear on social network platforms, which are highly correlated and we define them as the same event. The influence intensity of news about the same event in the social network changes over time. When an event suddenly occurs, it often has the strongest influence, but the influence intensity gradually decreases over time. Additionally, the influence can undergo nonlinear changes due to the impact of news from the same event as well as news from different events(Matsubara et al. 2017). The framework is shown in Figure 3. This paper represents the influence of news over time by improved Hawkes process as formula (1).

$$\lambda^*(t) = \lambda_0 + \kappa(t), \tag{1}$$

where $\lambda^*(t)$ is the influence intensity function at time t, $\lambda_0$ is the baseline intensity, which is a constant, $\kappa(t)$ represents the influence of past news on the current intensity.The equation of $\kappa(t)$ is specifically expressed as follows:

$$\kappa(t) = \begin{cases} \kappa^+ = \sum\limits_{t_i<t} \alpha e^{-\beta(t-t_i)}, & \text{same event} \\ \kappa^- = \sum\limits_{t_i<t} \gamma e^{-\beta(t-t_i)} \cdot \text{prop}(\cdot), & \text{diff event} \end{cases} \tag{2}$$

where $\kappa^+$ represents the positive correlated impact of news within the same event due to self-excitation, and $\kappa^-$ represents the impact of news from different events on the current event's news. prop$(\cdot)$ indicates the correlation between news items, with values ranging from $[-1, 1]$. When a highly popular news item with poor correlation arises, it negatively impacts news from other events. Therefore, the original correlation is mapped using the equation prop$(\cdot) = \tanh(2 \cdot \text{sim}(\cdot,\cdot) - 1)$. $\alpha$, $\beta$, and $\gamma$ are statistical parameters. The explanation of how to calculate prop(s) as follows:

During normal dissemination, older news tends to be forgotten more quickly. Consequently, the likelihood of people associating with related news decreases. To simulate this characteristic, this paper introduces Gaussian Noise based on news dissemination, described by the following equation[2]:

$$\begin{gathered} q\left(\mathbf{x}^1, ..., \mathbf{x}^T | \mathbf{x}^0\right) = \prod_{t=1}^{T} q\left(\mathbf{x}^t | \mathbf{x}^{t-1}\right) \\ q(\mathbf{x}^t | \mathbf{x}^{t-1}) = \mathcal{N}(\mathbf{x}^t; \sqrt{\eta_t}\mathbf{x}^{t-1}, (1-\eta_t)\mathbf{I}) \end{gathered}, \tag{3}$$

where $T$ is the number of steps to add noise, $\eta$ is a learning variation of time process. This noising process can be defined as a Markov process. To simplify the computation, it can be transformed into the following.

$$\begin{gathered} q\left(\mathbf{x}^t \mid \mathbf{x}^0\right) = \mathcal{N}\left(\mathbf{x}^t; \sqrt{\bar{\delta}_t}\mathbf{x}^0, \left(1-\bar{\delta}_t\right)\mathbf{I}\right) \\ \mathbf{x}^t = \sqrt{\bar{\delta}_t}\mathbf{x}^0 + \sqrt{1-\bar{\delta}_t}\boldsymbol{\epsilon} \end{gathered}, \tag{4}$$

where $\delta_t = 1 - \eta_t, \bar{\delta}_t = \prod_{s=1}^{t} \delta_s$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

After applying the noising process to news from other events, the similarity sim$(\cdot,\cdot)$ is calculated using the

[2] Distinct from stage 2, features are not fine-tuned.

Minkowski distance as follows:

$$\mathrm{prop}(\mathbf{x}_j) = \mathrm{sim}(\mathbf{x}_i^t, \mathbf{x}_j) = (\sum_{k=1}^{D} |\mathbf{x}_{i,k}^t - \mathbf{x}_{j,k}|^D)^{1/D}, \tag{5}$$

where $\mathbf{x}_i, \mathbf{x}_j$ are attributes of different nodes, $k$ is an iterator, and $D$ is the feature dimension.

To estimate the statistics parameters, the log likelihood function of the proposed improved news Hawkes process is as follows:

$$\mathscr{L} = \log \prod_{i=1}^{N} \lambda^*(t_i) \exp\left(-\int_0^T \lambda^*(t)dt\right), \tag{6}$$

where the approximation simplifies the calculation by approximating the integral with a summation over discrete time points.

To calculate the influence of the $i$-th news on the current news at time $t$, we have the following equation:

$$I(t)_i = \alpha e^{-\beta(t-t_i)} + \kappa^-(t), \tag{7}$$

where $\alpha e^{-\beta(t-t_i)}$ represents the influence of the $i$-th news on the current news' popularity within the same event, and $\kappa^-(t)$ is the influence of other news on the current news.

After calculating the influence $I(t)$ of news within the same event, we estimate the adjacency parameter matrix $\mathbf{A}_e$ for the event. At this point, $\mathbf{A}_e$ is an $[0,1]^{r\times r}$ matrix. $r$ is the news number of the event (different events have different $r$). This transformed matrix can serve as a prior condition for establishing the heterogeneous news relationship graph described in the next section.

**Heterogeneous Graph Construction** Unlike widely used heterogeneous graphs, the heterogeneous graph provided in this paper does not contain nodes of different attributes such as paper-author, paper-author-venue, or user-item, all nodes are news items. When news is published on social media platforms such as TikTok or YouTube, a large amount of information about the news and its authors emerges. This information possesses very distinct statistical and sociological characteristics. Hou et al. 2024a have statistically proven this point. For example, in the process of news dissemination, compared to real news, fake news often lacks user authentication and exhibits extremely severe differences in title characteristics. However, these features are often not effectively utilized in the complex heterogeneous multimodal fake news detection task, where only simple features are used for feature extraction. Therefore, classifying attributes is key to constructing the proposed heterogeneous graph.

Inspired by Qi et al. 2023a and Hou et al. 2024b, this paper constructs news using the top few most significant statistical features with the largest differences. As shown in Figure 4, these are the attribute relationships that can be constructed in the FakeSV dataset. Additionally, Some attribute relationships are composed of similarities. However, since this paper assumes no prior knowledge of the real network, a threshold $\rho$ is designed to map the $[0,1]^{n\times n}$ matrix to a $\{0,1\}^{n\times n}$ matrix. We have $f(v_i, v_j) = \begin{cases} 1 & \text{if sim} \geq \rho \\ 0 & \text{if otherwise} \end{cases}$. Among them, the matrices composed of different edge types are named $\mathbf{A}_i, i \in \mathcal{O}_E$. It should be noted that $\mathbf{A}_e$ is a submatrix of $\mathbf{A}_1$. And the setting of edges is shown in Fig. 4.

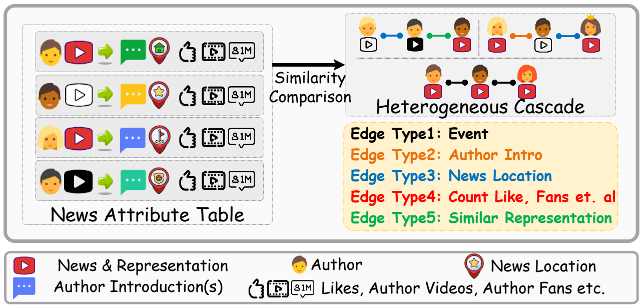


Figure 4: An example of the proposed Heterogeneous Graph Construction strategy in dataset FakeSV.

## Self-supervised-based Multimodal Content Learning (Stage 2)

**Pruning and Mask** Data imbalance is a very important issue that affects the performance of comparative learning. Nowadays, there have been many works to solve the problem of data imbalance (Jiang et al. 2021; Frankle and Carbin 2018). In this paper, we follow a pruning method for data augmentation on unbalanced data (Frankle and Carbin 2018).

To perform effective data augmentation, some modal features of the news are masked to reconstruct their initial features, resulting in new effective representations. The masking method for features can be described as $\hat{\mathbf{x}}_i = \begin{cases} \mathbf{x}_{[MASK]} & \text{if } v_i \in V^m \\ \mathbf{x}_i & \text{if } v_i \notin V^m \end{cases}$, $V^m$ is the news mask list.

**Unimodal Content Augmentation** This section will be devoted to the unimodal content augmentation strategy used. Unimodal content learning is mainly applied to the analysis of correlations within similar modalities, where the original modal features are compared with the augmented features to calculate the loss with the following equation:

$$\mathcal{L}_{\text{uni}} = -\log \sum_{\mathbf{x}_i \in \mathcal{X}} \frac{\exp\left[\mathrm{sim}\left(\widetilde{z}^i, \widetilde{z}_p^j\right) / \tau_{\text{uni}}\right]}{\sum_{p=1}^{2n} \nVdash_{[i\neq p]} \exp\left[\mathrm{sim}\left(\widetilde{z}^i, \widetilde{z}_p^p\right) / \tau_{\text{uni}}\right]}, \tag{8}$$

where $\widetilde{z}$ represents the processed correlation feature. $\widetilde{z}^i, \widetilde{z}^j, \widetilde{z}^p$ are unpruned data, $\widetilde{z}_p^j, \widetilde{z}_p^p$ are augmented data. $\tau_{\text{uni}}$ is the temperature value, which is a hyperparameter.

**Cross-modal Contrast Enhancement** The strategy is mainly used for the comparison between different modal data. All contrast losses are calculated and summed to get the cross-modal contrast loss,

$$\mathcal{L}_{\text{crs}} = -\sum_i \log \sum_{\mathbf{x}_i \in \mathcal{X}} \frac{\exp\left[\mathrm{sim}\left(\widetilde{z}^i, \widetilde{z}^j\right) / \tau_{\text{crs}}\right]}{\sum_{p=1}^{2n} \nVdash_{[i\neq p]} \exp\left[\mathrm{sim}\left(\widetilde{z}^i, \widetilde{z}^p\right) / \tau_{\text{crs}}\right]}. \tag{9}$$

**Multimodal Loss Integration** To integrate the above strategies, a hyperparameter $\lambda$ is set for joint loss to use for tuning the network.

$$\mathcal{L}_{CL} = \lambda \mathcal{L}_{\text{uni}} + (1-\lambda)\mathcal{L}_{\text{crs}} + ||\mathbf{\Theta}||_2, \tag{10}$$

where $||\mathbf{\Theta}||_2$ represents the L2 normalization. The parameters of above are updated through the above process to prepare for the next step.

## Network-Augmented Heterogeneous Multimodal Fake News Detection Task

According to the model after parameter update, new latent graph and attributes are obtained for the personalized node classification task.

**Edge Types Combination** The previous heterogeneous graph construction has generated edges of different types. These edges obviously have different contributions to the graph. Thus, it needs to be represented differently. We employ an attention mechanism to solve the difficulties. The aggregate method is formulated as:

$$\mathbf{A} = \sum_{i \in \mathcal{O}_E} \omega^{v \leftrightarrow v} \mathbf{A}_i, \tag{11}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the final heterogeneous adjacency matrix, $\omega^{v \leftrightarrow v}$ is the importance of different types of edges, which is learnable in attention mechanism.

Finally, to get the effective representation, this paper explains a useful mechanism of graph representation, graph transformer encoder (Zhu et al. 2024). Unlike other methods, this approach combines modal features and graph structures to obtain an effectively integrated feature representation. The equation is designed below:

$$\mathrm{Attn}^h(\mathbf{Q}, \mathbf{K}, \mathbf{A}) = \mathrm{softmax}\left(\mathcal{M}\left(\frac{\mathbf{Q}^h \mathbf{K}^{hT}}{\sqrt{d_H}}, \mathbf{A}\right)\right), \tag{12}$$

where $\mathcal{M}$ is masking function, defined as: $\mathcal{M}(u, v) = u + \zeta v$, $\zeta$ is a sufficiently large value.

**Classification** To improve the robustness of the model, a focal loss (Lin et al. 2017) is introduced in this paper, as shown in equation (13). By designing two hyperparameters, the nodes that are difficult to classify are centralized.

$$\mathcal{L}_{\text{focal}} = -\frac{1}{|\mathcal{V}_l|} \sum_{i \in \mathcal{V}_l} \sum_{c=0}^{C} \phi_c y_{ic} \left(1 - \hat{y}_{ic}\right)^{\psi} \log\left(\hat{y}_{ic}\right), \tag{13}$$

where $\mathcal{V}_l$ is the node containing the label, $C$ is the category, and $\phi_i, \psi \in [0, 1]$, is a tunable hyperparameter. $y_{ic}$ is the actual label value and $\hat{y}_{ic}$ is the predicted label value.

# Experiments

In this section, we conduct some extensive experiments to evaluate the proposed framework, HML.

## Benchmark Datasets

This paper applies two complex heterogeneous multimodal fake news detection datasets, FakeSV (Qi et al. 2023a) and FVC (Papadopoulou et al. 2019). Additionally, to prove the social latent graph inference method can be widely applied to multimodal fake news datasets, this paper additionally uses Twitter (Boididou et al. 2018) and Weibo (Wang et al. 2018) datasets to evaluate the inference task as a plugin.

## Baselines

In order to prove the superiority of the method, some of the more advanced algorithms are compared as follows:

**Unimodal** Traditional analytical methods mainly explore the expressive unimodal features. This paper mainly use **BERT** (Devlin et al. 2019), **VGGish** (Hershey et al. 2017), **VGG19** (Simonyan and Zisserman 2015), and **C3D** (Ji et al. 2013) to analyze the characteristics of their respective modalities.

**Multi-modal** Existing multimodal methods mainly focus on cross-modal methods , such as **Serrano et al. 2020**, **FANVM** (Choi and Ko 2021), **SV-FEND** (Qi et al. 2023a), **MMVD** (Zeng et al. 2024), **NEED** (Qi et al. 2023b), and **FakingRecipe** (Bu et al. 2024). However, fewer analyses have both semantic features and social features. This paper will compare the aforementioned work with the proposed method HML.

**Large Language Model** To explore the application ability of LLMs for the task of fake news detection, this paper designs the prompt method and conducts experiments through the APIs of **Doubao**[3] and **GPT-4o**[4]. Additionally, a baseline model, **ARG** (Hu et al. 2024), for detecting fake news in LLM has also been evaluated.

## Experimental Setup

**Metrics and Parameters** To prevent the model from capturing features of news evolution, this paper divides the dataset 80:20 following the FakeSV benchmark. To validate the effectiveness of the model, this paper uses a computational classification accuracy method to evaluate the effectiveness of the model using Accuracy, Precision, Recall, and F1-score with interval estimation and K-S test. For parameters, this paper uses 4 × NVIDIA RTX A6000 as GPU. Besides, the method achieves optimal performance with $\lambda = 0.5$ and pruning ratio $e = 0.6$.

## Overall Performance

We have compared our methods to unimodal methods, multimodal methods, and large language model prompt methods respectively, and the comparison results obtained are shown in Table 1, which are analyzed as follows,

**Performance** After the above experiments, all the metrics of our method reach more than 89% in the FakeSV dataset and 90% in the FVC dataset, which is an improvement of $0.12\% \sim 4.39\%$ compared with the previous best method in both datasets. After comparing with the unimodal method and LLM method, it is found that the proposed method substantially improves the relevant metrics.

**Analysis** For unimodal classification methods, only a very small portion of the news information can be learned. For prior multimodal methods, the semantic understanding methods for news are mainly unimodal and cross-modal simple feature extraction fusion and classification, which

[3] https://www.doubao.com

[4] https://chatgpt.com

| | Method | FakeSV | | | | FVC | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Acc.** | **Prec.** | **Rec.** | **F1** | **Acc.** | **Prec.** | **Rec.** | **F1** |
| Unimodal | Text (BERT) | 77.06 | 77.16 | 77.07 | 77.04 | 61.70 | 61.81 | 61.72 | 61.76 |
| | Audio (VGGish) | 66.72 | 67.05 | 66.70 | 66.53 | 58.44 | 58.48 | 58.63 | 58.61 |
| | Image (VGG19) | 69.66 | 69.78 | 69.78 | 69.59 | 65.79 | 65.49 | 66.08 | 65.81 |
| | Video (C3D) | 69.59 | 70.07 | 69.56 | 69.38 | 71.81 | 71.89 | 71.85 | 71.72 |
| Multimodal | Serrano et al. | 68.71 | 70.74 | 68.73 | 67.91 | 66.87 | 67.15 | 66.34 | 66.74 |
| | FANVM | 76.37 | 75.39 | 73.71 | 74.18 | 85.81 | 85.20 | 85.44 | 85.35 |
| | SV-FEND | 78.88 | 79.41 | 78.89 | 78.79 | 84.71 | 84.25 | 86.53 | 85.37 |
| | MMVD | 82.64 | 82.63 | 82.73 | 82.63 | 89.28 | 90.27 | **90.36** | 90.46 |
| | NEED | 84.62 | 84.81 | 84.64 | 84.61 | - | - | - | - |
| | FakingRecipe | 85.35 | 85.84 | 84.29 | 84.83 | - | - | - | - |
| LLM | Doubao | 70.25 | 72.60 | 70.24 | 69.45 | 45.89 | 44.40 | 45.89 | 44.82 |
| | GPT-4o | 73.07 | 73.31 | 73.06 | 72.99 | 50.72 | 49.94 | 50.72 | 50.17 |
| | ARG | 78.32 | 78.60 | 78.36 | 78.28 | 84.44 | 85.19 | 83.93 | 84.16 |
| | **HML(Ours)** | **89.14**(+3.79%) | **89.36** | **89.22** | **89.22** (+4.39%) | **91.02**(+1.74%) | **91.68** | 90.32 | **90.58**(+0.12%) |

Table 1: Experimental results of different methods. The experiments are conducted to compare unimodal, multimodal, and large language models. Besides, the last line shows the enhancement of the proposed method to baseline methods. Underline denotes the second-best metric, while **bold** denotes the best metric. (Acc.: accuracy, Prec.: precision, Rec.: recall)

| Method | Twitter | | Weibo | |
|---|---|---|---|---|
| | **Acc.** | **F1** | **Acc.** | **F1** |
| EANN | 64.86 | 63.92 | 79.56 | 80.03 |
| +*NI* | **74.72** | **74.63** | **84.89** | **84.38** |
| | (+9.86%) | (+10.71%) | (+5.33%) | (+4.35%) |
| MCAN | 74.55 | 74.85 | 89.96 | 89.33 |
| +*NI* | **77.45** | **77.92** | **90.27** | **90.60** |
| | (+2.90%) | (+3.07%) | (+0.31%) | (+1.27%) |
| CAFE | 80.62 | 80.38 | 84.13 | 83.77 |
| +*NI* | **83.40** | **82.94** | **85.91** | **85.68** |
| | (+2.78%) | (+2.56%) | (+1.78%) | (+1.91%) |

Table 2: Plugin Experiments on Twitter and Weibo.
NI: Network Inference

cannot effectively recognize the fine-grained differences between news. Our approach HML both alignes, fuses multimodal features, and mines latent networks between news. In addition, contrastive learning is used to amplify the differences between different samples. Thus, our method outperforms the state-of-the-art method.

**Compare with LLM** According to the experimental results, it can be seen that LLM performs poorly in the multimodal fake news detection task, mainly because the prompt task mainly calls the original features of the model, and it cannot learn from the newly generated news. Therefore, how to adjust LLM to adapt the multimodal fake news detection task may be a topic that can be investigated in the next step.

**Extensive Applications** To demonstrate the effectiveness of the method on a broad range of datasets, this section evaluates it on traditional image-text fake news detection datasets. The network inference method proposed in this paper is used as a plugin and combined with existing classical methods to create a improved fake news detection method. As is shown in Table 2, evaluations are conducted using the EANN (Wang et al. 2018), MCAN (Wu et al. 2021), and CAFE (Chen et al. 2022) methods, with improvements ranging from 0.3% to 10% relative to the baseline. Combined with the ablation experiment results in the next section, our analysis suggests that the proposed method for establishing the latent network can effectively capture features. On the same dataset, since the same latent network structure is obtained, the improvement in metrics depends on the feature extraction capability of the baseline method.

## Ablation Study

In order to prove the effectiveness of the proposed strategies, this section investigates the ablation experiments. Experiment results of FakeSV see Table 3.

**Ablation Study about Stage 1** To ensure that each component in the latent network inference has a positive effect on the whole framework, we designed 3 ablation experiments.

- **w/o Latent Graph (LG)** remove inference the latent (heterogeneous) graph.
  - **w/ LG w/o Event Inference** remove the event based latent inference.
  - **w/ LG w/o Edge Type Construction** remove performing the establishment of heterogeneous graph with edge types.

The experimental results prove the necessity of the latent network inference. Considering only the relationship inherent in the events, the experiment proved that the indicator was reduced by 0.52%, indicating that there is a latent relationship between different events. For removing the edge-type construction strategy, only the similarity relationship between the news is considered and the event factor is not taken into account. The experiment proves that the indicator is reduced by 2.29%. However, if the latent network inference method is completely removed, the experimental results drastically drop by 8%, effectively proving the importance of the proposed strategy.

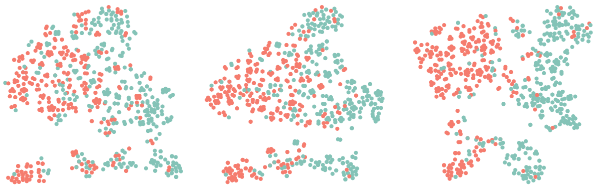
| Method | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|
| **HML(Ours)** | **89.14** | **89.36** | **89.22** | **89.22** |
| w/o Stage1 | 81.68 | 82.21 | 81.31 | 81.44 |
| - Event Inference | 86.84 | 86.91 | 86.87 | 86.93 |
| - Edge Type Cons. | 88.70 | 89.06 | 88.98 | 88.70 |
| w/o Stage 2 | 78.57 | 78.30 | 77.89 | 78.20 |
| - Unimodal Augmentation | 88.80 | 88.95 | 89.00 | 88.79 |
| - Cross-modal Enhancement | 88.40 | 88.41 | 88.51 | 88.39 |

Table 3: Ablation study on FakeSV.

**Ablation Study about Stage 2** To justify the Self-supervised-based Multimodal Content Learning, this section conducts ablation experiments on the proposed strategy by eliminating the Unimodal Content Augmentation and Cross-modal Contrast Enhancement, respectively.

- **w/o Multimodal Content Learning (MCL)** remove all the content learning strategy.
  - **w/ MCL w/o Unimodal Augmentation** remove the unimodal augmentation and retain only the cross-modal enhancement strategy.
  - **w/ MCL w/o Cross-modal Enhancement** remove the cross-modal enhancement strategy and retain only the unimodal augmentation.

Two content learning strategies are removed separately and it can be seen that there is a certain decrease of about 10% in metrics after removal. This indicates that relying solely on the latent network inference is insufficient for content learning to effectively capture the important feature relationships in news. Furthermore, though the metrics for removing content learning separately do not show more significant improvement compared to stage 1, it still achieves the purpose of proposing this strategy. For instance, for unlabeled data, the proposed strategy can also enhance the performance through augmentation and comparison.



(a) BERT (b) SV-FEND (c) HML

Figure 5: T-SNE Visualization of FakeSV dataset.

## Visualization and Case Study

**Visualization** This section presents a dimensionality reduction visualization of the classification features for fake news detection. Figure 5 shows the feature distributions obtained by different methods on the test set of FakeSV dataset, which contains a total of 717 news articles, with 352 being real and 365 being fake. It is clearly evident that the proposed method clusters news contents of the same category more compactly, qualitatively demonstrating the effectiveness of the method.

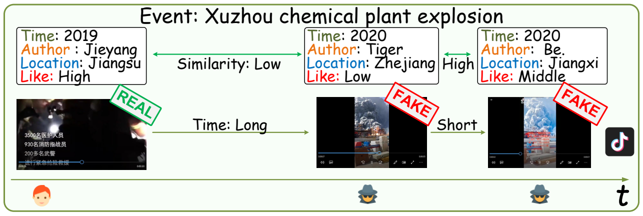


Figure 6: Case study of proposed method HML in Douyin social media platform.

**Case Study** This paper uses the explosion incident in Xuzhou, Jiangsu as an example to illustrate this proposed method. As shown in Figure 6, the chemical plant incident attracted widespread attention in a short period, and the related news received a large number of likes and shares. Official media typically release news immediately after an incident occurs, while fake news often does not meet these conditions. For example, the second piece of news was circulated a year after the event was forgotten, with some media editors reusing old news and adding their own commentary. These self-media sources are usually unofficial, with IP addresses inconsistent with the event location, and have poor relevance, making them identifiable as fake news. By reasoning through the latent network rather than the propagation relationship, fake news can be detected more accurately. The third video, similar to the second one, also meets the criteria for fake news, and when the second piece of news is detected, the third can also be identified.

## Conclusion

This paper explores the latent network inference and the implementation of modal enhancement methods for complex heterogeneous multimodal news. Specifically, the social latent network is obtained through event-based cascade inference and relation-based heterogeneous graph construction strategies. Additionally, a news modality learning method is designed to enhance news features and facilitate expansion to large-scale datasets. Extensive experimental results show that the proposed method can get effective network representation capabilities, and the fake news detection method achieves state-of-the-art performance on benchmark datasets.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 62271411, 62471403, U22A2098, 62261136549, U22B2036), the Fundamental Research Funds for the Central Universities (Nos. G2024WD0151, D5000240309), and the XPLORER PRIZE; Sponsored by the Practice and Innovation Funds for Graduate Students of Northwestern Polytechnical University.

# References

Bhargava, P.; MacDonald, K.; Newton, C.; Lin, H.; and Pennycook, G. 2023. How Effective are TikTok Misinformation Debunking Videos? *Harvard Kennedy School Misinformation Review*.

Boididou, C.; Papadopoulos, S.; Zampoglou, M.; Apostolidis, L.; Papadopoulou, O.; and Kompatsiaris, Y. 2018. Detection and Visualization of Misleading Content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1): 71–86.

Bu, Y.; Sheng, Q.; Cao, J.; Qi, P.; Wang, D.; and Li, J. 2023. Combating Online Misinformation Videos: Characterization, Detection, and Future Directions. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8770–8780.

Bu, Y.; Sheng, Q.; Cao, J.; Qi, P.; Wang, D.; and Li, J. 2024. FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1351–1360.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*, 2897–2905.

Cheng, L.; Guo, R.; Shu, K.; and Liu, H. 2021. Causal Understanding of Fake News Dissemination on Social Media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 148–157.

Choi, H.; and Ko, Y. 2021. Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2950–2954.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Frankle, J.; and Carbin, M. 2018. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *Proceedings of the 6th International Conference on Learning Representations*.

Gao, Y.; Liu, F.; and Gao, L. 2023. Echo Chamber Effects on Short Video Platforms. *Scientific Reports*, 13(1): 6282.

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. 2017. CNN Architectures for Large-scale Audio Classification. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 131–135.

Hou, D.; Gao, C.; Li, X.; and Wang, Z. 2024a. DAG-Aware Variational Autoencoder for Social Propagation Graph Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8508–8516.

Hou, D.; Yin, S.; Gao, C.; Li, X.; and Wang, Z. 2024b. Propagation Dynamics of Rumor vs. Non-rumor across Multiple Social Media Platforms Driven by User Characteristics. *arXiv preprint arXiv:2401.17840*.

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad actor, Good advisor: Exploring the Role of Large Language Models in Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22105–22113.

Hu, L.; Yang, T.; Zhang, L.; Zhong, W.; Tang, D.; Shi, C.; Duan, N.; and Zhou, M. 2021. Compare to the Knowledge: Graph Neural Fake News Detection with External Knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 754–763.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 221–231.

Jiang, Z.; Chen, T.; Mortazavi, B. J.; and Wang, Z. 2021. Self-damaging Contrastive Learning. In *International Conference on Machine Learning*, 4927–4939.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Matsubara, Y.; Sakurai, Y.; Prakash, B. A.; Li, L.; and Faloutsos, C. 2017. Nonlinear Dynamics of Information Diffusion in Social Networks. *ACM Transactions on the Web*, 11(2): 1–40.

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. MDFEND: Multi-domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3343–3347.

Nguyen, V.-H.; Sugiyama, K.; Nakov, P.; and Kan, M.-Y. 2020. FANG: Leveraging Social Context for Fake News Detection using Graph Representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1165–1174.

Papadopoulou, O.; Zampoglou, M.; Papadopoulos, S.; and Kompatsiaris, I. 2019. A Corpus of Debunked and Verified User-generated Videos. *Online Information Review*, 43(1): 72–88.

Phan, H. T.; Nguyen, N. T.; and Hwang, D. 2023. Fake News Detection: A Survey of Graph Neural Network Methods. *Applied Soft Computing*, 139: 110235.

Qi, P.; Bu, Y.; Cao, J.; Ji, W.; Shui, R.; Xiao, J.; Wang, D.; and Chua, T.-S. 2023a. 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*, 14444–14452.

Qi, P.; Cao, J.; Li, X.; Liu, H.; Sheng, Q.; Mi, X.; He, Q.; Lv, Y.; Guo, C.; and Yu, Y. 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*, 1212–1220.

Qi, P.; Zhao, Y.; Shen, Y.; Ji, W.; Cao, J.; and Chua, T.-S. 2023b. Two Heads Are Better Than One: Improving Fake News Video Detection by Correlating with Neighbors. In *Findings of the Association for Computational Linguistics 2023*, 11947–11959.

Serrano, J. C. M.; Papakyriakopoulos, O.; Hegelich, S.; and Hegelich, S. 2020. 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*.

Shang, L.; Kou, Z.; Zhang, Y.; and Wang, D. 2021. A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. In *Proceedings of the 2021 IEEE International Conference on Big Data*, 899–908.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big data*, 8(3): 171–188.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-scale Image Recognition. In *Proceedings of the 3rd International Conference on Learning Representations*.

Vo, N.; and Lee, K. 2018. The rise of guardians: Fact-checking URL Recommendation to Combat Fake News. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 275–284.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. EANN: Event Adversarial Neural Networks for Multi-modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 849–857.

Wu, L.; Liu, P.; and Zhang, Y. 2023. See How you Read? Multi-Reading Habits Fusion Reasoning for Multi-modal Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13736–13744.

Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; and Xu, Z. 2021. Multimodal Fusion with Co-attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2560–2569.

Yin, S.; Zhu, P.; Wu, L.; Gao, C.; and Wang, Z. 2024. GAMC: an Unsupervised Method for Fake News Detection using Graph Autoencoder with Masking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 347–355.

Zeng, Z.; Luo, M.; Kong, X.; Liu, H.; Guo, H.; Yang, H.; Ma, Z.; and Zhao, X. 2024. Mitigating World Biases: A Multimodal Multi-View Debiasing Framework for Fake News Video Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6492–6500.

Zhang, K.; Yu, J.; Shi, H.; Liang, J.; and Zhang, X.-Y. 2023. Rumor Detection with Diverse Counterfactual Evidence. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3321–3331.

Zheng, C. 2023. Research on the Flow Experience and Social Influences of Users of Short Online Videos. A Case Study of DouYin. *Scientific Reports*, 13(1): 3312.

Zhou, X.; Wu, J.; and Zafarani, R. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 354–367.

Zhu, J.; Gao, C.; Yin, Z.; Li, X.; and Kurths, J. 2024. Propagation Structure-Aware Graph Transformer for Robust and Interpretable Fake News Detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4652–4663.

Zhu, Y.; Sheng, Q.; Cao, J.; Nan, Q.; Shu, K.; Wu, M.; Wang, J.; and Zhuang, F. 2023. Memory-guided Multi-view Multi-domain Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(7): 7178–7191.