Short Video is not only Video: Multimodal Unified Social Hypergraph Contrastive Enhancement for Fake News Video Detection

Anonymous ACL submission

Abstract

Nowadays, fake short videos have seriously affected people's perception of news and situational awareness of event development. Previous work mainly focuses on the characteristics and dissemination of the news, and there is no in-depth mining of the social relationships and feature relationships of videos. This paper proposes a **M**ultimodal **U**nified Social **H**ypergraph **C**ontrastive Enhancement method (**MUHC**) for fake news videos detection. First, a unified social hypergraph is innovatively established for the representation of potential relationships in short videos. Meanwhile, a multimodal contrastive learning method for intra-modal and inter-modal relationships are designed to integrate different modalities. The above approach enhances data scalability while learning deeper about the potential relationships of the videos. Extensive experiments demonstrate that the method outperforms state-of-the-art on benchmark dataset.[1]

1 Introduction

The emergence of social networks has not only brought about the convenience of communication and access to knowledge, but also resulted in an abundance of fake news, which has caused significant impacts in politics(Fisher et al., 2016), economy(ElBoghdady, 2013), culture(Olan et al., 2024) and health(Chen et al., 2023). Nowadays, with the rise of short video platforms, the carrier of information exists not only in simple text, image, and audio, but in social relationships and user interactions, making the detection of fake news increasingly difficult(Comito et al., 2023). Therefore, it is urgent to design an automated fake news video detection method to decrease the negative impact.

Existing multimodal fake news detection methods are mainly classified into two aspects, social relationship-based(Shu et al., 2020) and semantic feature-based(Wang et al., 2018). For the first stage, social relationship is mainly represented between different news through graph. In recent years, the methods are mainly used to construct the propagation relationship graph through commenting and retweeting mechanism(Cheng et al., 2021; Zhang et al., 2023). 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 et al., 2023).For the other stage, existing methods mainly rely on extracting, fusing(Zhou et al., 2020; Nan et al., 2021), and enhancing different modal data(Zhu et al., 2022; 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 et al., 2023), external knowledge(Hu et al., 2021), etc. These methods have been confirmed to achieve good results.

However, there are certain challenges in the above works. For social relations, due to the different recommendation mechanisms, short video data social relations are more complex. It is hard to obtain the propagation relations of the videos

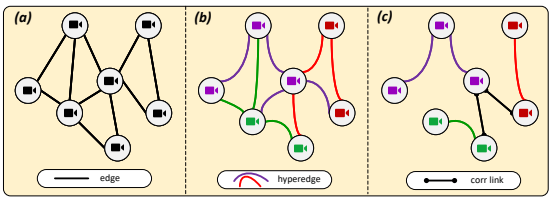

Figure 1: Illustration of a) video potential relationship graph, b) video potential relationship hypergraph and c) unified social hypergraph. **HINT**: where corr link connects event relationship. It can be clearly seen that the unified social hypergraph can represent the same video correlation using fewer edges.

[1]Code will release if manuscript acceptable.

in the same way as Twitter and Weibo. Therefore, the traditional cascade mining is difficult to obtain the complete short video relationships. Meanwhile, the more complex network structure provides challenges for short video potential relationship mining, and it is difficult for traditional graph to mine the global information of the nodes, which restricts the learning of video potential relationships and styles. For the feature aspect, the above methods are overly dependent on external features, and the features of the video itself only remain in the work of extraction and splicing using pre-trained models, without aligning the relationships of the modalities. In addition, due to the specificity of the fake news detection, there is not enough labeled data for experimental analysis. Overall, the **motivation** lies in design a model that can both deeply mine large-scale video content features and extract video social relationship features.

To solve the above issues, this paper designs a **M**ultimodal **U**nified Social **H**ypergraph **C**ontrastive Enhancement method (**MUHC**) for Fake News Video Detection. To address the first challenge, this paper designs a unified social hypergraph establishment method for short videos. Specifically, unlike propagation graph, this paper combining ordinary graph and hypergraph based on the similarity attributes and event attributes between videos for representing the relationships between short videos. This approach can represent social relationships with different graph structures more effectively and reduces the effect of redundant edges. For the second challenge, this paper designs a multimodal contrastive learning method. Combining the unified social hyoergraph constructed above, intra-modal is performed by data augmentation of different modal features, and then multimodal consistency alignment is performed for inter-modal features for inter-modal contrastive learning and network tuning. Finally, fine-tuning is performed to learn the video semantics based on the existing tags. Overall, the main contributions of this work are as follows.

- **Unified Social Potential Hypergraph.** Based on video social relationships, a novelty unified social potential hypergraph is built for addressing the shortcomings of graph structure, and a more effective short video potential hypergraph representation is obtained.
- **Multimodal Contrastive Learning with Hypergraph.** Combined with the above unified hypergraph, a multimodal contrastive learning method is designed for intra-modal and inter-modal feature alignment and enhancement, and the self-supervised learning method lays the foundation for the extension of unlabeled datasets.
- **Better Performance.** Experimental results show that the method proposed in this paper outperform the existing fake news detection methods on the benchmark dataset and achieves state-of-the-art.

2 Related Works

To understand the existing works, this section investigates the application of graph representation learning and multimodal contrastive learning in fake news detection.

2.1 Graph Representation Learning

Fake news detection methods based on graph representation learning mainly focus on the propagation structure. Vosoughi et al. explored the diffusion patterns of information on Twitter, revealing that misinformation tends to spread more extensively than real news(Vosoughi et al., 2018). Zhou et al. subsequently detailed the characteristics of how fake news propagates(Zhou and Zafarani, 2019). Moreover, the GCNFN model enhances comment embedding by incorporating users' profiles(Monti et al., 2019), while the UPFD model considers users' past posts to reflect their inherent biases(Dou et al., 2021). These approaches demonstrate robust detection capabilities but require high-quality social network data. Furthermore, some researchers have focused on leveraging graph-structured data. Wang et al. introduced KMGCN, a framework that employs graph convolution networks for extracting textual features(Wang et al., 2020). Rosenfeld et al. developed a Weisfeiler-Lehman graph kernel that is independent of text, user, and time inputs, illustrating that structural encoding of information cascades can significantly aid in assessing the credibility of news(Rosenfeld et al., 2020). However, the above methods can rely on relatively few features, and the detection efficiency is lower than that of multi-dimension feature methods. To mine the potential social relationships of news, Qi et al. designed a short video social graph characterization method based on correlating with neighbors, which achieved good results.(Qi et al., 2023b) But the method is augmented by debunk videos, which is

not able to achieve the purpose of early detection. In addition, inspired by hypergraphs(Feng et al., 2019; Gao et al., 2022), Sun et al. designed HGSL for learning user spreading behavior(Sun et al., 2023), but still have problems.

2.2 Multimodal Contrastive Learning

Contrastive learning has been widely used in multimodal representation and learning(Qian et al., 2022). Radford et al. designed a text-image pre-training model using contrast learning, CLIP(Radford et al., 2021). Based on this, Zhou et al. proposed a CLIP-guided learning multimodal fake news detection method for image-text consistency detection(Zhou et al., 2023). Moreover, leveraging different contrastive properties, ALBEF(Li et al., 2021) introduced a strategy for mining hard negatives based on the distribution of contrastive similarities, while BLIP(Li et al., 2022) and VLMo(Bao et al., 2022) implemented a comparable approach. To deepen the exploration of contrastive information, Chen et al. addressed the cross-modal ambiguity learning challenge from an information theory standpoint, adopting specific methods for detection depending on the ambiguities present across various modalities(Chen et al., 2022). Based on this, Wang et al. established a cross-modal contrastive learning fake news detection method, which realized more accurate image-text alignment(Wang et al., 2023). However, the above method mainly focuses on the graphic domain and mainly compares the traditional modalities with limited learnable features. To solve above problems, Yin et al. designed a contrast loss and graph autoencoder to effectively learn the potential features of the propagation graph(Yin et al., 2024). However, it still did not solve the problem of contrast relationship between traditional modalities and social relation modalities.

3 Paramilitary

3.1 Problem Definition

Definition 1 (Unified Hypergraph) *We define the unified hypergraph $U = (\mathcal{V}, \mathcal{E})$ as an extended graph structure where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges. Each edge $e \in \mathcal{E}$ is a non-empty subset $e \subseteq \mathcal{V}$ such that $|e| \geq 2$.*

Problem 1 (Unified Social Potential Hypergraph) *Define the **Unified Potential Relationship Hypergraph** as $\mathcal{H} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W}, \boldsymbol{H})$, where $\mathcal{V}$ represents the nodes of the hypergraph. The hyperedges, $\mathcal{E} \subseteq \mathcal{P}(\mathcal{V})$, represent the complex relationships among videos and events, where $\mathcal{P}(\mathcal{V})$ denotes the power set of $\mathcal{V}$. The node attributes $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N \in \mathbb{R}^{N \times D}$, describe the features of the videos, with N being the number of videos/nodes and D the dimension of the attribute features. The adjacency matrix $\boldsymbol{H} \in \mathbb{N}^{|\mathcal{V}| \times |\mathcal{E}|}$ represents the connectivity within the hypergraph, reflecting the relationships among nodes. Further details see section 4.*

Problem 2 (Fake News Video Detection) *The video dataset is defined as $V = \{V_1, V_2, \ldots, V_n\}$, where each V_i is a video instance. Each instance V_i is modeled as a video potential relationship processes. We aim to learn an self-supervised function, g, defined as:*

$$g : V \rightarrow Y,$$

where V represents video instances with their potential relationship processes and $Y \in \{Fake, Real\}$ denotes fake or real videos.

3.2 Video Unimodal Features Representations

Short videos have rich modalities, such as text, audio, image and frame et al. In social networks, short videos also have social contexts (which are also named metadata), such as user information and comments.

For modal extractions, this paper use pre-trained models, and the specific extraction model is shown in Figure 2(a).

4 Methodology

This section introduce the proposed method **MUGC** in detail. As shown in Figure 2, we first perform feature extraction and aggregation of different modalities of the video. After that, the unified social potential hypergraph is built based on the video **Social Link** and **Corr Link**, and the hypergraph structure is embedded with the representation. Furthermore, intra-modal and inter-modal contrastive learning is performed on the representations to perform relevance enhancement of the different features, with the pruning optimization of network parameters. Finally, the network is tuned by labeled data to complete the supervised fake news video detection. The specific details and computational process are shown below.

4.1 Video Feature Integration

There may be correlations between different modals of short video. This subsection follows

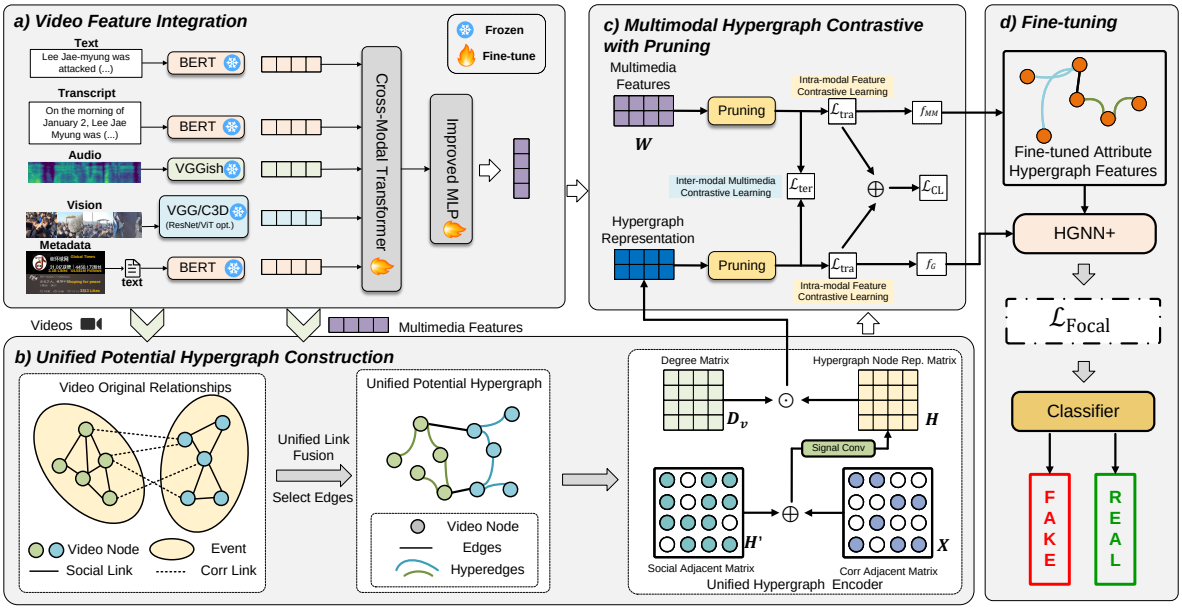

Figure 2: Framework of MUHC. The whole framework is divided into four modules.

the feature extraction method of Section 3.2, and expanded from co-attention(Lu et al., 2016) to a **cross-modal transformer** method. For consistency modeling for different modalities, which is used to learn the differences and correlations between different modalities. The specific equation is eq(1).

$$\widetilde{h_{x1}}, ... \widetilde{h_{xn}} = \text{cross-modal-transformer}\{h_{x1}, ..., h_{xn}\}, \tag{1}$$

where $h_{x1}, ..., h_{xn}$ is the original representation of the different modes, $\widetilde{h_{x1}}, ... \widetilde{h_{xn}}$ is the computed new representation with modal alignment. The specific combining method is shown in Fig. 2.

After computation, the corresponding new modal information is aggregated using the transformer to obtain the corresponding modal aggregation representation as shown in equation 2.

$$\mathbf{X} = \text{MEAN}\{\widetilde{h_T}, \widetilde{h_O}, \widetilde{h_A}, \widetilde{h_I}, \widetilde{h_F}, \widetilde{h_M}\}, \tag{2}$$

where MEAN is an average computation based on the modal representation, which is computed and then passed into a modified MLP for vector representation and parameter augmentation.

4.2 Unified Social Potential Hypergraph Construction

In this section, this paper builds a special kind of hypergraph, Unified Hypergraph, for representation of different relations. For Social Link, hyperedge is used for the representation, and for Corr Link, the ordinary graph edge is used. Finally the potential hypergraph representation of short videos with stronger performance combining the hyperedge and the ordinary edge is built.

4.2.1 Social Link

When a outbreak news appears, short video platforms will have amount of videos of the same type in a very short period of time, which is defined as an event. Different videos in the same event are fake and real, but generally there will be a strong correlation. The correlation of the same event is defined as a Social Link. We build a hypergraph of the Social Link with hyperedges, where all the video nodes of the same event are connected by a single hyper edge. The set of these hyperedges is defined as $\mathcal{E}_s$. Equation (3) indicates the method of hypergraph building.

$$\begin{cases} \mathbf{w}_c = \text{copy}\left(\text{sigmoid}\left(w_c\right), M_c\right) \\ \mathbf{W} = \text{diag}\left(\mathbf{w}_1^1, \cdots, \mathbf{w}_1^{M_1}, \cdots, \mathbf{w}_C^1, \cdots, \mathbf{w}_C^{M_C}\right), \\ \mathbf{H}' = \mathbf{H}'_1 \,\|\mathbf{H}'_2\| \cdots \|\mathbf{H}'_C \end{cases} \tag{3}$$

where $\mathbf{w}_c \in R$ represent a hyperparameter and all hyperedges within a specific hyperedge group c. The $\text{sigmoid}(\cdot)$ function is used for element-wise normalization. The representation $\mathbf{w}_j = \left(\mathbf{w}_C^1, \cdots, \mathbf{w}_C^{M_c}\right) \in R^{M_c}$ denotes the weight embedding for hyperedge group c. The $\text{copy}(a, b)$ function generates a embedding of size b, with the

value a replicated b times. The total number of hyperedges, M, is given by the sum $M_1 + M_2 + \cdots + M_c$. The weight matrix $\mathbf{W} \in R^{M \times M}$ is diagonal, with each diagonal entry $\mathbf{W}^{ii}$ representing the weight of hyperedge e_i. The incidence matrix $\mathbf{H}' \in \mathbb{N}^{N \times M}$ is formed by concatenating the incidence matrices of various hyperedge groups.

4.2.2 Corr Link

However, for the above hypergraph building method, events cannot be linked to each other. Therefore, we built Corr Link, which establishes a link based on the similarity between videos and videos, and is calculated using the Minkowski distance formula, which is shown below:

$$\operatorname{sim}(\mathbf{v}_i, \mathbf{v}_j) = (\sum_{k=1}^{D} |\mathbf{x}_{i,k} - \mathbf{x}_{j,k}|^D)^{1/D}, \tag{4}$$

where $x_{i,k}$,$x_{j,k}$ are representations of different videos.

However, a huge number of correlation edges can be obtained through the equation. To ensure the stronger correlation between different videos, this paper ensures that each node can select at most K edges. The specific equation is as follows.

$$\mathcal{E}_w = \left\{ (\mathbf{v}_i, \mathbf{v}_j) \mid \operatorname{sim}(\mathbf{v}_i, \mathbf{v}_j) \text{ in top-K of } [\operatorname{sim}(\mathbf{v}_i, \mathbf{v}_j)]_{p=1}^{N} \right\}, \tag{5}$$

where K is a trainable hyperparameter. Note that the following relationship $\forall \mathbf{v}_i \in \mathcal{V}$ holds.

4.2.3 Unified Link Fusion

Combining the hyperedges created above with ordinary edges results in a newly defined Unified Hypergraph, and the set of edges can be denoted by $\mathbf{H}$.

$$\mathbf{H} = (\mathbf{H'}_1 \| \mathbf{H'}_2 \| \cdots \| \mathbf{H'}_C) \oplus (\mathbf{A}_1 \| \mathbf{A}_2 \| \cdots \| \mathbf{A}_C), \tag{6}$$

where $\mathbf{H'}_1 \| \mathbf{H'}_2 \| \cdots \| \mathbf{H'}_C$ denotes hyperedges set. $\mathbf{A}_1 \| \mathbf{A}_2 \| \cdots \| \mathbf{A}_C$ denotes corr edges set.

The graph is a special form of hypergraph, to better characterize the above graph structure, this paper introduces an iterative version of the hypergraph convolution method, HGNN+, which can be efficiently fused for different nodes and hyperedges, and is calculated as follows.

$$\mathbf{X}^{t+1} = \sigma\left(\mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D}\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{D}_v^{-1/2}\mathbf{X}^t\mathbf{\Theta}\right), \tag{7}$$

where parameter $\mathbf{\Theta}$ is learnable. This filter $\mathbf{\Theta}$ is used on the nodes of the hypergraph to extract features. and $\mathbf{D}_v$ is the diagonal matrix of degrees.

After the above equations are calculated, a unified social potential hypergraph representation, named h_{UG}, can be obtained and used as input for subsequent contrastive learning task.

4.3 Multimodal Hypergraph Contrastive Learning with Pruning

4.3.1 Pruning

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).

4.3.2 Intra-modal Contrastive Learning

Contrastive learning is an effective self-supervised learning method, and this and the next subsubsection will be devoted to the contrastive learning method used. Intra-modal contrastive learning is mainly applied to the analysis of correlations within similar modalities, where the original modal features are compared with the pruning enhanced features to calculate the contrast loss with the following equation:

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

where $\widetilde{z}$,$\widetilde{z}_p$ are the features before and after pruning. τ is temperature which is a hyperparameter.

4.3.3 Inter-modal Contrastive Learning

Inter-modal contrastive learning is mainly used for the comparison between the graph structure and the features of the video itself for the alignment between different modal data. All contrast losses are calculated and summed to get the inter-modal contrast loss.

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

4.3.4 Contrastive Loss Integration

To integrate contrast learning loss, a hyperparameter λ is set for joint contrast loss to use for tuning the network.

$$\mathcal{L}_{CL} = \lambda\mathcal{L}_{\text{tra}} + (1-\lambda)\mathcal{L}_{\text{ter}}. \tag{10}$$

4.4 Fine-tuning

To obtain the final model prediction structure, a fine-tune attribute graph classification method is designed in this paper. Both graph structure and feature representation are derived from the comparison learning trained structure, and a more robust loss function Focal is used for model training inference and prediction.

$$\mathcal{L}_{\text{focal}} = -\frac{1}{|\mathcal{V}_l|} \sum_{i \in \mathcal{V}_l} \sum_{c=0}^{C} \alpha_c y_{ic} \left(1 - \hat{y}_{ic}\right)^{\gamma} \log\left(\hat{y}_{ic}\right), \quad (11)$$

where α, γ are learnable hyperparameters.

5 Experiments and Analysis

In this section, we will design extensive experiments to measure the effectiveness of our proposed method MUGC. The following 3 research questions are proposed for subsequent experimental investigations.

- **RQ1**. How does MUGC perform in the fake news video detection?
- **RQ2**. What is the performance of the proposed unified hypergraph and contrastive learning method?
- **RQ3**. What is the interpretability significance of the method proposed and why not use other modules to solve the task?

5.1 Dataset

Due to the lack of datasets, this paper adopts the **only** avaliable benchmark of short social network videos, FakeSV(Qi et al., 2023a). The dataset is obtained from Douyin and Kuaishou, which has a total of 738 events, 1827 real videos and 1827 videos are fake. The video has semantic features such as original text, audio, and visual, and social features such as comment and user.

5.2 Experiment Settings

5.2.1 Metrics

The experiments follow the criteria of benchmark dataset, also using 5-fold cross-validation with interval estimation. For the metrics, for fake news detection which is essentially a classification task, Accuracy, Recall, Precision, F1-score are used.

5.2.2 Parameter Setups

The experiments in this paper use AMD Ryzen Threadripper PRO 5995WX 64-Cores CPU and NVIDIA RTX A6000 as the experimental environment, and the maximum memory usage for data loading is 29G. textual feature extraction is performed using BERT-base-uncased, and the batch size is 64. Indicating that the features are 768, and the K parameter of Corr Link is set to 10. Other hyperparameters are analyzed in Appendix.

5.3 Baselines

In order to prove the superiority of the method, some of the more advanced algorithms are compared in terms of unimodal, multimodal and large language models as follows:

5.3.1 Uni-modal

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. In addition, Zhang et al. designed a news text detection method for news content and comments unified about emotions(Zhang et al., 2021).

5.3.2 Multi-modal

Existing multimodal methods mainly focus on text-image methods, such as EANN(Wang et al., 2018), and video semantic modal methods, such as Serrano et al.(Serrano et al., 2020) and FANVM(Choi and Ko, 2021). But there are fewer analysis that have both semantic features and social features. In this paper, we use the benchmark of FakeSV dataset, SV-FEND, as a baseline to compare with our proposed method(Qi et al., 2023a).

5.3.3 Large Language Model

With the development of the large language model, it is more capable of giving more positively correlated responses to linguistic information. To explore the application ability of GPT for the task of fake news detection, this paper designs the prompt method and conducts experiments through the APIs of GPT3.5-turbo (OpenAI, 2022) and GPT4.0-turbo (OpenAI, 2023), respectively. For GPT3.5, the text is inputted. For the corresponding multi-modal large model of GPT4.0, the content of text and video pumping frames are inputted to obtain positive incentive feedback.

Table 1: Experimental results of different methods. (Acc.: accuracy, Prec.: precision, Rec.: recall)

	Method	Acc.	Prec.	Rec.	F1
Uni-modal	Text(BERT)	77.14 ±2.75	77.18 ±2.73	77.12 ±2.72	77.10 ±2.76
	Audio(VGGish)	66.78 ±1.29	67.10 ±1.20	66.74 ±1.26	66.58 ±1.34
	Image(VGG19)	69.72 ±3.21	69.80 ±3.30	69.82 ±3.29	69.65 ±3.21
	Video(C3D)	69.59 ±1.77	70.07 ±1.81	69.56 ±1.75	69.38 ±1.77
	Text(Dual Emotion)	75.68 ±2.04	77.18 ±2.74	77.09 ±2.74	77.05 ±2.77
Multi-modal	Serrano et al.	68.72 ±2.30	70.75 ±2.11	68.73 ±2.30	67.92 ±2.53
	FANVM	76.39 ±1.14	75.41 ±1.54	73.73 ±1.24	74.19 ±1.18
	EANN	77.87 ±2.03	77.91 ±2.02	77.87 ±2.02	77.86 ±2.03
	SV-FEND	78.88 ±1.98	79.41 ±1.85	78.89 ±1.93	78.79 ±2.01
LLM	GPT3.5-turbo	57.49 ±3.39	60.82 ±4.12	57.48 ±2.94	58.42 ±3.24
	GPT4-turbo	65.05±1.98	66.02 ±2.13	65.05±2.16	65.31 ±2.06
	MUHC(Ours)	**89.82** ±1.33	**89.96** ±1.91	**89.82** ±1.81	**89.80** ±1.27
	Improve(%)	10.94 ↑	10.55 ↑	10.93 ↑	11.01 ↑

5.4 Experiment Performance (RQ1)

We completed the experiments with the above baselines and our method separately and performed the interval estimation. The experimental results are shown in Table 1. In the following, we will analyze the experimental results in depth.

Performance Analysis: The experimental results show that the experimental metrics of our method exceeds 89.5%, which is an improvement of 10% compared to the benchmark method. This result is compared with other unimodal, multimodal methods. In terms of unimodal methods, it can be seen that the modality that determines whether a short video is real or fake is mainly text, which can express more semantic than other features. In addition, although the method Zhang et al. (2021) proposed combines the news text features with the sentiment of the comments, it still does not have better results due to the semantic features. In terms of multi-modal methods, such as FANVM and EANN have a greater performance improvement than unimodal methods, but they still have a certain gap with our benchmark. For the method proposed in this paper, the main reason is that it is able to enhance the inter-modal features using contrastive learning while obtaining the features within the video modalities. Moreover, the short video social relationship mining innovatively proposed is able to better obtain the potential relationship between videos, which improves the feature mining ability.

Compare with LLM: Due to the limitation of data and resources, the common method of applying LLM at this stage is mainly prompt without fine-tuning. However, the LLMs are mainly generative models, which cannot achieve better results in the classification task of learning features mentioned above. As a result, the general metrics of the LLM are very low. The future work can consider the enhancement and expansion of the data to accomplish the fine-tuning for large language models.

5.5 Ablation Studies (RQ2)

In this subsection, the Unified Social Potential Hypergraph and the contrastive learning approach are ablated and analyzed, respectively. Besides, the practical implications of the above modules are explained below.

5.5.1 Study on Unified Social Potential Hypergraph

- **w/o PH.** the proposed unified social potential hypergraph is removed altogether, leaving only the basic features of the video and comparative learning.
- **w/o Social Link.** remove the hyperedge module and leave the PH module with only Corr Link aka similarity module.
- **w/o Corr Link.** remove the similarity module and leave only the hyperedge module.

Analysis: Experimental results can clearly prove that the unified social potential hypergraph has very important performance in fake news video detection based on social relationships. When all the potential hypergraphs are removed, it can be seen that the experimental metrics are reduced by about 1%, which is a significant difference. Additionally, Social Link and Corr Link both show an overall performance decrease of more than 1% when removed. Furthermore, to demonstrate the advantage of unified hypergraph performance, this experiment compares the inference speed and performance of general graph modeling. The results indicate that the unified hypergraph has a slight improvement in both inference speed and performance compared to graph.

Table 2: Ablation study on unified social potential hypergraph(PH) and its various components. HINT: Time is inference time.

Method	Acc.	Prec.	Rec.	F1	Time
Ours	**89.83**	**89.97**	**89.91**	**89.91**	1.0X
w/o PH	82.44	82.16	82.34	82.17	-
w/o Social Link	87.99	88.07	88.00	88.01	-
w/o Corr Link	88.84	88.80	88.88	88.85	-
w/o Hyperedge	89.14	89.36	89.22	89.22	0.8X

Table 3: Ablation result of Multimodal Hypergraph Contrastive Enhancement

Method	Acc.	Prec.	Rec.	F1
Ours	**89.82**	**89.96**	**89.82**	**89.80**
w/o intra-modal	88.83	88.80	88.89	88.82
w/o inter-modal	87.84	87.89	87.86	87.87

5.5.2 Study on Contrastive Learning

In this section we will analyze the multimodal contrastive learning methods. To prove that inter-modal and intra-modal methods are both effective, above modules are removed respectively, with the following details:

- **w/o intra-modal** remove intra-modal contrastive learning module, compute only the relationship between video features and hypergraph features.
- **w/o inter-modal** remove inter-modal contrastive learning module, only intra-modal relations are computed after data augmentation.

Analysis: Two contrastive learning methods are removed separately and it can be seen that there is a certain decrease about 2% in metrics after removed. This indicates that contrastive learning can better learn the relationship and modalities. However, it can be seen that in fact the metrics of contrastive learning are not particularly significantly improved compared to the potential hypergraph. The main reason for proposing the method is to propose a self-supervised learning short video social relationship detection mechanism. For unlabeled data, our proposed method can also enhance the performance through augmentation and comparison. The method is not limited to the FakeSV dataset, but can be extended to other related fields as well in the future.

5.6 Interpretability Studies (RQ3)

The above experiments show that some of the feature extraction techniques and methods used are might outdated, but the method is the best strategy obtained by combining the ROI analysis. In this subsection, we will take the vision module as an example to briefly analyze the performance metrics. The experimental results are shown in Table 4.

Table 4: Result of module feasibility and interpretability study. SVA. is F1-score of SV-FEND.

Method	SVA.	F1
Ours	**78.88**	**89.80**
w/ ResNet50	78.45	89.63
w/ ViT	78.46	89.74
w/ VST	78.12	89.06

For image, this paper selects ResNet(He et al., 2016) and ViT(Dosovitskiy et al., 2021) as feature extractors. Additionally, to demonstrate the method's validity, the video processing unit's C3D extracted content is replaced with Video Swin Transformer(Liu et al., 2022). It can be seen that the performance of the method in this paper slightly surpasses newer methods. Moreover, methods such as ViT and Video Swin Transformer have significantly more parameters compared to VGG19. Considering the return on investment, the aforementioned methods are not suitable for this framework. This proves that newer methods do not necessarily lead to better results.

6 Conclusion

This paper explores the establishment of potential relationships and the implementation of modal enhancement methods for short videos. Specifically, this paper designs a multimodal fake news detection algorithm based on unified social hypergraph enhancement. Separately, this paper proposes a potential social relationship hypergraph establishment without retweet information for short video social relationship mining. Among them, a Social Link building method based on hyperedge and a Corr Link building method based on ordinary edge are designed. The combination of both can be better for the representation of the proposed relationship hypergraph. Additionally, for multimodal data, we propose an intra-modal and inter-modal contrastive learning method for enhancing the relationships between different modal features and facilitating the extension to large-scale datasets. Experimental results show that the detection method proposed in this paper achieves sota results on the benchmark dataset with good efficiency.

Limitations

The dataset available for this method is currently limited; When performing the design of potential hypergraphs, the data needs to be labeled and designed according to the method proposed. Besides, the method may not work best in the multimodal data aggregation module. This paper only provides a benchmark method for data mining in the corresponding scenario, and its metrics are not guaranteed to be the highest. It should be noted that the feature extraction module of the method proposed in this paper can be discussed in depth, but it is not specifically built in this paper based on such task.

References

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Mu-Yen Chen, Yi-Wei Lai, and Jiunn-Woei Lian. 2023. Using deep learning models to detect fake news about covid-19. *ACM Transactions on Internet Technology*, 23(2):1–23.

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.

Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. 2021. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 148–157.

Hyewon Choi and Youngjoong Ko. 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*, pages 2950–2954.

Carmela Comito, Luciano Caroprese, and Ester Zumpano. 2023. Multimodal fake news detection on social media: a survey of deep learning techniques. *Social Network Analysis and Mining*, 13(1):101.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2051–2055.

Dina ElBoghdady. 2013. Market quavers after fake ap tweet says obama was hurt in white house explosions. *The Washington Post*, 23.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 33, pages 3558–3565.

Marc Fisher, John Woodrow Cox, and Peter Hermann. 2016. Pizzagate: From rumor, to hashtag, to gunfire in dc. *Washington Post*, 6:8410–8415.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. 2022. Hgnn+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3181–3199.

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.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 131–135.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 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*, pages 754–763.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. 2021. Self-damaging contrastive learning. In *International Conference on Machine Learning*, pages 4927–4939.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

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.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1165–1174.

Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. 2024. Fake news on social media: the impact on society. *Information Systems Frontiers*, 26(2):443–458.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Huyen Trang Phan, Ngoc Thanh Nguyen, and Dosam Hwang. 2023. Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, page 110235.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 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*, pages 14444–14452.

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

Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. 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*, pages 11947–11959.

Yiyue Qian, Chunhui Zhang, Yiming Zhang, Qianlong Wen, Yanfang Ye, and Chuxu Zhang. 2022. Comodality graph contrastive learning for imbalanced node classification. *Advances in Neural Information Processing Systems*, 35:15862–15874.

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

Nir Rosenfeld, Aron Szanto, and David C Parkes. 2020. A kernel of truth: Determining rumor veracity on twitter by diffusion pattern alone. In *Proceedings of The Web Conference 2020*, pages 1018–1028.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 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*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 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.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*.

Ling Sun, Yuan Rao, Yuqian Lan, Bingcan Xia, and Yangyang Li. 2023. HG-SL: Jointly learning of global and local user spreading behavior for fake news early detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5248–5256.

Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to

combat fake news. In *The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 275–284.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Longzheng Wang, Chuang Zhang, Hongbo Xu, Yongxiu Xu, Xiaohan Xu, and Siqi Wang. 2023. Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5696–5704.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 540–547.

Lianwei Wu, Pusheng Liu, and Yanning Zhang. 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, pages 13736–13744.

Shu Yin, Peican Zhu, Lianwei Wu, Chao Gao, and Zhen Wang. 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, pages 347–355.

Kaiwei Zhang, Junchi Yu, Haichao Shi, Jian Liang, and Xiao-Yu Zhang. 2023. Rumor detection with diverse counterfactual evidence. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3321–3331.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 354–367.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter*, 21(2):48–60.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo*, pages 2825–2830.

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*.

A Baseline Details

A.1 Uni-Modal (5 Methods):

- **BERT** (Devlin et al., 2019): applied for analyzing textual data extracted from video titles and transcripts.
- **VGGish** (Hershey et al., 2017): processes audio segments from videos to compute spectrograms, crucial inputs for the VGGish model.
- **VGG19** (Simonyan and Zisserman, 2015): used to encode individual video frames, capturing intricate visual details essential for precise video content analysis.
- **C3D** (Ji et al., 2013): combined with an MLP layer for generating predictions, particularly effective in recognizing temporal patterns in video sequences.
- **Dual Emotion**(Zhang et al., 2021): introduces dual emotion features to improve existing fake news detection systems by distinguishing between publisher and social emotions.

A.2 Multi-Modal (4 Methods):

- **Serrano et al.** 2020 extract TF-IDF features from video titles and the first 100 comments, employed in logistic regression and SVM classifiers.
- **FANVM** (Choi and Ko, 2021) aims to detect fake news videos across various topics, estimating topic distributions from video descriptions and comments.
- **EANN** (Wang et al., 2018) detects fake news using both image and text data, leveraging an event adversarial neural network to learn invariant features across different events.
- **SV-FEND** (Qi et al., 2023a) integrates multimodal data using cross-modal transformers to capture correlations among text, audio, visual data, and social contextual features.

A.3 Large Language Models (2 Methods):

- **GPT3.5-turbo**(OpenAI, 2022) is an optimized iteration of the GPT-3.5 model, engineered to deliver rapid responses without compromising text quality. Tailored for applications needing real-time natural language processing, it reduces latency and computational demands through fine-tuned efficiencies.
- **GPT-4-turbo**(OpenAI, 2023) expands on GPT-3.5-turbo's foundation by incorporating advanced training methodologies and a broader dataset. This evolution enhances its ability to produce nuanced and contextually precise outputs, making it ideal for addressing intricate conversational contexts and diverse linguistic tasks.

B Hyperparameter Details

To identify the optimal combination of model hyperparameters, the hyperparameters of corr link, feature dimension, contrastive learning, and pruning are thoroughly examined. The specific results are presented in Table 5.

To maximize potential relationship mining between videos, for selecting K for corr link, the range of $\{1, 5, 10, 15, 20, 25\}$ was considered in this study, and it was found that the best performance occurs at $K = 10$. To balance efficiency and performance, feature dimension was analyzed, with the range of $\{128, 256, 512, 768, 1024\}$ chosen. The best performance efficiency was observed at $D = 768$. In multimodal contrastive learning, to balance the weights of intra-modal and inter-modal comparisons, the hyperparameter λ was examined, and it was determined that $\lambda = 0.4$, indicating the best performance is achieved when intra-modal and inter-modal weights are balanced. During the pruning process, to find the optimal pruning ratio, the pruning ratio e was analyzed, and it was found that good performance was achieved at $e = 0.2$. Throughout the hyperparameter analysis, all control variables were maintained constant.

C Interpretability Note and Case Note

For relationships between videos, Qi et al. 2023b have explored that establishing potential relationships in the use of disinformation videos can help in the detection of false short videos. However, in daily life, it does not make sense to perform the detection of fake short videos after obtaining a debunk video. The importance of this work is to quickly perform the detection of whether a short video is fake or not based on the video features and potential relationships through the key nodes of the event outbreak without debunk.

In normal news outbreaks, most videos of the same event are in the same category. This can be very obvious in the FakeSV dataset, which is available at (Qi et al., 2023a) and will not be repeated

Table 5: Hyperparameter Analysis Results

Hyperparameter	Range	Optimal Value
Corr Link K	{1, 5, 10, 15, 20, 25}	10
Feature Dimension D	{128, 256, 512, 768, 1024}	768
Contrastive Learning λ	{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}	0.4
Pruning Ratio e	{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}	0.2

here. Therefore, discovering the nature of short videos based on events is more representative than learning features based on single video relationships.