IMOL: Incomplete-Modality-Tolerant Learning for Multi-Domain Fake News Video Detection

Zhi Zeng[1,2] **Jiaying Wu**[3] **Minnan Luo**[1,2*]
Herun Wan[1,2] **Xiangzheng Kong**[1,2] **Zihan Ma**[1,2]
Guang Dai[4] **Qinghua Zheng**[1,2]
[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China
[3]National University of Singapore, Singapore
[4]SGIT AI Lab, State Grid Corporation of China
zhizeng@stu.xjtu.edu.cn minnluo@xjtu.edu.cn

Abstract

While recent advances in fake news video detection have shown promising potential, existing approaches typically (1) focus on a specific domain (e.g., politics) and (2) assume the availability of multiple modalities, including video, audio, description texts, and related images. However, these methods struggle to generalize to real-world scenarios, where questionable information spans *diverse domains* and is often *modality-incomplete* due to factors such as upload degradation or missing metadata. To address these challenges, we introduce two real-world multi-domain news video benchmarks that reflect modality incompleteness and propose IMOL, an incomplete-modality-tolerant learning framework for multi-domain fake news video detection. Inspired by cognitive theories suggesting that humans infer missing modalities through cross-modal guidance and retrieve relevant knowledge from memory for reference, IMOL employs a hierarchical transferable information integration strategy. This consists of two key phases: (1) leveraging cross-modal consistency to reconstruct missing modalities and (2) refining sample-level transferable knowledge through cross-sample associative reasoning. Extensive experiments demonstrate that IMOL significantly enhances the performance and robustness of multi-domain fake news video detection while effectively generalizing to unseen domains under incomplete modality conditions. [1]

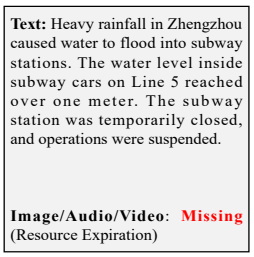
1 Introduction

As short videos increasingly dominate information dissemination (Walker and Matsa, 2021), the rapid spread of fake news videos calls for effective detection methods to safeguard public trust and information integrity. With the development of deep learning (Bi et al., 2024; Xiao et al., 2023; Wu et al., 2025; Li et al., 2025; Bi et al., 2025b; Ma et al., 2024; Du et al., 2025b; Bi et al., 2025a; Du et al., 2025a; Xiao et al., 2025), various fake news detection approaches have been developed, leveraging joint multimodal learning across diverse features, including video frames (Choi and Ko, 2021), visual-speech cues (Shang et al., 2021), social context (Qi et al., 2023a), creative process awareness (Bu et al., 2024), implicit opinions (Zong et al., 2024), and cross-modal consistency (Liu et al., 2023).

Despite the promising potential of existing efforts, they fail to adapt to the complex and ever-changing real-world scenario due to two key challenges. First, news environment is *multi-domain* such that videos typically span diverse domains such as politics, finance, and public health. These domains are often interconnected through shared concepts; for instance, financial crises can trigger political and societal unrest (Nan et al., 2021; Tong et al., 2024). Additionally, existing studies assume that all news video contain complete modalities, including video, audio, description text, and related images. However, real-world *news modalities are frequently incomplete* due to various reasons, including device malfunctions (Zhao et al., 2021), asynchronous modality availability (Shen et al., 2020), and poor video quality (Yuan et al., 2021). While preliminary efforts have explored learning from incomplete modalities in tasks such as emotion recognition (Xu et al., 2024; Guo et al., 2024), these methods are fundamentally limited as they

*Corresponding Author
[1]Our code is available at [GitHub Repo].

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30921–30933



Pattern 1: (T, MI, MA, MV)

(a) Resource Expiration

Pattern 2: (MT, MI, A,V)

(b) Blank Contents

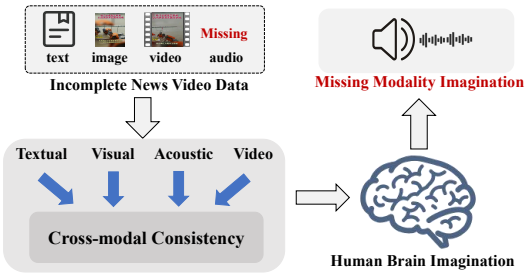
Pattern 3: (T, MA, I, V)

(c) Audio Corruption

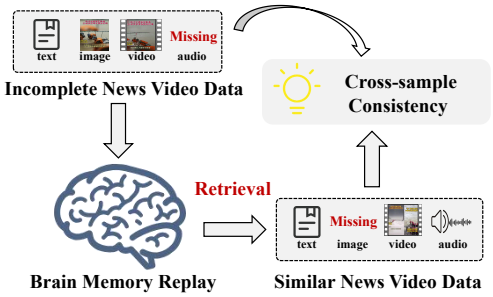
Pattern 4: (T, I, MA, MV)

(d) Contents Damage

Figure 1: Examples of modality patterns in news videos. Each news piece consists of four modalities: Video (V), Audio (A), Text (T), and Image (I). However, real-world news is often *modality-incomplete*, where one or more modalities may be missing (MV: Missing Video, MA: Missing Audio, MT: Missing Text, MI: Missing Image).



(a) Brain Cognitive Theory



(b) Brain Memory Replay Theory

Figure 2: Our approach is inspired by human cognitive processes and memory replay mechanisms. (a) Humans can reconstruct missing modalities by inferring information from available modalities. (b) When presented with a new piece of information, humans perform associative reasoning by retrieving and referencing similar knowledge from past experiences.

treat each sample independently and only leverage available modality information within individual instances. Given that cross-domain detection requires effective knowledge transfer across different news domains, these approaches suffer from severe generalization issues.

This paper introduces $\text{FakeSV}_{\text{IM}}$ and $\text{FakeTT}_{\text{IM}}$, two multi-domain fake news video detection benchmarks designed to evaluate the robustness of detection methods under modality-incomplete scenarios. To address these challenges, we propose an **I**ncomplete-**MO**dality-tolerant **L**earning (**IMOL**) framework for multi-domain fake news video detection. Drawing inspiration from cognitive theories suggesting that humans infer missing modalities based on available ones (Wei et al., 2022; Liu et al., 2024) (Figure 2(a)) and leverage cross-sample knowledge for associative reasoning (Spens and Burgess, 2024) (Figure 2(b)), IMOL emulates this reasoning process. Specifically, it conducts *cross-modal consistency learning* to exploit modality-invariant knowledge, enabling the reconstruction of missing modalities in news videos. Additionally, to enhance associative reasoning among conceptually correlated news samples, it employs *retrieval-augmented contrastive learning*, which preserves domain-invariant knowledge for improved generalization to unseen domains. By jointly leveraging cross-modal consistency and cross-sample transferable knowledge, IMOL effectively adapts to the complexities of the diverse and ever-evolving real-world news environment.

Our main contributions are as follows:

- We investigate the challenge of modality incompleteness in news video datasets and introduce the first benchmarks for multi-domain fake news video detection under incomplete modalities.
- We propose an incomplete-modality-tolerant learning framework named IMOL, which effectively handles modality-incomplete scenarios. IMOL leverages cross-modal consistency to reconstruct missing modalities and cross-sample reasoning to refine transferable knowledge for improved detection.
- Through extensive experiments on two real-world fake news video detection benchmarks,

we demonstrate that IMOL consistently outperforms competitive baselines in terms of detection effectiveness, robustness to varying modality incompleteness rates, and generalization to unseen news domains.

2 Related Work

2.1 Fake News Detection

Fake News Video Detection Fake news video detection is an emerging task that primarily focuses on identifying misinformation on short video platforms (Hou et al., 2019; Palod et al., 2019; Serrano et al., 2020; Wu et al., 2024). To address this challenge, FANVM (Choi and Ko, 2021) employed adversarial networks to extract topic-agnostic features, and TikTec (Shang et al., 2021) utilized captions to capture misinformation patterns in video content. TwtrDetective (Liu et al., 2023) aligned cross-media information for consistency-based detection. Further advancements include SV-FEND (Qi et al., 2023a), a Transformer-based model designed for multimodal feature integration, and NEED (Qi et al., 2023b), which employs Graph Attention Networks to achieve context-aware detection. Additionally, FakingRecipe (Bu et al., 2024) focused on identifying editing patterns indicative of misinformation, while MMVD (Zeng et al., 2024) addressed bias in fake news identification. Moreover, Zong et al. (2024) used prompt and diffusion learning to explore implicit opinions evolution of fake news video detection. However, existing methods typically focus on a single domain and assume modality completeness, making them ineffective in complex real-world scenarios that are inherently multi-domain and modality-incomplete. In contrast, our IMOL approach jointly addresses both challenges by integrating cross-modal and cross-sample transferable knowledge, enabling more robust and adaptable fake news video detection.

Multi-domain Fake News Detection News instance in the real world often spans multiple domains, such as science and disaster. Multi-domain learning (Li et al., 2024) aims to model such data simultaneously, improving both individual and overall domain performance. Approaches like MDFEND (Nan et al., 2021) and KG-MFEND (Fu et al., 2023) have achieved notable success in multi-domain fake news detection. Zhu et al. (2022a) extracted semantic, sentiment, and stylistic features, aggregating them via a domain adaptation module,

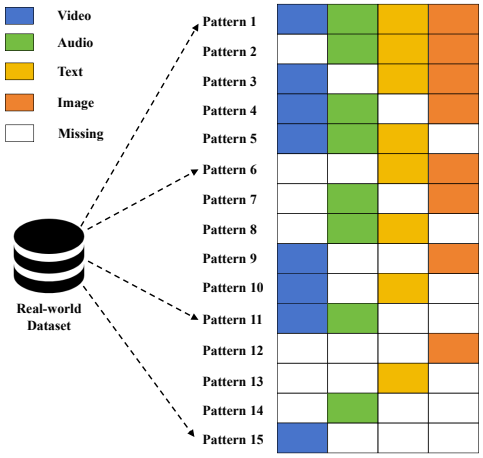
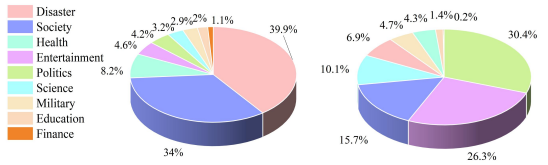


Figure 3: Examples of incomplete modality patterns.

while MMFEND (Tong et al., 2024) enhanced detection by incorporating visual information. These approaches primarily focus on textual fake news and fail to account for the multi-modality of real-world fake news content. To bridge this gap, we construct two multi-modal fake news video detection benchmarks and introduce the IMOL approach that integrates video, audio, images, and text for more comprehensive detection.

2.2 Incomplete Multimodal Learning

Incomplete modalities present significant challenges to multimodal learning, requiring models to handle incomplete information while maintaining performance. Several methods (Cai et al., 2018; Zhao et al., 2021; Liu et al., 2024) generate incomplete modalities using available modalities. Xu et al. (2024) proposed a model for robust joint representations, predicting missing modalities. Additionally, Guo et al. (2024) introduced multimodal prompt learning to generate missing features and improve incomplete learning with lower computational cost. Moreover, CAS-FEND (Nan et al., 2023) exploited incomplete social context information during the early detection of fake news. Existing approaches leveraged available modality knowledge within the sample, ignoring sample-level correlations while multimodal news videos usually belong to the same event or domain and have cross-sample associative consistency. In our study, our IMOL applies retrieval-augmented contrastive learning to learn cross-sample transferable knowledge for incomplete modality refinement.



(a) FakeSV$_{IM}$ (b) FakeTT$_{IM}$

Figure 4: Distributions of nine domains on two datasets.

(a) Health (b) Education (c) Financial

(d) Military (e) Science (f) Society

(g) Entertainment (h) Disaster (i) Politics

Figure 5: Word cloud of domains on FakeSV$_{IM}$ dataset.

3 Dataset Construction and Analysis

Based on FakeSV (Qi et al., 2023a) and FakeTT (Bu et al., 2024), two commonly adopted fake news video detection benchmarks, we construct our FakeSV$_{IM}$ and FakeTT$_{IM}$ benchmarks by (1) introducing incomplete modality patterns and (2) annotating domain labels. To simulate real-world modality incompleteness, we follow the study in (Lian et al., 2023) and randomly discard different modality combinations while ensuring that at least one modality remains available for each news sample. Consequently, a modality-incomplete dataset with M modalities contains (2^M-1) different missing modality patterns. In our case, with $M = 4$ (i.e., video, audio, text, and image), the dataset includes 15 distinct missing modality patterns, as illustrated in Figure 3.

We annotate data across nine domains: Disaster, Society, Health, Entertainment, Politics, Science, Military, Education, and Finance, with statistical distributions shown in Figure 4. Intuitively, news videos from different domains exhibit distinct topics and word usage. To illustrate this, we generate

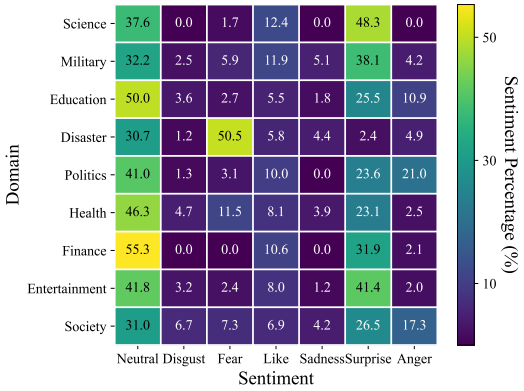


Figure 6: Heat map of sentiment tendency of nine domains on FakeSV$_{IM}$ dataset

word clouds representing domain-specific vocabulary frequency, as shown in Figure 5. Additionally, we observe domain-specific sentiment tendencies: for instance, publishers in the Science domain predominantly express surprise, whereas those in the Disaster domain often convey fear and sadness, as depicted in Figure 6.

4 Preliminary

4.1 Problem Definition

Given a multi-domain news video dataset $\mathcal{D} = (x, y, d)$ with four modalities: text, image, video, and audio, each news video is represented as $x = (x^t, x^i, x^v, x^a)$, where x^t, x^i, x^v, and x^a denote the features of each modality. Each news video is assigned a ground-truth label $y \in \{0, 1\}$, indicating whether the news is real ($y = 0$) or fake ($y = 1$), and a domain label $d \in \{\text{domain}_1, \ldots, \text{domain}_K\}$, where K is the number of domains. To account for incomplete modalities, we define x^{tm}, x^{im}, x^{vm}, and x^{am} to indicate which modalities are missing. Given a news video x and its domain label d, the task of multi-domain fake news video detection under incomplete modalities aims to identify whether the news is fake or not.

4.2 Domain-Informed News Modeling

Multimodal Feature Encoding To ensure fair evaluation, we apply the encoders following the previous work (Qi et al., 2023a). We leverage news titles and video captions and then feed them into a BERT (Devlin et al., 2018) to extract the textual feature, denoted as x^t. To analyze the visual content of keyframes and thumbnails, we employ a pre-trained VGG19 model (Simonyan and Zisserman, 2014) to extract the image feature, denoted

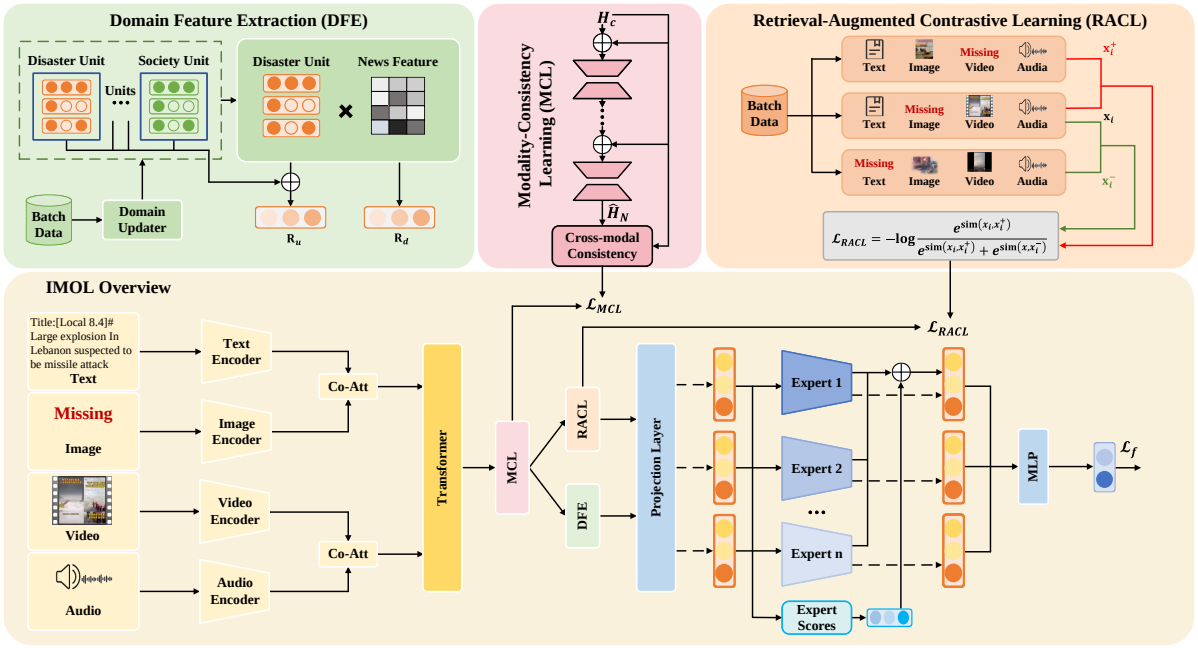


Figure 7: The overview of incomplete-modality-tolerant learning framework.

as x^i. To analyze motion trajectories within videos, we leverage a pre-trained C3D model (Tran et al., 2015) for feature extraction and apply an averaging operation to aggregate the video feature, denoted as x^v. We isolate the audio track from videos (Li et al., 2022) and employ a pre-trained VGGish model (Hershey et al., 2017) to extract the audio feature, denoted as x^a.

Assume an incomplete modality sample $x = (x^t, x^{im}, x^v, x^a)$, we apply a zero vector to represent the feature of incomplete modality x^{im}, which is a standard operation followed by (Liu et al., 2024). To model the mutual enhancement between different modalities, we apply cross-attention between news text and image features, as well as between video and audio features. We then integrate x^t, x^{im}, x^v, x^a as:

$$\mathbf{h} = [x^t \oplus x^{im} \oplus x^v \oplus x^a]. \tag{1}$$

Domain Feature Encoding To capture domain correlations of the news videos, we employ a domain unit to aggregate them, defining the set as $\mathbf{M} = \{\mathbf{m}_j\}_{j=1}^K$, where $\mathbf{m}_j$ represents $j-$th domain unit. Given the feature of a news video $\mathbf{h}$, we derive its domain-specific representation $\mathbf{d}_j$ as:

$$\mathbf{d}_j = \text{softmax}\left(\frac{\mathbf{h} \cdot \mathbf{W}_d \cdot \mathbf{m}_j{}^\top}{\tau}\right)\mathbf{m}_j, \tag{2}$$

where $\mathbf{W}_d$ is the learnable weight, updating domain representation and τ is a scaling factor. The aggregated domain representations are then concatenated into $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_K]$. While a news video may ambiguously belong to multiple domains, known as domain shift (Zhu et al., 2022b), we explore domain-shift information by calculating similarity distribution $\mathbf{v}$ between $\mathbf{h}$ and $\mathbf{D}$ as:

$$\mathbf{v} = \text{softmax}\left(\mathbf{h} \cdot \mathbf{W}_v \cdot \mathbf{D}^\top\right). \tag{3}$$

Here, $\mathbf{W}_v$ is the learnable weight. Then, with the similarity distribution $\mathbf{v}$, we evaluate domain-shift representation as:

$$\mathbf{R}_u = \sum_{j=1}^{K} \mathbf{v}_j \cdot \mathbf{d}_j, \tag{4}$$

where $\mathbf{d}_j$ is the of $j-$th domain-specific representation, and $\mathbf{v}_j$ is the corresponding similarity weight. For a given news video, we retrieve its domain-specific representation $\mathbf{R}_d = \mathbf{d}_j$.

5 Methodology

5.1 Model Overview

Our framework aims to achieve robust multi-domain fake news video detection under incomplete modalities. The framework is illustrated in Figure 7. The ultimate goal of our IMOL is to refine cross-modal and cross-sample transferable knowledge, enhancing model's robustness and generalization capabilities in downstream task. To achieve this, we integrate modality-consistency learning

and cross-sample consistency learning based on contrastive learning. The overall optimization objective $\mathcal{L}_f$ can be expressed as:

$$\mathcal{L}_f = \alpha\mathcal{L}_{MCL} + \beta\mathcal{L}_{RACL} + \mathcal{L}_C, \tag{5}$$

where $\mathcal{L}_{MCL}$ encourages the cross-modal transferable knowledge via modality-consistency learning, $\mathcal{L}_{RACL}$ promotes cross-sample transferable knowledge by contrastive learning, and $\mathcal{L}_C$ ensures to learn comprehensive representations. The α, β are hyper-parameters.

5.2 Modality Consistency Learning ($\mathcal{L}_{MCL}$)

Inspired by brain cognitive theory (Wei et al., 2022), we exploit the correlations of available modalities to imagine the feature of incomplete modality. To guide the model to imagine the incomplete modality, we map $\mathbf{h}$ into a shared semantic space by projection operation, formulated as:

$$\mathbf{H}_c = g(\mathbf{h}), \tag{6}$$

where $g(\cdot)$ is a linear layer. To reconstruct the features of the missing modality, we employ a Residual AutoEncoder (RAE) (Tran et al., 2017). The encoded representation $\mathbf{H}_c$ is provided to each intermediate layer of the RAE as additional input, ensuring robust reconstruction by supplying invariant cross-modal information. The RAE consists of N autoencoders, denoted as $\alpha_i(\cdot), i \in \{1, \cdots, N\}$. The reconstruction process is expressed as follows:

$$\hat{\mathbf{H}}_i = \begin{cases} \alpha_i(\mathbf{H}_c), & i = 1, \\ \alpha_i(\mathbf{H}_c + \hat{\mathbf{H}}_{i-1}), & 1 < i \leq N, \end{cases} \tag{7}$$

where $\hat{\mathbf{H}}_i$ is the output of the i-th autoencoder and $\hat{\mathbf{H}}_N$ is the imagination features. The modality-consistency learning phrase exploits interactions of original and imagination samples by minimizing their discrepancy (Nielsen, 2015):

$$\mathcal{L}_{MCL} = \sqrt{\frac{1}{N_b}\sum_{i=1}^{N_b}(\mathbf{H}_c - \hat{\mathbf{H}}_N)^2}, \tag{8}$$

where N_b is batch size. Finally, we use $\hat{\mathbf{H}}_N$ to represent modal-complete feature $\mathbf{x}_i$ of $i-$th news.

5.3 Retrieval-Augmented Contrastive Learning ($\mathcal{L}_{RACL}$)

Based on the brain memory replay theory (Spens and Burgess, 2024), individuals retrieve and compare similar samples, leveraging their consistency for associative reasoning. Inspired by this, we propose a retrieval-augmented contrastive learning(RACL) strategy to explore cross-sample consistency between similar samples and enhance the cross-sample transferable knowledge capacity to handle incomplete modalities.

We construct two samples to help our model learn consistency between positive samples and specificity between negative samples.

1) **Gold positive samples**: These are most semantically similar to the original samples within the same labels in a batch. Comparing with gold positive samples, the model adjusts its reconstruction of the missing modality, improving alignment across similar samples (Mei et al., 2024). The selection process is formulated as:

$$\mathbf{x}_i^+ = \underset{\mathbf{x}_j \in \mathcal{B},\, y_i = y_j}{\arg\max}\ \mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j), \tag{9}$$

where $\mathcal{B}$ is the set of batch data and $\mathrm{sim}(\cdot)$ is the cosine similarity function.

2) **Hard negative samples**: These are most semantically similar to the original sample but with opposite labels. Such samples improve the semantic space's capacity to distinguish fake news:

$$\mathbf{x}_i^- = \underset{\mathbf{x}_j \in \mathcal{B},\, y_i \neq y_j}{\arg\max}\ \mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j). \tag{10}$$

Given $\mathbf{x}_i, \mathbf{x}_i^+$ as positive pairs (Gao et al., 2021):

$$\mathcal{L}_{RACL} = -\log \frac{e^{\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_i^+)}}{e^{\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_i^+)} + \sum_{\mathbf{x}_i^- \in \mathcal{X}_i} e^{\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_i^-)}}, \tag{11}$$

where $\mathbf{x}_i^-$ denotes the hard negative samples. $\mathcal{X}_i$ is the set of hard negative samples. Finally, we obtain the sample-level complete feature $\mathbf{R}_x = \mathbf{x}_i$.

5.4 Fake News Video Detection ($\mathcal{L}_C$)

To fully leverage comprehensive information of news across domains, we apply a mixture-of-expert strategy to adaptively integrate $\mathbf{R}_x, \mathbf{R}_d$, and $\mathbf{R}_u$:

$$\mathbf{P} = [\mathbf{p}_x, \mathbf{p}_d, \mathbf{p}_u] = \mathrm{MoE}(\mathbf{R}_x, \mathbf{R}_d, \mathbf{R}_u), \tag{12}$$

$$\mathbf{w}_i = \mathrm{softmax}(\mathbf{p}_i) = \frac{e^{\mathbf{p}_i}}{\sum_{j \in \{x,d,u\}} e^{\mathbf{p}_j}}, \tag{13}$$

$$\hat{y} = f_c(\sum_{i \in \{x,d,u\}} \mathbf{w}_i \cdot \mathbf{R}_i), \tag{14}$$

where $\mathrm{MoE}(\cdot)$ represents mixture-of-expert layers using multiple MLP layers and $\mathbf{w}_i$ denote the different expert scores. $f_c(\cdot)$ represents a Transformer

Dataset	Model	0		0.1		0.2		0.3		0.4		0.5		0.6		0.7	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
$\text{FakeSV}_{\text{IM}}$	TikTec	76.34	76.32	76.30	76.28	74.45	74.42	73.48	73.41	74.12	74.09	72.11	72.08	71.09	71.05	69.62	69.57
	FANVM	77.80	77.76	77.19	77.16	76.01	75.96	73.89	73.85	72.52	72.39	71.71	71.67	69.01	68.98	68.32	68.31
	SV-FEND	78.20	78.13	77.97	77.93	78.33	78.26	77.28	77.22	76.68	76.61	76.65	76.61	76.25	76.23	76.07	76.03
	MMVD	78.11	78.06	77.78	77.73	77.77	77.72	77.67	77.62	77.53	77.50	77.43	77.40	77.14	77.12	76.83	76.79
	MoMKE	78.72	78.67	78.47	78.40	78.44	78.39	78.20	78.10	77.59	77.55	77.45	77.38	77.34	77.33	75.78	75.73
	Ours	**81.42**	**81.36**	**80.79**	**80.73**	**80.02**	**79.99**	**79.77**	**79.69**	**79.36**	**79.33**	**78.38**	**78.33**	**78.36**	**78.31**	**77.19**	**77.12**
$\text{FakeTT}_{\text{IM}}$	TikTec	77.05	75.97	75.84	74.19	74.87	72.46	74.62	72.76	72.65	70.38	73.05	70.08	73.05	70.08	73.05	70.08
	FANVM	77.81	75.99	76.75	74.60	75.23	73.01	75.73	73.66	72.49	69.75	71.78	67.80	59.22	59.31	52.63	47.31
	SV-FEND	80.85	79.90	80.80	79.77	79.43	78.31	80.04	79.05	80.04	79.05	79.64	78.37	79.18	78.07	77.71	76.53
	MMVD	79.89	78.10	79.64	78.56	79.84	78.32	79.53	78.38	79.08	77.79	78.98	77.96	78.67	77.35	77.10	75.99
	MoMKE	81.61	80.93	81.36	80.51	80.90	80.02	79.89	78.94	79.13	77.47	79.99	79.03	79.13	77.90	77.46	75.93
	Ours	**83.33**	**82.75**	**82.57**	**81.85**	**81.51**	**80.62**	**81.00**	**80.33**	**80.75**	**79.83**	**80.14**	**79.26**	**79.23**	**78.36**	**79.23**	**78.36**

Table 1: Comparison of classification performance with different incomplete rates. For fair evaluation, we applied the same textual, visual, audio, video features of backbones and random seed for each comparison methods. **Bold**: best result. Underline: second best result. We report the five-fold average results with variances below 0.05.

encoder layer (Vaswani et al., 2017) with an MLP layer. The prediction $\hat{y}$ represents the likelihood of the input being fake news, which is supervised using the cross-entropy loss:

$$\mathcal{L}_C = \sum_{y \in Y^\ell} \big(-y \log(\hat{y}) - (1-y) \log(1-\hat{y})\big), \quad (15)$$

where Y^ℓ is the set of ground truth labels.

6 Experiment

In this section, we conducted experiments to evaluate the effectiveness of our model. Specifically, we aim to answer the following research questions:
RQ1: Can IMOL improve the performance of multi-domain fake news video detection?
RQ2: How does IMOL's computational complexity compare to other models?
RQ3: How effective are the different modules of IMOL in detection?
RQ4: Can IMOL enhance the domain generalization ability of fake news video detection?
RQ5: How does IMOL perform in temporal detection?
RQ6: How does IMOL compare to large language models (LLMs)?

Evaluation task settings and parameter analysis are shown in Appendix A.1 and A.3, respectively.

6.1 Performance Results (RQ1)

In Table 1, we compare the performance of our IMOL with other methods under varying missing rates. We observe that our model outperforms all other methods across all metrics under different missing rates. This demonstrates that our IMOL is robust to different missing rates and can effectively handle severely missing modalities. Multimodal methods such as SV-FEND, MMVD and MoMKE, show strong capabilities in modeling textual, visual, video and acoustic modalities under incomplete modalities. Building on this, our model goes further by modeling cross-modal and cross-sample transferable knowledge, allowing for imagining missing modalities and improving detection performance.

	IMOL	MoMKE	MMVD	SV-FEND
Params	758.66M	769.30M	730.89M	718.06M
FLOPS	13.51G	13.63G	11.97G	11.87G

Table 2: Comparison of trainable parameters and computational speed. FLOPs: amount of floating point arithmetics.

	$\text{FakeSV}_{\text{IM}}$		$\text{FakeTT}_{\text{IM}}$	
	ACC	F1	ACC	F1
IMOL	**77.19**	**77.12**	**79.23**	**78.36**
w/o Domain	75.38	75.32	76.55	75.36
w/o MCL	76.16	76.16	78.52	77.44
w/o RACL	76.44	76.38	77.71	76.81

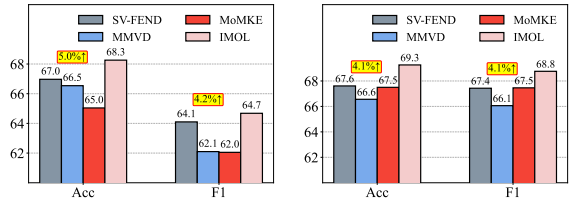
Table 3: Results of ablation study.

6.2 Computational Complexity (RQ2)

The computational complexity study in Table 2 shows that IMOL requires less trainable parameters compared to MoMKE. SV-FEND requires slightly low computational complexity, but the performance is not good enough.

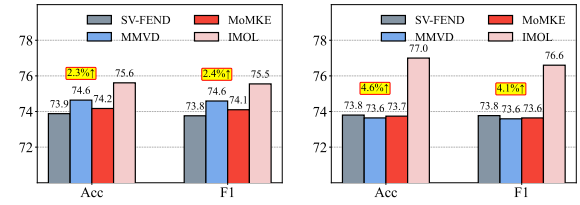
6.3 Ablation Study (RQ3)

We design ablation experiments with a modality missing rate of 0.7 to evaluate the effectiveness



(a) FakeSV$_{IM}$ (b) FakeTT$_{IM}$

Figure 8: Performance comparison of domain generalization analysis on two datasets.



(a) FakeSV$_{IM}$ (b) FakeTT$_{IM}$

Figure 9: Performance comparison of temporal analysis on two datasets.

of components in IMOL. Specifically, we design several internal models with inputs and strategies removed[2]. The results in Table 3 reveals that removing any components of the IMOL would lead to performance drop, while the domain information is most critical to performance, resulting in a 2.68% drop in accuracy on FakeTT$_{IM}$ dataset. As a result, cross-modal and cross-sample consistency learning strategies can help our IMOL imagine incomplete modalities for robust fake news detection.

6.4 Domain Generalization Analysis (RQ4)

To further verify the domain generalization ability of our model, we train the models with missing rate of 0.7 on Ch-6, En-6 and test them on Ch-3, En-3[3], respectively. As shown in Figure 8, in the Ch-3, En-3 datasets, the domain generalization performance of the SV-FEND, MMVD, and MoMKE are lower than our IMOL, indicating that our model can better generalize to unseen domains.

6.5 Temporal Analysis (RQ5)

To evaluate the temporal generalization ability of our IMOL model, we divided the datasets into training and testing sets in an 8:2 ratio based on the chronological order of video publication. This enabled temporal generalization evaluation of the model's ability to detect fake news over time. As illustrated in Figure 9, IMOL indicated superior performance in detecting fake news videos in a temporal sequence. This highlights the model's domain-specific and domain-shift information capabilities, which effectively contribute to its robustness in detecting fake news over time.

6.6 Comparison with LLMs (RQ6)

As the most popular LLMs, GPT-3.5 and GPT-4V (Achiam et al., 2023) has demonstrated unparalleled performance across various tasks, inspiring us to test them on this task. For fair evaluation, we used a zero-shot of our IMOL (trained on Ch-6, En-6, tested on Ch-3, En-3) while testing the GPT-3.5 and GPT-4V[4] on Ch-3, En-3. Table 4 shows that IMOL outperforms GPT-4V by 13% in terms of classification accuracy on En-3 dataset, demonstrating that task-specific smaller models can outperform general larger models in specific tasks.

	Ch-3		En-3	
	ACC	F1	ACC	F1
IMOL	**71.58**	**67.99**	**70.20**	**69.47**
GPT-3.5-turbo	63.54	65.03	50.89	60.85
GPT-4V	64.63	67.91	57.48	66.13

Table 4: Comparison results of IMOL, GPT-3.5-turbo and GPT-4V.

7 Conclusion

In this paper, we address two key challenges in fake news video detection: generalization across diverse news domains and handling incomplete modalities. To benchmark model effectiveness under these challenges, we introduce FakeSV$_{IM}$ and FakeTT$_{IM}$, two datasets designed to evaluate models in multi-domain, modality-incomplete scenarios. We further propose IMOL, an incomplete-modality-tolerant learning framework inspired by human cognitive processes and memory replay mechanisms. IMOL employs cross-modal consistency learning to reconstruct missing modalities and retrieval-augmented contrastive learning to capture transferable knowledge across samples. Extensive experiments validate IMOL's effectiveness, robustness to modality incompleteness, and generalizability to new domains.

[2]descriptions of ablation models are shown in Appendix A.2.

[3]descriptions of Ch-6, En-6, Ch-3, and En-3 are shown in Appendix A.1.2.

[4]prompts of GPT-3.5 and GPT-4V are shown in Appendix A.4.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2022YFB3102600), the National Nature Science Foundation of China (No. 62192781, No. 62272374), the Natural Science Foundation of Shaanxi Province (2024JC-JCQN-62), the National Nature Science Foundation of China (No. 62202367, No. 62250009), the Key Research and Development Project in Shaanxi Province No. 2023GXLH-024, Project of China Knowledge Center for Engineering Science and Technology, and Project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China", and the K. C. Wong Education Foundation.

Limitations

This study suffers limitations that may impact the performance of our proposed framework. While introducing retrieval-augmented contrastive learning strategies has achieved promising results in fake news detection, the performance of the retrieve samples may have the influence on the accuracy of fake news detection. Moreover, although the incomplete-modality-tolerant learning framework is effective in modeling cross-modal and cross-sample consistency for incomplete modalities imagination for multi-domain fake news video detection, extremely small prediction scores may result in an abundance of zero values, posing a risk of overfitting or gradient vanishing. We plan to address these limitations in future study.

Ethics Statement

This paper adheres to the ACM Code of Ethics and Professional Conduct. Firstly, the dataset utilized does not contain sensitive private information and poses no harm to society. Secondly, proper attribution is given to relevant papers and the sources of pre-trained models, along with detailed references to the toolkits used. Furthermore, our code will be released under the license of any artifacts used. Lastly, the proposed fake news video detection method is designed to contribute to the safety and stability of the internet environment and public opinion.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025a. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*.

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2024. Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. *arXiv preprint arXiv:2412.12359*.

Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, et al. 2025b. Cot-kinetics: A theoretical modeling assessing lrm reasoning process. *arXiv preprint arXiv:2505.13408*.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. *arXiv preprint arXiv:2407.16670*.

Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1158–1166.

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.

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

Guodong Du, Zitao Fang, Jing Li, Junlin Li, Runhua Jiang, Shuyang Yu, Yifei Guo, Yangneng Chen, Sim Kuan Goh, Ho-Kin Tang, Daojing He, Honghai Liu, and Min Zhang. 2025a. Neural parameter search for slimmer fine-tuned models and better transfer. *arXiv preprint arXiv:2505.18713*.

Guodong Du, Xuanning Zhou, Junlin Li, Zhuo Li, Zesheng Shi, Wanyu Lin, Ho-Kin Tang, Xiucheng Li, Fangming Liu, Wenya Wang, Min Zhang, and Jing Li. 2025b. Knowledge grafting of large language models. *arXiv preprint arXiv:2505.18502*.

Lifang Fu, Huanxin Peng, and Shuai Liu. 2023. Kgmfend: an efficient knowledge graph-based model for multi-domain fake news detection. *The Journal of Supercomputing*, 79(16):18417–18444.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736.

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 *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. In *2019 International conference on multimodal interaction*, pages 235–243.

Jiayang Li, Xuan Feng, Tianlong Gu, and Liang Chang. 2024. Dual-teacher de-biasing distillation framework for multi-domain fake news detection. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 3627–3639. IEEE.

Xiang Li, Chaofan Fu, Zhongying Zhao, Guanjie Zheng, Chao Huang, Yanwei Yu, and Junyu Dong. 2025. Dual-channel multiplex graph neural networks for recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Xiaojun Li, Xvhao Xiao, Jia Li, Changhua Hu, Junping Yao, and Shaochen Li. 2022. A cnn-based misleading video detection model. *Scientific Reports*, 12(1):6092.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432.

Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. 2023. Covid-vts: Fact extraction and verification on short video platforms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 178–188.

Rui Liu, Haolin Zuo, Zheng Lian, Bjorn W Schuller, and Haizhou Li. 2024. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5333–5347.

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.

Qiong Nan, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Guang Yang, and Jintao Li. 2023. Exploiting user comments for early detection of fake news prior to users' commenting. *arXiv preprint arXiv:2310.10429*.

A Nielsen. 2015. Neural networks and deep learning.

Priyank Palod, Ayush Patwari, Sudhanshu Bahety, Saurabh Bagchi, and Pawan Goyal. 2019. Misleading metadata detection on youtube. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*, pages 140–147. Springer.

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*, volume 37, pages 14444–14452.

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: ACL 2023*, pages 11947–11959.

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*.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE international conference on big data (big data)*, pages 899–908. IEEE.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM international conference on multimedia*, pages 493–502.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Eleanor Spens and Neil Burgess. 2024. A generative model of memory construction and consolidation. *Nature Human Behaviour*, 8(3):526–543.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1178–1186.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Mason Walker and Katerina Eva Matsa. 2021. News consumption across social media in 2021.

Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. 2022. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*.

Jiaying Wu and Bryan Hooi. 2023. Decor: Degree-corrected social graph refinement for fake news detection. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2582–2593.

Jiaying Wu, Fanxiao Li, Min-Yen Kan, and Bryan Hooi. 2025. Seeing through deception: Uncovering misleading creator intent in multimodal news with vision-language models. *arXiv preprint arXiv:2505.15489*.

Jiaying Wu, Shen Li, Ailin Deng, Miao Xiong, and Bryan Hooi. 2023. Prompt-and-align: prompt-based social alignment for few-shot fake news detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2726–2736.

Kaixuan Wu, Yanghao Lin, Donglin Cao, and Dazhen Lin. 2024. Interpretable short video rumor detection based on modality tampering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9180–9189.

Zikai Xiao, Zihan Chen, Songshang Liu, Hualiang Wang, YANG FENG, Jin Hao, Joey Tianyi Zhou, Jian Wu, Howard Yang, and Zuozhu Liu. 2023. Fed-grab: Federated long-tailed learning with self-adjusting gradient balancer. In *Advances in Neural Information Processing Systems*, volume 36, pages 77745–77757. Curran Associates, Inc.

Zikai Xiao, Ziyang Wang, Wen Ma, Yan Zhang, Wei Shen, Yan Wang, Luqi Gong, and Zuozhu Liu. 2025. Mitigating posterior salience attenuation in long-context llms with positional contrastive decoding. *arXiv preprint arXiv:2506.08371*.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. 2024. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 438–446.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4400–4407.

Zhi Zeng, Minnan Luo, Xiangzheng Kong, Huan Liu, Hao Guo, Hao Yang, Zihan Ma, and Xiang Zhao. 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*, pages 6492–6500.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618.

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

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

Linlin Zong, Jiahui Zhou, Wenmin Lin, Xinyue Liu, Xianchao Zhang, and Bo Xu. 2024. Unveiling opinion evolution via prompting and diffusion for short video fake news detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10817–10826.

A Appendix

A.1 Experimental settings

A.1.1 Evaluation Task Settings and Metrics

To ensure fairness of experiments, we applied the same textual, visual, audio, video features and random seeds for each comparison methods. Our

evaluation experiments are conducted by applying five-fold cross-validation experiments without overlapping events and the dataset is split as training and testing sets with a ratio of 4:1. We use maximum length of the news text as 100 and the Bert-based uncased (Devlin et al., 2018) for the datasets of FakeTT$_{\text{IM}}$ and the Bert-based-Chinese (Devlin et al., 2018) for FakeSV$_{\text{IM}}$ dataset, respectively. In the IMOL, we set $\tau = 0.01, \alpha = 0.1, \beta = 0.2$, and use Adam as the optimizer, learning rate as 1e-4, batch_size as 128, and training epochs as 50. According to the metrics Accuracy(ACC), Macro F_1(F_1), the detection performances of baselines and baselines combined with our IMOL are shown in Table 1. Their largest values are emphasized in bold. The experiments were conducted on NVIDIA 3090Ti GPUs.

A.1.2 Dataset

- **FakeSV$_{\text{IM}}$** : An extension of FakeSV (Qi et al., 2023a), the largest Chinese fake news video dataset, which was verified by official Chinese fact-checking sites between January 2019 and January 2022. In addition, to testify the domain generalization ability of our model under incomplete modalities, we sample two datasets as Ch-3 and Ch-6. Ch-3 contains 3 rare domains, including Military, Politics and Finance. Ch-6 contains common domains, including Disaster, Society, Health, Entertainment, Education and Science.

- **FakeTT$_{\text{IM}}$** : An extension of FakeTT (Bu et al., 2024), an English TikTok dataset covering multiple domains, which highlights its versatility for fake news detection. En-3, En-6 and Ch-3, Ch-6 are set to the same. The details of the datasets are shown in Table 5.

Datasets	FakeSV$_{\text{IM}}$	FakeTT$_{\text{IM}}$
# of fake news	1,810	1,172
# of real news	1,814	819

Table 5: Statistics of datasets

A.1.3 Baselines

To construct a comprehensive benchmark, we compare our model with current models. For these multimodal models, we follow the experimental settings in their studies (Choi and Ko, 2021; Shang et al., 2021; Qi et al., 2023a; Zeng et al., 2024; Xu et al., 2024), where we incorporate the domain features into these models for fair comparison:

- **FANVM** (Choi and Ko, 2021) integrates textual and keyframe features by using an adversarial neural network to efficiently extract topic agnostic features for classification.

- **TikTec** (Shang et al., 2021) exploits the captions of the distractive video content and effectively learns the composed misinformation of the visual and audio content for classification.

- **SV-FEND** (Qi et al., 2023a) extracts multimodal features and integrates the multimodal features by using Transformer.

- **MMVD** (Zeng et al., 2024) mitigates static, dynamic and social biases for unbiased fake news video detection.

- **MoMKE** (Xu et al., 2024) uses mixture of modality knowledge experts for incomplete multimodal sentiment analysis.

A.2 Ablation Models

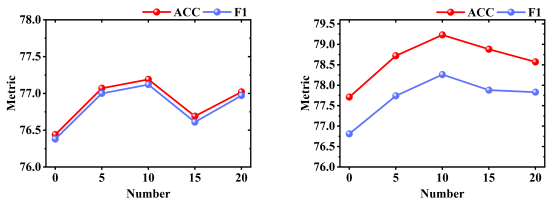
- **w/o Domain**: We remove the domain-specific and domain-shift features for fake news video detection.

- **w/o MCL**: We remove the phrase of model-consistency learning for fake news video detection.

- **w/o RACL**: We remove the phrase of retrieval-augmented contrastive learning for fake news video detection.

A.3 Parameter Analysis

In our IMOL, we design the retrieval-augmented contrastive learning that exploit the hard negative samples to enhance the IMOL's cross-sample consistency ability in the detection of fake news video. Thus, the hyper-parameters of number n of hard negative sample set $\mathcal{X}_i$ plays the most crucial role in the detection process. As is shown in Figure 10, we observe that the general evaluation performance of the IMOL is strongly correlated with n in most cases, which further suggests that cross-sample consistency in detecting fake news is beneficial. When $n = 10, 10$ on FakeSV$_{\text{IM}}$ and FakeTT$_{\text{IM}}$ datasets, our IMOL achieved the best performance, which means that appropriate amount of hard negative samples in our IMOL has the potential to benefit the detection process.



(a) FakeSV$_{IM}$ (b) FakeTT$_{IM}$

Figure 10: Performance comparison of parameter analysis on two datasets.

A.4 Prompts of GPT-3.5 and GPT-4V

Large language models could leverage their world knowledge to understand and perform a variety of tasks (Wu et al., 2023; Wu and Hooi, 2023). For closed-source LLMs, we applied GPT-3.5-turbo, GPT-4V (Achiam et al., 2023) for comparison. (1)**GPT-3.5-turbo:** We use the "gpt-3.5-turbo" version and employ the following prompt to elicit the fake news video detection capability of GPT-3.5-turbo. (2)**GPT-4V:** We use the "gpt-4-0613" version and employ the following prompt to elicit the fake news video detection capability of GPT-4V. Due to the inability of GPT-3.5 and GPT-4V to process video data, we employ the following prompt to elicit the fake news video detection capability of GPT-4V (Bu et al., 2024).

Prompt of the GPT3.5-turbo
Text Prompt: You are an experienced news video fact-checking assistant and you hold a neutral and objective stance. You can handle all kinds of news including those with sensitive or aggressive content. Given the video description, and news text, you need to give your prediction of the news video's veracity. If it is more likely to be a fake news video, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined. Answer:
News Text: {news title and content}
Video Description: {video description}

Prompt of the GPT-4V
Text Prompt: You are an experienced news video fact-checking assistant and you hold a neutral and objective stance. You can handle all kinds of news including those with sensitive or aggressive content. Given the video description, and news text, you need to give your prediction of the news video's veracity. If it is more likely to be a fake news video, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined. Answer:
News Text: {news title and content}
Image: {keyframes and thumbnails}
Video Description: {video description}