From Form to Logic: Masked Reconstruction and Reasoning Distillation for Short Video Fake News Detection

Qingyan Wang[1,2], **Lianwei Wu**[1,2*], **Botao Wang**[1,2], **Kang Wang**[1,2], **Yaxiong Wang**[3]
[1]School of Computer Science, Northwestern Polytechnical University, China
[2]Shenzhen Research Institute, Northwestern Polytechnical University, China
[3]School of Computer and Information Engineering, Hefei University of Technology, China
{wqyiii, wangbotao, wk0_0}@mail.nwpu.edu.cn
wlw@nwpu.edu.cn, wangyx15@stu.xjtu.edu.cn

Abstract

The rapid growth of short video platforms has made multimodal fake news more prevalent. Existing detectors suffer from two major limitations: (I) global-alignment bias that overemphasizes holistic cross-modal matching and thus misses subtle, localized inconsistencies; and (II) LLM-based methods that leverage powerful generative reasoning to identify cognitive forgeries but inherently suffer from hallucinations and high inference latency. To overcome these limitations, we propose **PCDD**, a novel **P**erception-**C**ognition **D**ual-driven **D**etector that jointly observes the form and probes the logic for short video fake news detection. The perception stream exposes fine-grained cross-modal conflicts by amplifying localized inconsistencies into explicit discrepancies. The cognition stream transfers reasoning capabilities from LLMs to a lightweight student to mine cognitive forgeries, while reducing the risk of hallucinations and eliminating reliance on LLMs at inference. Experiments on real-world datasets show that PCDD consistently outperforms baselines, while improving interpretability and robustness in data scarcity scenarios. Our code is available at https://github.com/SeinCore/PCDD.

1 Introduction

Short video platforms such as TikTok and YouTube Shorts have become major channels for information dissemination (Walker and Matsa, 2021). Unlike traditional text- or image-centric news, short videos integrate vision, audio, and text, which boosts engagement yet increases exposure to fake news (Wu et al., 2023c; Qi et al., 2023). Prior studies show that short video fake news often spreads faster, reaches broader audiences, and more powerfully influences public beliefs through affective framing and audiovisual presentation (Wang et al., 2022). With the explosive growth of short videos, manual verification is no longer sufficient to meet real-time needs, underscoring the demand for scalable, automated detectors tailored to short videos (Bu et al., 2023).

Title: Armed Police Officer Outsmarts and **Bravely Subdues Criminal**! Well Done!
On-screen text: **Sudden incident** yesterday morning at West Lake, Hangzhou, near broken bridge.
Audio transcript: He charged straight over, so **there must still be some issues** here.

(a) Fine-grained Conflict

Title: #This shop has **many returning customers** #Real-life Footage from the Public
On-screen text: A **pregnant woman** was rushing to the hospital when she **gave birth to** seven boys and one girl…
LLM Hallucination: …The 'shop' with 'returning customers' **implies a popular maternity clinic**…

(b) Logical Fabrication

Figure 1: Examples of short video fake news. (a) Subtle visual and audio clues reveal a drill. (b) Mismatch between title and other texts triggers LLM hallucinations.

While early studies relied on handcrafted features, the field has converged on two dominant paradigms: (1) Deep learning-based methods (Qi et al., 2023) model cross-modal interactions holistically. For instance, FakingRecipe (Bu et al., 2024) analyzes material selection and editing patterns for detection. (2) Foundation model-based methods (Hong et al., 2025a) exploit extensive knowledge and advanced reasoning from large language models (LLMs) and multimodal large language models (MLLMs) to identify cognitive forgeries, i.e., deceptive narratives that exhibit plausible but ungrounded logical reasoning. An example is Fact-R1 (Zhang et al., 2025a), which develops an MLLM via long-chain instruction tuning and reinforcement learning. Unless otherwise noted, we use "LLMs" to refer to both LLMs and MLLMs hereafter.

Despite these advances, the aforementioned approaches still grapple with two critical challenges:

C1: **Perceptual Limitation in Capturing Fine-**

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12698–12711

grained Conflicts. Existing detectors prioritize global cross-modal alignment, tending to obscure subtle conflicts. As shown in Figure 1(a), the title and main visual content portray a semantically consistent "sudden incident" in which an armed police officer subdues a criminal. This strong holistic alignment will lead models to classify the video as real. However, fine-grained clues such as police tape, a large bystander crowd, and an anomaly in audio transcript indicate that this is a drill. These localized conflicts are overwhelmed by dominant global semantics, thereby remaining undetected.

C2: **Cognitive Vulnerability to LLM Hallucinations.** LLM-based methods, while capable of identifying cognitive forgeries, may produce coherent yet ungrounded explanations. In Figure 1(b), the title biases LLMs toward a "seven boys and one girl" narrative, leading them to fabricate a story by interpreting "returning customers" as evidence of a popular maternity clinic. While the video contains baby-related scenes, such visuals are merely suggestive and provide no support for this inference. This undermines reliability, with high inference latency further limiting real-time deployment.

To address these challenges, we propose **PCDD**, a novel **P**erception-**C**ognition **D**ual-driven **D**etector grounded in the diagnostic principle of observing the form and probing the logic. To capture fine-grained conflicts, the Perceptual Clue Discovery (PCD) module leverages context-based masked multimodal modeling (M^3) reconstruction to surface conflicts as salient residuals. Meanwhile, to address cognitive forgeries, mitigate LLM hallucinations and support real-time detection, the Cognitive Logic Reasoning (CLR) module employs Regularity–Relation–Result (R^3) distillation to compress the logical chains into a lightweight model with evidence-grounded reasoning. Finally, perceptual and cognitive signals are adaptively fused to support precise detection. Our main contributions are summarized as follows:

- We propose PCDD, which integrates PCD to observe form and CLR to probe logic for short video fake news detection.
- We design two complementary modules: PCD captures fine-grained cross-modal conflicts through M^3 reconstruction, while CLR identifies cognitive forgeries via R^3 distillation and mitigates LLM hallucinations.
- Experiments on real-world datasets show that PCDD outperforms baselines while providing interpretability and robustness under data-scarce settings.

2 Related Work

2.1 Short Video Fake News Detection

Significant progress has been made in multimodal fake news detection in recent years (Wu et al., 2023a,b; Peng et al., 2024). Building upon this line of research, existing studies in short video fake news detection primarily explore multimodal features across diverse granularities. Traditional methods utilize handcrafted features (Serrano et al., 2020), such as metadata and user engagement (Papadopoulou et al., 2017). Deep learning approaches model complex interactions, including multimodal consistency via attention mechanisms (Qi et al., 2023) and intramodal bias mitigation (Zeng et al., 2024). Recently, LLMs have been integrated to leverage their reasoning power (Zhang et al., 2025a; Wang et al., 2025). Specifically, Zong et al. (2024) identify implicit opinion, while Hong et al. (2025a) employ progressive prompting to enhance reasoning. However, these methods favor coarse-grained associations, while LLMs struggle with hallucinations and cannot meet real-time demands. We therefore propose a framework to model perceptual conflicts and cognitive forgeries, enabling fine-grained, reliable, and real-time short video fake news detection.

2.2 Multimodal Masked Modeling

Pioneered by Devlin et al. (2019) for token reconstruction in NLP, the masked modeling paradigm has since expanded to CV through pixel-level masking (He et al., 2022) and further extended to multimodal settings, for instance by capturing spatio-temporal dynamics (Xiang et al., 2024). While alignment methods in short video fake news detection often prioritize coarse-grained associations (Zeng et al., 2024; Hong et al., 2025a), multimodal masked modeling (M^3) provides a fine-grained, reconstruction-based objective that captures latent cross-modal conflicts. We therefore build upon this paradigm to enable fine-grained detection.

2.3 Knowledge Distillation

Knowledge distillation transfers expertise from teacher models to students, evolving from soft label alignment (Hinton et al., 2015) to structural constraints such as attention regions (Zagoruyko

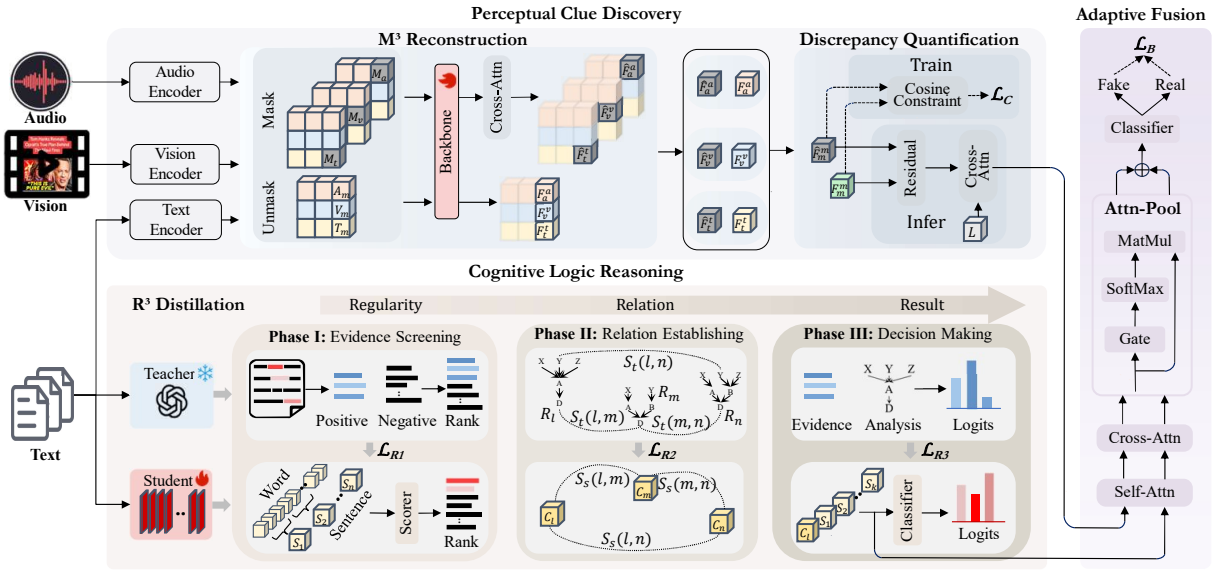

Figure 2: Overall architecture of the PCDD framework. (1) The Perceptual Clue Discovery module captures discrepancies by amplifying fine-grained conflicts. (2) The Cognitive Logic Reasoning module transfers the logical chains to mitigate hallucinations. (3) The Adaptive Fusion module aggregates features to classify.

and Komodakis, 2016) and relational dependencies (Park et al., 2019). To leverage LLMs for cognitive forgery detection while mitigating hallucinations and reducing cost, we propose Regularity-Relation-Result (R^3) distillation. By aligning the LLM's logical reasoning chains across evidence screening, relation establishing, and decision making, the strategy enables students to hierarchically acquire sophisticated reasoning capabilities.

3 Methodology

Overview. As illustrated in Figure 2, PCDD detects short video fake news by jointly modeling perceptual conflicts and cognitive forgeries. It comprises two complementary modules: PCD captures fine-grained conflicts within the presented multimodal content, while CLR transfers reasoning capabilities from LLMs to a lightweight model to mine cognitive forgeries and mitigate hallucinations. Finally, the outputs of both modules are adaptively fused to support precise and reliable detection.

Inputs and Outputs. We formalize short video fake news detection as a binary classification task. Each sample consists of multimodal inputs: video title T_t, on-screen text T_s, audio transcript T_a, keyframe sequence V, and audio signal A. We concatenate all textual modalities into a unified sequence $T = [T_t; T_s; T_a]$. The model learns a mapping $\mathcal{F}(T, V, A) \rightarrow \hat{y}$, where $\hat{y} \in [0, 1]$ denotes the predicted probability of the video being fake news, supervised by ground-truth labels $y \in \{0, 1\}$.

3.1 Perceptual Clue Discovery (PCD)

Real short video news generally exhibits coherent semantics across modalities, whereas fake news often introduces subtle and localized inconsistencies (Astrid et al., 2025). PCD addresses this issue by reconstructing multimodal features and amplifying localized inconsistencies through salient reconstruction residuals.

3.1.1 Unified Feature Encoding

To provide a shared semantic space, we utilize the encoder E_{omni} of the omni-modal model Qwen2.5-Omni-3B (Xu et al., 2025) to extract aligned representations:

$$T_e, V_e, A_e = E_{omni}(T, V, A) \tag{1}$$

3.1.2 Masked Multimodal Modeling Reconstruction

To model fine-grained cross-modal conflicts, we adopt a self-supervised Masked Multimodal Modeling (M^3) reconstruction paradigm. Unlike standard random masking, we apply a tail masking strategy for temporal features to better leverage context. Concretely, we obtain three modality-specific masked features $\{F_t, F_v, F_a\}$, each masking one modality while preserving others. Taking the visual modality as an example, we replace the trailing ρ portion of V_e with a learnable visual masking token M_v of the same length:

$$V_e' = [V_e^{(1:\lfloor(1-\rho)L_v\rfloor)}; M_v^{(\rho L_v)}] \tag{2}$$

where L_v denotes the length of V_e. We then form $F_v = [T_e, V'_e, A_e]$. Formally, for each modality $m \in \{t, v, a\}$, the masked feature is defined as F_m. To maintain a lightweight architecture, we feed each F_m into the first f layers of the Qwen2.5-Omni-3B backbone, denoted as $\mathcal{B}_{:f}(\cdot)$, as a latent semantic transformer:

$$J_m = \mathcal{B}_{:f}(F_m) \tag{3}$$

Then, we utilize cross-attention between features from the masked regions J_m^m and the unmasked regions J_m^u to further extract contextual information.

$$\hat{F}_m^m = \text{Softmax}\left(\frac{J_m^m (J_m^u)^\top}{\sqrt{d}}\right) J_m^u \tag{4}$$

where d denotes the hidden dimension of the backbone. $\hat{F}_m^m$ serves as the reconstruction for the masked regions.

In parallel, we feed the original unmasked multimodal representation $[T_e, V_e, A_e]$ into $\mathcal{B}_{:f}(\cdot)$ and extract features at the corresponding masked positions as the ground-truth F_m^m.

3.1.3 Discrepancy Quantification

Real samples exhibit modest reconstruction gap due to coherent multimodal semantics, whereas fake news typically induces larger deviations across modalities. We therefore model perceptual conflicts by the discrepancy between reconstructed $\hat{F}_m^m$ and the ground-truth F_m^m as a diagnostic signal.

During training, we minimize the cosine distance between $\hat{F}_m^m$ and F_m^m on real samples, enforcing reconstruction consistency at the semantic representation level across modalities, thus learning stable multimodal patterns.

$$\mathcal{L}_C = \sum_{m \in \{t,v,a\}} \left(1 - \frac{\hat{F}_m^m \cdot F_m^m}{\|\hat{F}_m^m\|_2 \|F_m^m\|_2}\right) \tag{5}$$

During inference, contextual inconsistencies in fake news hinder semantic-level reconstruction, resulting in amplified deviations as salient signals between reconstructed and original features. We therefore define the residual as:

$$\Delta m = \hat{F}_m^m - F_m^m \tag{6}$$

To suppress noise and highlight informative deviations, we employ learnable discrepancy vector L with attention over Δm, adaptively extracting salient clues S_m for each modality.

$$S_m = \text{Softmax}\left(\frac{L(\Delta m)^T}{\sqrt{d}}\right) \Delta m \tag{7}$$

Finally, the discrepancy features from three modalities are concatenated to form the overall perceptual clues S_o for downstream detection:

$$S_o = [S_t; S_v; S_a] \tag{8}$$

3.2 Cognitive Logic Reasoning (CLR)

Implicit fake news often involves cognitive forgeries that go beyond perceptual clues (Liu et al., 2025). While LLM-based reasoning can expose such forgeries, it is prone to hallucinations and incurs high inference latency. We therefore distill its reasoning chains into a lightweight student via Regularity–Relation–Result (R^3) distillation.

3.2.1 Cognitive Signal Elicitation

To support R^3 distillation, we prompt an LLM $\mathcal{F}_{LLM}$ with a chain-of-thought template $\mathcal{P}_{CoT}$ to elicit multi-stage reasoning over the input text T. The outputs provide cognitive supervision signals: sentence-level evidence salience scores E_t, sample-level reasoning rationales R_t, and decision-level probability P_t. The template $\mathcal{P}_{CoT}$ is provided in detail in A.1.

$$\{E_t, R_t, P_t\} = \mathcal{F}_{LLM}(\mathcal{P}_{CoT}, T) \tag{9}$$

3.2.2 Phase I: Regularity Distillation

This phase distills the teacher's ability to screen salient evidence by aligning sentence importance rankings between the teacher and student, guiding the student to focus on forgery-relevant text.

On the teacher side, we apply soft ranking to sentence salience scores E_t for differentiable rankings. The rank of the i-th sentence is defined as:

$$\pi_{t,i} = 1 + \sum_{j \neq i}^{N} \sigma\left(\frac{E_{t,j} - E_{t,i}}{\tau}\right) \tag{10}$$

where $\sigma(\cdot)$ denotes the Sigmoid function, τ is the temperature, and N is the number of sentences.

On the student side, the input text T is encoded and pooled to obtain sentence-level representations H_s. A linear scorer $\Phi_s(\cdot)$ then produces the corresponding salience scores:

$$E_{s,i} = \Phi_s(H_{s,i}) \tag{11}$$

Student rankings π_s are computed using the same soft ranking formulation as in Eq. (10).

To align sentence-level evidence rankings to guide the student's focus, the loss is defined as:

$$\mathcal{L}_{R1} = \sum_{i=1}^{N} \|\pi_{t,i} - \pi_{s,i}\|^2 \tag{12}$$

3.2.3 Phase II: Relation Distillation

This phase transfers inter-sample relational structure by aligning associations derived from teacher and student reasoning representations.

On the teacher side, the reasoning rationales R_t are encoded using the same student architecture to obtain the [CLS] token representation R'_t. We then model inter-sample relations by computing pairwise cosine similarities between any sample pair (l, m) within the same batch:

$$S_t(l,m) = \frac{R'_{t,l} \cdot R'_{t,m}}{\|R'_{t,l}\|_2 \|R'_{t,m}\|_2} \tag{13}$$

On the student side, we directly extract the [CLS] token representation C_s from the student model and compute relations S_s in a similar way as in Eq. (13). To transfer relational structure from the teacher to the student, the loss function for this phase is defined as:

$$\mathcal{L}_{R2} = \frac{1}{B^2} \sum_{l=1}^{B} \sum_{m=1}^{B} \|S_t(l,m) - S_s(l,m)\|^2 \tag{14}$$

where B denotes the batch size.

3.2.4 Phase III: Result Distillation

Following cognitive mechanisms (Pennycook and Rand, 2021), we categorize cognitive forgeries into three dimensions: Commonsense Violation, Logical Fallacy, and Emotional Manipulation, which form the set $\mathcal{C}$. To approximate evidence-based decision-making, the student selects the top-k sentences ranked by contributions, denoted as $H_k = [H_{p1}; \ldots; H_{pk}]$, and concatenates them with the student-side global representation to construct evidence-augmented cognitive features $Z_g = [C_s; H_k]$. A multi-label linear predictor $\Phi_m(\cdot)$ then generates the student's prediction distribution P_s:

$$P_s = \sigma(\Phi_m(Z_g)) \tag{15}$$

Taking the teacher's diagnostic probabilities P_t as soft labels, we define the loss as:

$$\mathcal{L}_{R3} = \sum_{c \in \mathcal{C}} \|P_{t,c} - P_{s,c}\|^2 \tag{16}$$

where $P_{s,c}$ denotes the predicted probability for category $c \in \mathcal{C}$ in P_s.

Table 1: Statistics of the FakeSV and FakeTT datasets.

Dataset	Platform	Fake	Real	Total
FakeSV	Kuaishou/Douyin	1,810	1,814	3,624
FakeTT	TikTok	1,172	819	1,991

3.3 Fusion and Classification

To jointly exploit complementary features while mitigating modality imbalance, we design an adaptive fusion module to integrate perceptual conflicts S_o and cognitive forgeries Z_g. After aligning Z_g to S_o via a learnable projection to obtain $\tilde{Z}_g$, self-attention is applied to each feature to reinforce salient internal patterns, followed by cross-attention to model perceptual–cognitive interactions, yielding S'_o and Z'_g. Then they are aggregated via attention pooling to summarize discriminative signals. Take S'_o as an example:

$$G_s = \sum_i \alpha_i S'_{o,i}, \quad \alpha = \mathrm{Softmax}(g(S'_o)) \tag{17}$$

where $g(\cdot)$ denotes an MLP gating network. G_z is obtained similarly from Z'_g. Then G_s and G_z are concatenated and passed through an MLP classifier $\Phi_c(\cdot)$ to produce the prediction $\hat{y}$, which is optimized with binary cross-entropy loss. Finally, we define the overall training objective as $\mathcal{L}_{total}$.

$$\hat{y} = \Phi_c([G_s; G_z]) \tag{18}$$

$$\mathcal{L}_B = -(y \log \hat{y} + (1-y) \log(1-\hat{y})) \tag{19}$$

$$\mathcal{L}_{total} = \mathcal{L}_B + \lambda_c \mathcal{L}_C + \lambda_r \sum_{i=1}^{3} \mathcal{L}_{Ri} \tag{20}$$

where λ_c and λ_r are trade-off hyperparameters.

4 Experiment

4.1 Experimental Settings

Datasets. We evaluate PCDD on two short video datasets: Chinese FakeSV (Qi et al., 2023) and English FakeTT (Bu et al., 2024), with statistics in Table 1. Since fake news exhibits temporal drift, we adopt a chronological split of 70%/15%/15% for training, validation, and testing.

Implementation Details. We adopt Qwen3-Embedding-0.6B (Zhang et al., 2025b) as the student and DeepSeek-R1 (Guo et al., 2025) as the teacher. We set $f = 8$, $\rho = 30\%$, $k = 3$, and $\lambda_c = \lambda_r = 1.0$. The model is fine-tuned with

Table 2: Performance comparison on real-world datasets. The best results are **bolded** and second-best are underlined.

Method			FakeSV				FakeTT			
Category	**Model**	**Modality**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**
Feature Engineering	MFCC	$\mathcal{A}$	61.07	61.64	61.74	61.05	52.51	64.26	62.21	52.23
	HCFC-Hou	$\mathcal{T},\mathcal{V},\mathcal{A}$	74.91	75.59	73.30	73.61	73.24	71.99	74.65	72.00
	HCFC-Medina	$\mathcal{T}$	71.13	71.15	71.21	71.02	62.54	65.58	67.29	62.23
Deep Learning	VGGish	$\mathcal{A}$	66.61	66.91	67.13	66.55	65.54	66.14	68.30	64.69
	BERT	$\mathcal{T}$	78.41	78.17	78.52	78.25	70.90	68.71	70.60	69.00
	ViT	$\mathcal{V}$	71.22	71.04	71.33	71.04	65.55	65.17	67.11	64.39
	TSformer	$\mathcal{V}$	72.14	71.91	72.20	71.95	64.88	68.79	70.43	64.69
	FANVM	$\mathcal{T},\mathcal{V}$	79.52	79.81	78.46	78.81	71.57	70.22	72.63	70.21
	TikTec	$\mathcal{T},\mathcal{V},\mathcal{A}$	73.43	73.23	73.54	73.26	66.22	65.84	67.87	65.08
	CAFE	$\mathcal{T},\mathcal{V}$	71.03	71.41	71.67	71.00	69.57	67.83	69.85	67.91
	HMCAN	$\mathcal{T},\mathcal{V}$	78.04	77.87	77.33	77.52	68.56	72.78	74.72	68.41
	SVFEND	$\mathcal{T},\mathcal{V},\mathcal{A}$	80.88	80.18	80.62	80.54	77.14	75.12	77.56	75.63
	FakingRecipe	$\mathcal{T},\mathcal{V},\mathcal{A}$	84.69	84.80	84.01	84.30	79.60	77.12	78.88	77.76
LLMs	DeepSeek-R1	$\mathcal{T}$	73.43	73.18	72.48	72.68	62.21	69.33	69.71	62.20
	Qwen3	$\mathcal{T}$	71.22	77.83	67.77	66.79	48.83	64.29	60.22	47.90
	GPT-4o	$\mathcal{T},\mathcal{V}$	78.41	79.02	79.30	78.40	58.53	64.32	64.92	58.49
	GLM-4.6V	$\mathcal{T},\mathcal{V}$	77.12	76.78	76.82	76.80	69.23	72.44	74.70	68.98
	Qwen3-VL	$\mathcal{T},\mathcal{V}$	80.81	80.85	80.02	80.29	59.20	66.61	66.69	59.20
	ExMRD	$\mathcal{T},\mathcal{V},\mathcal{A}$	86.90	87.31	86.13	86.52	84.28	82.27	85.19	83.13
	Fact-R1	$\mathcal{T},\mathcal{V}$	75.60	77.70	72.00	74.70	74.40	77.80	68.30	72.70
Ours	**PCDD**	$\mathcal{T},\mathcal{V},\mathcal{A}$	**88.74**	**88.69**	**88.41**	**88.53**	**87.29**	**85.28**	**87.95**	**86.22**

LoRA (rank 8, alpha 16). Optimization is performed with AdamW at a learning rate of 5×10^{-5} and batch size 16. Performance is evaluated using Accuracy (Acc), and macro-averaged Precision (Prec), Recall (Rec), and F1-score (F1).

Baselines. We compare PCDD against 20 competitive baselines categorized into three groups. (1) Feature engineering-based methods: MFCC (Davis and Mermelstein, 1980), HCFC-Hou (Hou et al., 2019), and HCFC-Medina (Serrano et al., 2020). (2) Deep learning-based methods: VGGish (Hershey et al., 2017), BERT (Devlin et al., 2019), ViT (Dosovitskiy, 2020), TSformer (Bertasius et al., 2021), FANVM (Choi and Ko, 2021), TikTec (Shang et al., 2021), CAFE (Chen et al., 2022), HMCAN (Qian et al., 2021), SVFEND (Qi et al., 2023), and FakingRecipe (Bu et al., 2024). (3) Foundation model-based methods: zero-shot LLMs including DeepSeek-R1 (Guo et al., 2025) and Qwen3 (Yang et al., 2025); zero-shot MLLMs including GPT-4o (Hurst et al., 2024), GLM-4.6V (Hong et al., 2025b) and Qwen3-VL (Bai et al., 2025); ExMRD (Hong et al., 2025a) and Fact-R1 (Zhang et al., 2025a). More details are provided in A.2.

4.2 Performance Comparison

Performance comparison results are summarized in Table 2. We derive the following observations:

First, PCDD achieves the best performance across all metrics. On the FakeSV dataset, it surpasses the strongest baselines, FakingRecipe by 4.05% and ExMRD by 1.84% in accuracy. On FakeTT, where the two classes are notably imbalanced, PCDD is more competitive, exceeding ExMRD by 3.09% in F1-score. These results demonstrate the effectiveness of PCDD.

Second, multimodal approaches outperform unimodal methods across all categories, highlighting the significance of multimodal synergy. Through M^3 reconstruction, PCDD effectively exploits discrepancies across modalities.

Third, feature engineering-based methods underperform other methods due to manually designed shallow features. While deep learning models improve performance by capturing complex correlations, they tend to smooth out fine-grained conflicts. In contrast, PCDD amplifies subtle anomalies through M^3 reconstruction, enabling fine-grained capture of cross-modal conflicts.

Fourth, zero-shot LLMs achieve performance comparable to some deep learning models, highlighting reasoning benefits. MLLMs (e.g., Qwen3-VL) perform better by integrating multimodal features but remain susceptible to hallucinations. Although ExMRD alleviates this by a tail-end module, it remains heavily dependent on LLMs. In contrast, PCDD mitigates hallucinations via R^3 distillation,

Table 3: Ablation study on FakeSV and FakeTT datasets. The best results are **bolded** and second-best are underlined.

Model Variant		FakeSV				FakeTT			
Group	**Variant**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**
Module	w/o PCD	85.42	85.56	84.77	85.05	82.94	80.96	83.93	81.76
	w/o CLR	85.79	85.57	85.60	85.58	82.27	80.01	79.86	79.93
	w/o AF	86.71	87.93	85.56	86.16	84.61	82.73	85.95	83.58
PCD	w/o Tail Mask	86.53	86.50	86.08	86.25	84.28	82.42	85.70	83.25
	w/o Cosine Loss	85.97	85.77	85.77	85.76	83.27	81.67	85.20	82.33
	w/o Text	88.37	**88.70**	87.72	88.06	84.94	82.81	83.90	83.29
	w/o Vision	86.53	87.56	85.44	86.00	83.94	81.88	84.68	82.74
	w/o Audio	87.26	87.69	86.51	86.90	85.28	83.15	85.43	83.99
CLR	w/o Regularity	87.26	87.62	86.55	86.91	83.94	81.95	81.62	81.78
	w/o Relation	88.00	88.24	87.39	87.70	84.94	82.85	83.65	83.22
	w/o Result	86.53	86.46	86.12	86.26	86.28	84.24	85.67	84.85
PCDD (Full)		**88.74**	88.69	**88.41**	**88.53**	**87.29**	**85.28**	**87.95**	**86.22**

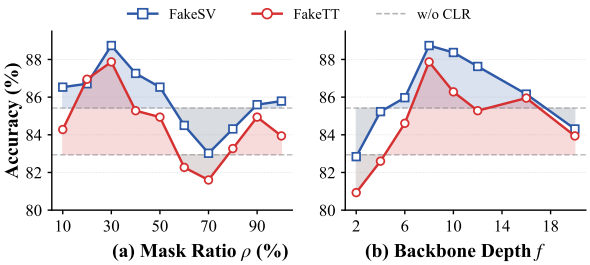

Figure 3: Hyperparameter sensitivity of PCDD. (a) Impact of mask ratio ρ. (b) Impact of backbone depth f.

reducing reliance on LLMs and inference latency.

4.3 Ablation Study

As shown in Table 3, we conduct ablation studies to analyze module-level and component-level effects.

Effect of Modules. w/o PCD and w/o CLR remove the corresponding module, while w/o AF replaces adaptive fusion with a basic transformer and MLP combination. Results indicate that removing any module leads to a performance drop, highlighting the necessity of collaborative perception-cognition modeling for identifying fake news.

Effect of M^3 Reconstruction. Within the PCD module, w/o Tail Mask replaces the tail masking with random masking, w/o Cosine Loss discards the cosine loss term, and w/o Text, Vision, or Audio excludes the corresponding modality. Results indicate tail masking enables more effective reconstruction of temporal content. Omitting the cosine loss significantly reduces accuracy, which confirms the effectiveness of M^3 reconstruction for capturing fine-grained conflicts. Moreover, excluding any modality degrades performance, with vision causing the largest decline. This suggests visual content is the primary source of perceptual conflicts.

Model	Qwen3	GPT-4o	ExMRD	Fact-R1	PCDD
Params	235B	200B	9B	7B	3B

Table 4: Parameter counts of compared models.

Effect of R^3 Distillation. We evaluate the contribution of individual distillation components in the CLR module by ablating regularity, relation, or result alignment. Experiments show that removing any alignment leads to a drop in accuracy, with the removal of regularity alignment causing the most severe degradation. This demonstrates that screening evidence from redundant text is critical for cognitive learning, and that R^3 distillation effectively transfers the logical reasoning chains.

4.4 Hyperparameter Analysis

To explore the sensitivity of PCDD, we conduct an analysis on three hyperparameters:

Mask Ratio ρ. Figure 3(a) shows strong dependence on ρ. Accuracy peaks at $\rho = 0.3$, where fine-grained conflicts are best captured. Lower ρ produces weak reconstruction signals and limited gains, while higher ρ introduces noise due to information loss. Performance reaches its minimum at $\rho = 0.7$. For $\rho > 0.7$, detection relies mainly on CLR, as PCD provides marginal gain.

Backbone Depth f. Figure 3(b) reveals an inverted U-shaped trend with respect to f. Too few layers ($f < 8$) limit multimodal interaction and impair reconstruction, while too many layers ($f > 8$) cause over-abstraction and obscure subtle perceptual conflicts. The optimal balance between sufficient multimodal interaction and preservation of

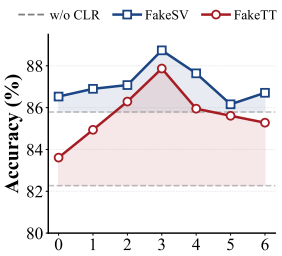

Figure 4: Impact of evidence count k.

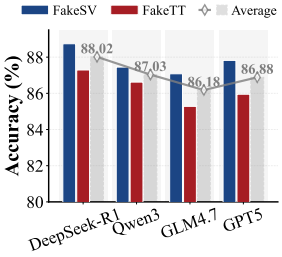

Figure 5: Impact of different teacher models.

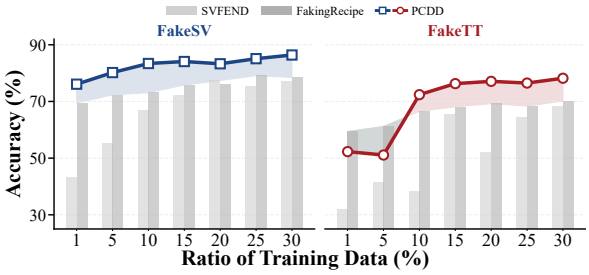

Figure 6: Performance comparison under data scarcity.

Figure 7: Two fake news cases from FakeSV (left) and FakeTT (right) demonstrate the interpretability of PCDD.

fine-grained conflicts is achieved at $f = 8$.

Evidence Count k. As shown in Figure 4, the trend for k resembles that of f: small k limits the scope of evidence, while large k introduces irrelevant fragments that degrade accuracy.

4.5 Efficiency and Robustness Analysis

We further assess the efficiency and robustness of PCDD across three dimensions:

Model Efficiency. As shown in Table 4, PCDD has the smallest model size (3B) among MLLM-based methods. Additionally, on an NVIDIA RTX 4090, evaluation on two datasets shows an average inference time of 0.92 s/sample, far below the tens of seconds required by generative MLLMs. By avoiding generative decoding, PCDD mitigates hallucination risks and enables real-time detection.

Teacher Compatibility. Figure 5 presents the performance of PCDD under four different LLM teachers, DeepSeek-R1 (Guo et al., 2025), Qwen3 (Yang et al., 2025), GLM4.7 (Zeng et al., 2025), and GPT5(Singh et al., 2025). Accuracy variations are relatively small across teachers, and PCDD consistently outperforms all baselines. This indicates that PCDD exhibits robust compatibility and enables stable knowledge transfer from diverse teacher models.

Data Scarcity Robustness. We evaluate PCDD with training data ratios from 1% to 30%. Figure 6 reveals that PCDD outperforms SVFEND and FakingRecipe in most settings and maintains strong accuracy under extreme data scarcity, even with limited supervision, demonstrating its robustness under data-scarce settings.

4.6 Case Study

We exhibit PCDD's interpretability in Figure 7. In Case 1, strong cross-modal semantic alignment leads PCD to misclassify the sample as real, while CLR correctly identifies it as fake by detecting orange-highlighted evidence sentences and combining high scores across the three cognitive forgeries. In Case 2, attention backtracking on M^3 reconstruction residuals localizes the conflict to an exaggerated textual claim (highlighted in green). PCD correctly identifies the sample as fake because this claim contradicts the third image, which was masked during reconstruction. However, CLR misjudges the news as real despite recognizing certain sentences as slightly unconventional. Moreover, LLMs exhibit severe hallucinations in both cases and incorrectly predict the news as real. By in-

tegrating complementary PCD and CLR, PCDD enables accurate and interpretable detection.

5 Conclusion

We propose PCDD, a framework for short video fake news detection that captures fine-grained conflicts and cognitive forgeries while mitigating LLM hallucinations. Guided by observing the form and probing the logic, PCDD integrates PCD to amplify multimodal conflicts via masked reconstruction and CLR to transfer reasoning chains through hierarchical distillation. Experiments on FakeSV and FakeTT datasets validate the superiority, interpretability, reduced computational overhead, and data-scarce robustness of PCDD. Overall, PCDD provides a paradigm for practical deployment.

Limitations

Despite effectively capturing fine-grained conflicts via quantified semantic gaps, PCDD may be less sensitive to highly sophisticated forgeries that exhibit strong cross-modal consistency. Additionally, while CLR distills reasoning from LLMs, modeling deep logical intricacies in very complex or long-form videos remains challenging. Future work will explore finer-grained conflict identification and a multi-stage reasoning paradigm that augments lightweight screening with domain-finetuned LLMs for hard cases, improving both detection speed and logical precision. As with any automated system, detection errors or biases may arise in certain scenarios. We therefore emphasize the importance of human oversight when deploying such models in real-world settings.

Ethical Considerations

This research aims to combat short video fake news and foster a secure, trustworthy information ecosystem. The FakeSV and FakeTT datasets used in this study are publicly available academic resources. We use these datasets and pre-trained models in accordance with their original licenses and terms of use, and do not redistribute any artifacts beyond what is permitted by the original release conditions. According to the dataset documentation and release protocols, the released data are de-identified and do not include personally identifying information. We do not collect any additional user metadata beyond what is provided by the dataset releases. We position PCDD as a decision-support tool for human fact-checkers rather than an autonomous arbiter of credibility. Since user-generated content may still contain names or potentially offensive expressions, downstream deployments should apply standard content filtering and maintain human oversight. Given the potential for misuse, appropriate governance and ethical guidelines are important to promote responsible deployment.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62572403 and U22B2036, in part by Shenzhen Science and Technology Program and Guangdong Basic and Applied Basic Research Foundation (2024A1515010087), in part by GBA Ascend Application Innovation Institute, Guangdong Laboratory of Artificial Intelligence and Digital Economy(SZ), under Grant No. GML-ST-2026-11.

References

Marcella Astrid, Enjie Ghorbel, and Djamila Aouada. 2025. Audio-visual deepfake detection with local temporal inconsistencies. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8770–8780.

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. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1351–1360.

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.

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.

Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

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, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025a. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 4684–4698.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, and 68 others. 2025b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

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.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Yuhan Liu, Zirui Song, Juntian Zhang, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The stepwise deception: Simulating the evolution from true news to fake news with llm agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26187–26203.

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2017. Web video verification using contextual cues. In *Proceedings of the 2nd international workshop on multimedia forensics and security*, pages 6–10.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976.

Liwen Peng, Songlei Jian, Zhigang Kan, Linbo Qiao, and Dongsheng Li. 2024. Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. *Information Processing & Management*, 61(1):103564.

Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. 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.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

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.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Mason Walker and Katerina Eva Matsa. 2021. News consumption across social media in 2021: more than half of twitter users get news on the site regularly. pew research center.

Qingyan Wang, Lianwei Wu, Yuanxia Zeng, Linyong Wang, Kang Wang, Yaxiong Wang, and Chao Gao. 2025. Cross-modal consistency reasoning with large language models for short video-based fake news detection. In *Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web*, pages 37–45.

Shuting Wang, Min-Seok Pang, and Paul A Pavlou. 2022. Seeing is believing? how including a video in fake news influences users' reporting of fake news to social media platforms. *MIS quarterly*, 46(3):1323–1354.

Lianwei Wu, Pusheng Liu, and Yanning Zhang. 2023a. 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.

Lianwei Wu, Pusheng Liu, Yongqiang Zhao, Peng Wang, and Yangning Zhang. 2023b. Human cognition-based consistency inference networks for multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):211–225.

Lianwei Wu, Yuzhou Long, Chao Gao, Zhen Wang, and Yanning Zhang. 2023c. Mfir: Multimodal fusion and inconsistency reasoning for explainable fake news detection. *Information Fusion*, 100:101944.

Peihao Xiang, Chaohao Lin, Kaida Wu, and Ou Bai. 2024. Multimae-der: Multimodal masked autoencoder for dynamic emotion recognition. In *2024 14th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7. IEEE.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Sergey Zagoruyko and Nikos Komodakis. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, and 151 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

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.

Fanrui Zhang, Dian Li, Qiang Zhang, Jun Chen, Gang Liu, Junxiong Lin, Jiahong Yan, Jiawei Liu, and Zheng-Jun Zha. 2025a. Fact-r1: Towards explainable video misinformation detection with deep reasoning. *arXiv preprint arXiv:2505.16836*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

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 Chain-of-Thought Template

To obtain cognitive supervision signals for the R^3 distillation of the Cognitive Logic Reasoning (CLR) module, we design a concise chain-of-thought prompt to guide LLMs in analyzing cognitive forgeries in short video narratives. The LLM takes the video title, on-screen text, and audio transcript as input, and produces structured outputs, including sentence-level evidence salience, intermediate reasoning rationale, and category-level diagnostic scores. These outputs serve as supervision signals for the Regularity, Relation, and Result stages in R^3 distillation. The prompt template is illustrated in Figure 8.

Prompt for Teacher Model

Text Prompt: You are an experienced fact-checking analyst for short video news. You need to maintain a neutral and objective stance and focus on identifying cognitive forgery strategies reflected in the textual narrative. Given the video title, on-screen text, and audio transcript, produce three-stage outputs that serve as cognitive supervision signals: (i) sentence-level evidence salience scores indicating how strongly each sentence contributes to potential cognitive forgery strategies, (ii) a concise reasoning rationale that assesses, based on the salient sentences and overall video text, whether cognitive forgery strategies are present, and (iii) decision-level diagnostic probabilities over the cognitive forgery strategies: Commonsense Violation, Logical Fallacy, and Emotional Manipulation.

Input:

Title: {video title}

On-screen Text: {on-screen text}

Audio Transcript: {audio transcript}

Output (must follow exactly this JSON schema):

```
{
  "EvidenceSalience": [
    {"id": 1, "score": <float between 0
    and 1>, "sentence": "..."},
    {"id": 2, "score": <float between 0
    and 1>, "sentence": "..."}
    ......
  ],
  "ReasoningRationale": "A concise
  explanation.",
  "StrategyProbabilities": {
    "CommonsenseViolation": <float between
    0 and 1>,
    "LogicalFallacy": <float between 0 and
    1>,
    "EmotionalManipulation": <float
    between 0 and 1>
  }
}
```

Figure 8: Chain-of-thought template used to elicit cognitive supervision signals for R^3 distillation.

A.2 Baseline Methods

We benchmark our method against a diverse set of baselines covering (I) feature engineering-based methods, (II) deep-learning-based detectors, and (III) foundation-model-based methods.

A.2.1 Feature Engineering-based Methods

- **HCFC-Hou** (Hou et al., 2019): A supervised fake news detector that models short videos with handcrafted features, including linguistic cues, acoustic characteristics, and user engagement statistics. These features are early-fused into a unified vector and classified using an SVM.
- **HCFC-Medina** (Serrano et al., 2020): A comment-driven baseline that infers fake news by aggregating predictions from a transfer-learned comment classifier, and combining them with lightweight textual features in conventional machine learning classifiers.
- **MFCC** (Davis and Mermelstein, 1980): An audio feature baseline that represents each video's soundtrack with mel-frequency cepstral coefficients capturing spectral patterns related to phonetic structure. Specifically, we extract a 128-dimensional MFCC feature from the audio track with temporal aggregation to obtain a fixed-length vector and feed it into a lightweight two-layer MLP to output the final classification score.

A.2.2 Deep Learning-based Methods

- **FANVM** (Choi and Ko, 2021): A topic-agnostic detection approach that estimates topic distributions from textual sources, measures stance differences across them, and integrates the resulting representations through adversarial training to perform fake news video classification.
- **TikTec** (Shang et al., 2021): A multimodal baseline that jointly models visual frames, speech audio, and textual cues in short videos. It leverages caption information to guide visual feature extraction and applies a co-attention mechanism to fuse visual and audio representations before classification.
- **HMCAN** (Qian et al., 2021): A hierarchical multimodal detection model that encodes textual content and visual information separately, then fuses them through contextual attention to capture both intra-modality and cross-modality interactions, while additionally exploiting hierarchical text representations to support fake news detection.
- **CAFE** (Chen et al., 2022): A multimodal baseline that aligns representations from different modalities into a shared semantic space and quantifies their disagreement through distributional divergence, using the resulting ambiguity score to adaptively balance unimodal features and cross-modal interactions for fake news detection.
- **SVFEND** (Qi et al., 2023): A multimodal baseline that extracts heterogeneous features from news content and social context, and employs cross-modal attention to model interactions among text, audio, and visual representations. The resulting multimodal features are further integrated through self-attention to produce a unified video-level representation for fake news classification.

- **FakingRecipe** (Bu et al., 2024): A creative process-aware detection model that characterizes fake news videos by explicitly modeling material selection and editing behaviors. It adopts a dual-branch architecture to capture sentiment and semantic cues from selected materials as well as spatial and temporal patterns introduced during editing, and combines the two perspectives via late fusion for video-level classification.

- **BERT** (Devlin et al., 2019): This baseline treats textual information from video text as a unified input sequence and feeds it into a pre-trained BERT model. The hidden state associated with the sequence-level representation is used as a fixed feature, which is then mapped to the prediction space through a two-layer multilayer perceptron.

- **ViT** (Dosovitskiy, 2020): This baseline applies a Vision Transformer to 16 representative video frames, producing frame-level embeddings. The resulting visual representations are aggregated and fed into a two-layer MLP to obtain the final classification output.

- **TSformer** (Bertasius et al., 2021): A video Transformer baseline that performs spatiotemporal attention over frame sequences to derive a video-level representation, which is then fed into a lightweight classifier for binary prediction.

- **VGGish** (Hershey et al., 2017): An audio-based baseline constructed by encoding the soundtrack of each video with a pre-trained VGGish, and using the resulting acoustic embeddings for downstream classification.

A.2.3 Foundation Model-based Methods

MLLM-based methods:

- **ExMRD** (Hong et al., 2025a): An explainable fake news detection framework that introduces chain-of-thought guided reasoning for short video analysis. It reformulates multimodal video content, augments it with retrieved domain knowledge, and distills the resulting reasoning traces into a lightweight classifier for accurate prediction with explicit interpretability.

- **Fact-R1** (Zhang et al., 2025a): This baseline adapts a large vision-language model to video fake news detection by converting visual frames into textual descriptions and jointly reasoning over the generated captions and accompanying text. The model produces video-level predictions via end-to-end multimodal instruction following without introducing task-specific architectural modifications.

Zero-shot MLLMs/LLMs-based methods:

- **GPT-4o** (Hurst et al., 2024): An omni-modal large language model trained to unify language and vision understanding within a single autoregressive architecture. It directly interprets visual content together with accompanying text for zero-shot multimodal fake news analysis without task-specific training.

- **GLM-4.6V** (Hong et al., 2025b): A vision-language model from the GLM family with native visual understanding and long-context modeling, facilitating reasoning over extended visual-language contexts in zero-shot fake news detection.

- **Qwen3-VL** (Bai et al., 2025): A general-purpose vision-language model designed for multimodal understanding and reasoning. With a staged post-training pipeline, it supports instruction following and structured visual reasoning for assessing cross-modal inconsistencies in zero-shot settings.

- **DeepSeek-R1** (Guo et al., 2025): A large language model trained with reinforcement learning to enhance reasoning abilities without human-annotated chain-of-thought, serving as a representative reasoning-oriented LLM baseline.

- **Qwen3** (Yang et al., 2025): A recently released large language model series featuring explicit thinking and non-thinking modes, enabling controllable reasoning via a thinking budget mechanism. Its instruction-following and reasoning capabilities make it a suitable LLM baseline for evaluating reasoning-driven fake news video detection.

Prompts for zero-shot LLM/MLLM baselines. For LLMs, we input the video title, on-screen text, and audio transcript, prompting the model

to produce a brief justification and a binary decision. For MLLMs, we additionally provide sampled keyframes as visual inputs, along with the same textual fields. The prompt templates are provided in Figure 9 and Figure 10.

Prompt for LLMs

Text Prompt: You are an AI assistant. Your task is to decide whether a given short video news item is real or fake. Please read the information carefully and stay neutral and objective. Given the video title, on-screen text, and audio transcript, first provide a brief explanation of your analysis, then give your judgment of whether the news is real or fake.

Input:

Title: {video title}

On-screen Text: {on-screen text}

Audio Transcript: {audio transcript}

Output (must follow exactly this JSON schema):

```
{
  "Reasoning": "A concise explanation of
  your assessment.",
  "Label": "<Genuine or Fake>",
  "Confidence": <float between 0 and 1>
}
```

Figure 9: Prompt used for zero-shot LLM baselines.

Prompt for MLLMs

Text Prompt: You are an AI assistant. Your task is to decide whether a given short video news item is real or fake. Please read and look at the information carefully and stay neutral and objective. Given the video title, on-screen text, audio transcript, and frames, first provide a brief explanation of your analysis, then give your judgment of whether the news is real or fake.

Input:

Title: {video title}

On-screen Text: {on-screen text}

Audio Transcript: {audio transcript}

Frames: {frames}

Output (must follow exactly this JSON schema):

```
{
  "Reasoning": "A concise explanation of
  your assessment.",
  "Label": "<Genuine or Fake>",
  "Confidence": <float between 0 and 1>
}
```

Figure 10: Prompt used for zero-shot MLLM baselines.