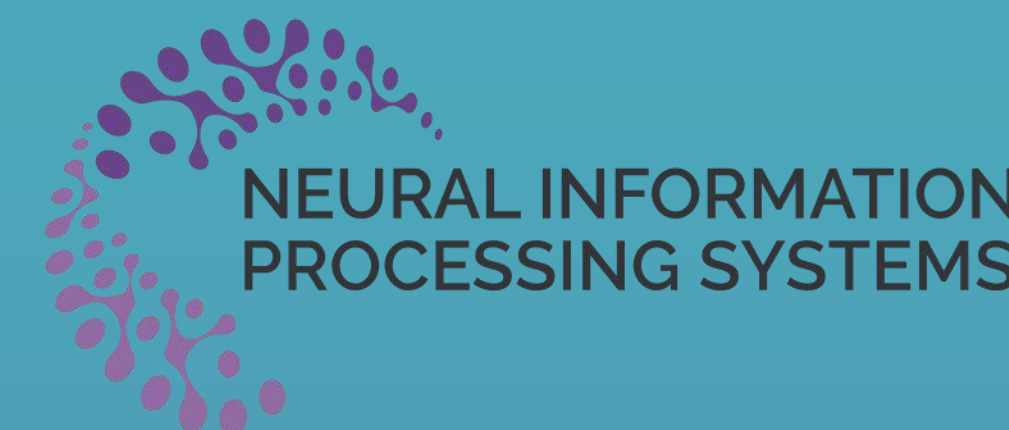




上海创智学院
Shanghai Innovation Institute

Tencent 腾讯



FACT-R1: Towards Explainable Video Misinformation Detection with Deep Reasoning

Fanrui Zhang^{1,2*}, Dian Li^{3†}, Qiang Zhang^{1*}, Jun Chen³, Gang Liu³, Junxiong Lin⁴, Jiahong Yan³, Jiawei Liu^{1*}, Zheng-Jun Zha^{1†}

¹MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, USTC

²Shanghai Innovation Institute ³Tencent QQ ⁴Fudan University

Abstract

1. FakeVV benchmark: 100k+ video–text pairs with fine-grained, interpretable labels, tackling the shortage of large, diverse multimodal misinformation datasets.
2. Fact-R1 framework: Combines deep reasoning with collaborative rule-based RL to reduce template overfitting and support interpretable verification of deceptive content.
3. Three-stage training: (1) long CoT instruction tuning, (2) preference alignment via DPO, and (3) GRPO with a verifiable reward—yielding emergent reasoning comparable to advanced text RL systems in a tougher multimodal setting.

Motivation

While state-of-the-art multi-modal models like GPT-4o fail to consistently detect video misinformation, and template-finetuned systems such as QwenVL remain constrained by rigid response formats, Fact-R1 establishes a novel paradigm by enabling deep, structured reasoning tailored for misinformation detection.

News Title: Decision in TikTok appeal against EU law which could ban app is due.

News Video

News Audio

now the Fate of tiktok in North America could be decided this week in a decision due today a federal appeals Court is set to say whether or not they uphold an appeal by tiktok over a US law which demands the social media app is sold by its Chinese parent company bite dance so what can we expect today that's what I put to social media expert Jasmine engberg earlier it's not looking good for tiktok right now given the difficult day that it had in court in September and as you mentioned we ...

Human

The news title is 'Decision in TikTok appeal against EU law which could ban app is due.', the news audio is 'now the Fate of tiktok in North America could be decided this week in a decision due today a federal appeals Court is set to say whether or not they uphold an appeal ...', is this multimodal content manipulated?

GPT-4o

The images depict various users engaging with TikTok content—showing both individuals using the app and scenes that presumably relate to TikTok culture. **There are no obvious technical alterations or digital manipulations presented in the content that would imply deceit or misinformation...** Instead, it appears to visually support the news topic by highlighting TikTok features and user engagement, ... The answer is **No**. ❌

QwenVL

<answer>Yes.</answer> ?

Fact-R1

<think>The video description indicates that it features a news segment from BBC News about the upcoming decision in the TikTok appeal against a US law that could lead to the app's ban. The video shows scenes from TikTok, a reporter discussing the implications. The discussion centers around the potential ban in the US. However, the news title, "Decision in TikTok appeal against EU law which could ban app is due," mentions the EU law, which is inconsistent with the content of the video description. The video description clearly states that the discussion and the decision are about a US law, not an EU law. Thus, the news title does not accurately reflect the content of the video description as it misleadingly refers to EU law instead of the US law discussed in the video. In conclusion, the event discussed in the news title does not match the specific focus of the video description. n</think>n<answer>Yes.</answer> ✅

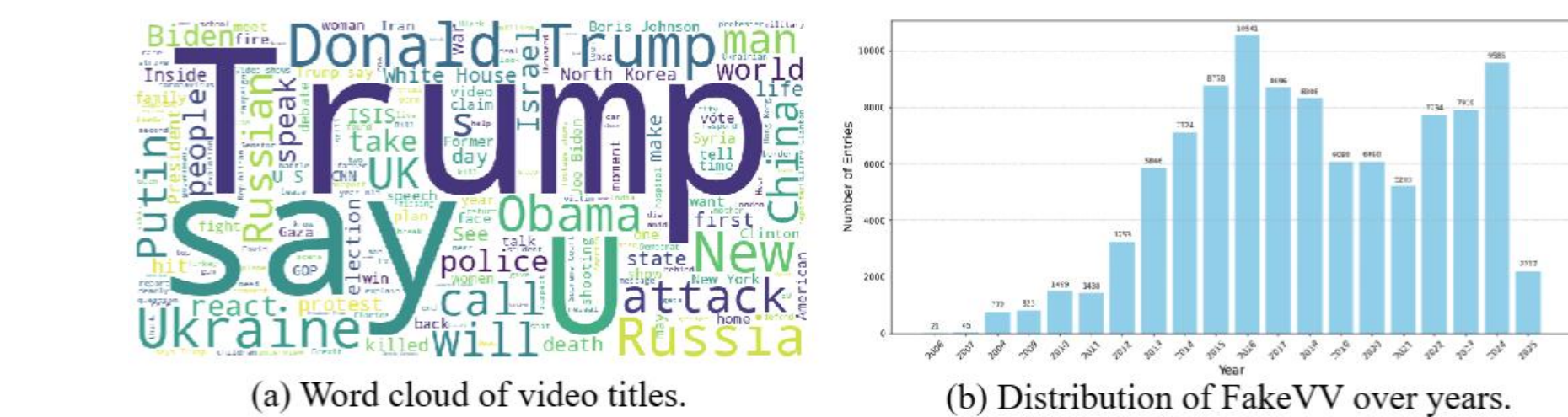


Figure 2: The statistics of FakeVV dataset.

Fact-R1 Pipeline

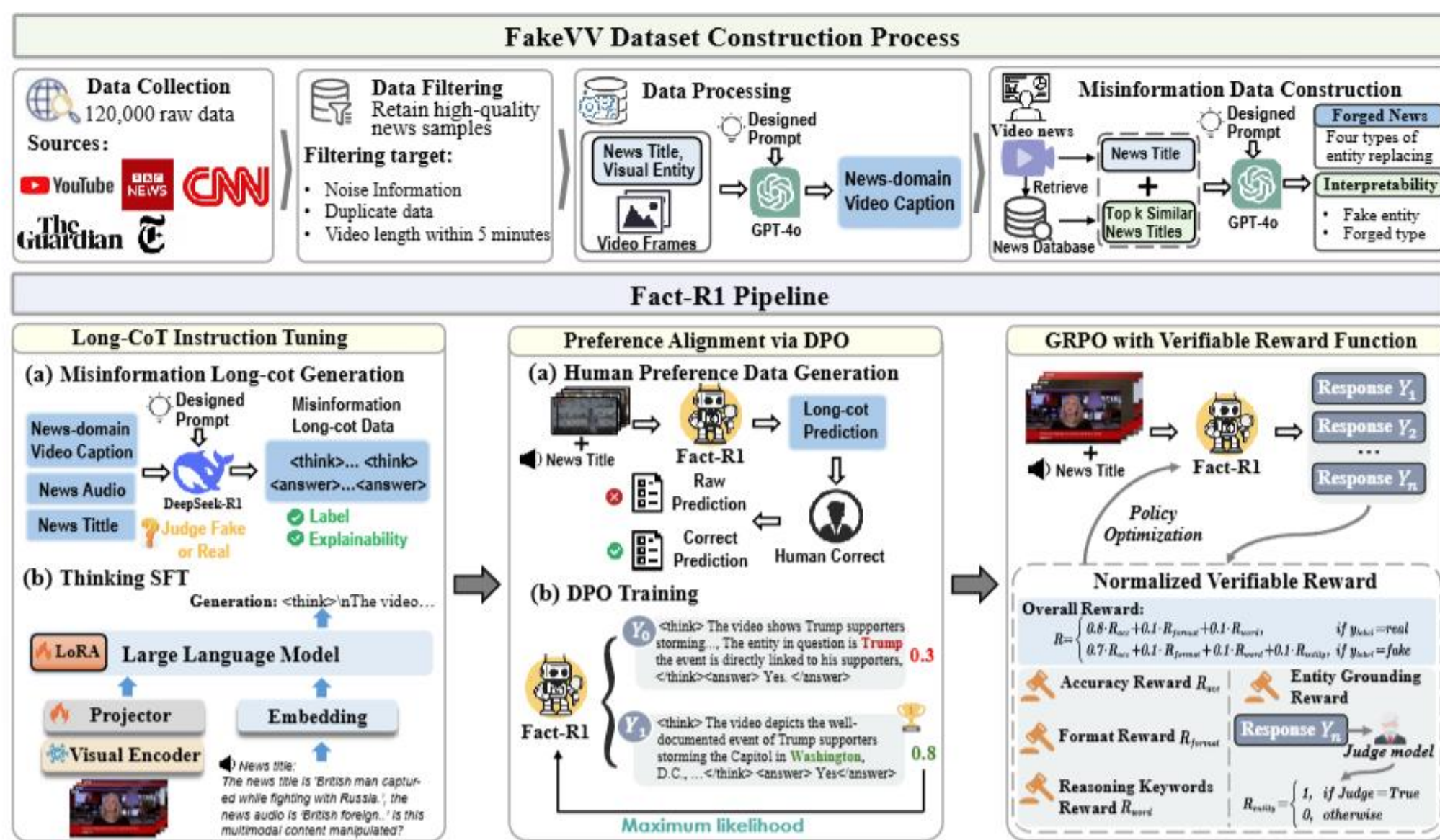
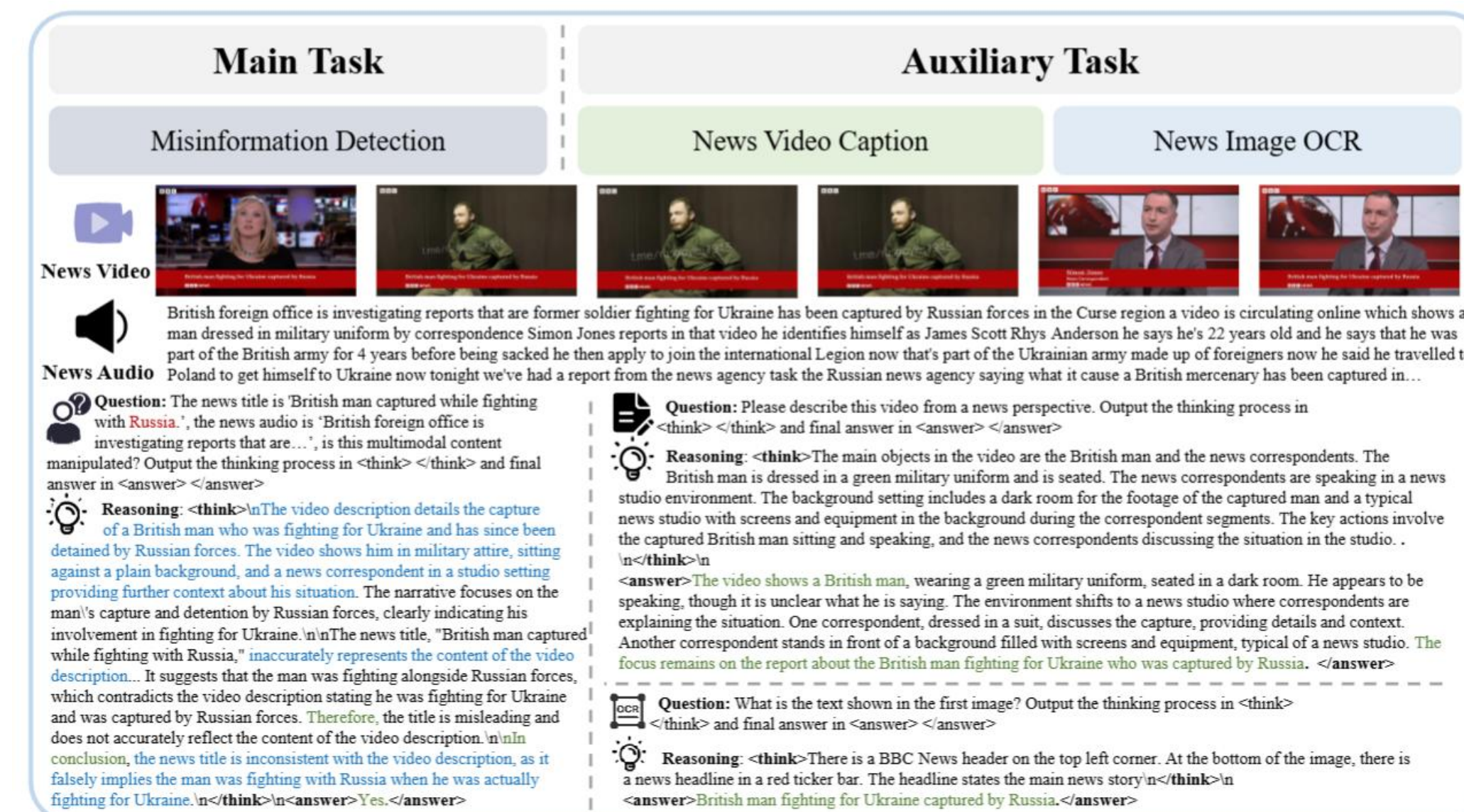


Figure 3: The overall architecture of the Fact-R1 is illustrated, with the upper part showing the FakeVV dataset construction process and the lower part presenting the training pipeline of Fact-R1.



Experiments

Table 2: Performance comparison on three real-world datasets. The best results are in **red bold**.

Model	FakeSV				FakeTT				FakeVV			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
BERT [10]	65.4	66.0	66.5	66.2	68.7	67.5	67.5	67.5	60.4	57.9	56.8	57.3
TikTec [30]	64.8	63.2	61.9	62.5	61.1	64.8	64.2	64.5	59.3	59.1	59.5	59.3
FANVM [9]	65.4	66.1	64.3	65.2	68.9	64.7	68.8	67.1	61.9	60.7	60.8	60.8
SV-FEND [24]	67.1	67.4	66.3	66.8	67.6	72.2	69.0	70.6	70.9	71.4	71.3	71.3
FakingRec [6]	69.5	69.7	70.4	70.0	71.0	71.9	72.0	72.0	72.1	72.4	71.6	72.0
Gemini2-thinking [35]	63.1	61.8	61.9	61.9	56.6	55.2	55.3	55.3	51.5	46.0	46.0	48.6
GPT-4o [1]	66.6	65.2	64.7	64.9	57.9	57.8	62.9	63.7	56.0	60.4	35.0	44.3
GPT-o1-mini [16]	60.3	57.7	56.5	57.1	52.5	51.6	51.7	51.7	47.5	46.9	37.6	41.8
DeepSeek-R1 [14]	61.8	60.4	60.3	60.3	49.8	52.6	52.5	52.6	53.5	58.1	25.2	35.1
Qwen2.5-VL-7B [3]	55.6	55.5	55.7	55.6	54.9	54.0	54.1	54.0	52.9	51.1	51.1	51.1
Qwen2.5-VL-72B [3]	57.6	55.4	55.2	55.3	59.2	58.1	58.3	58.2	54.0	60.0	24.0	34.3
QVQ-72B-preview [36]	60.8	59.0	58.8	58.9	58.1	54.0	52.8	53.4	53.5	52.6	52.6	52.6
InternVL2.5-8B [8]	49.8	52.6	52.5	52.6	43.9	44.0	44.0	44.0	53.5	58.5	24.0	34.0
InternVL2.5-78B-MPO [39]	57.5	53.0	52.0	52.5	59.2	57.1	56.7	56.9	54.0	60.0	24.0	34.3
Fact-R1	75.6	77.7	72.0	74.7	74.4	77.8	68.3	72.7	81.2	84.5	76.4	80.3

Table 3: Ablation study on the contribution of key components in Fact-R1.

Variant	FakeTT		FakeVV	
	ACC	F1	ACC	F1
w/o SFT	70.9	71.7	66.8	66.8
w/o DPO	72.1	72.4	80.4	79.9
w/o GRPO	70.7	70.6	79.8	79.1
w/o Audio	73.0	72.3	79.0	77.7
Fact-R1	74.4	72.7	81.2	80.3

Table 4: Evaluating the Impact of the Reward Function in Fact-R1.

Variant	FakeTT		FakeVV	
	ACC	F1	ACC	F1
w/o Keywords	71.1	72.0	78.6	79.9
w/o Entity	70.4	71.4	79.4	80.0
w/o Ocr	71.9	71.8	80.8	80.2
w/o Caption	71.6	71.6	75.5	77.7
Fact-R1	74.4	72.7	81.2	80.3

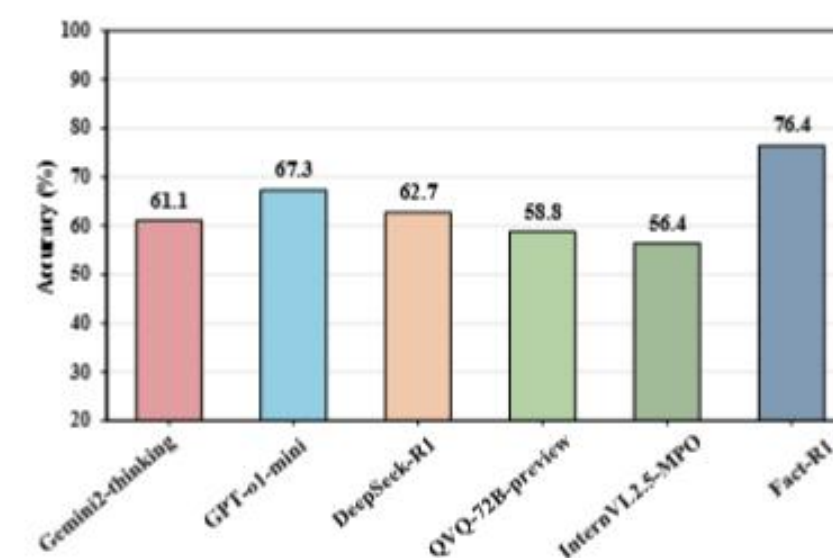
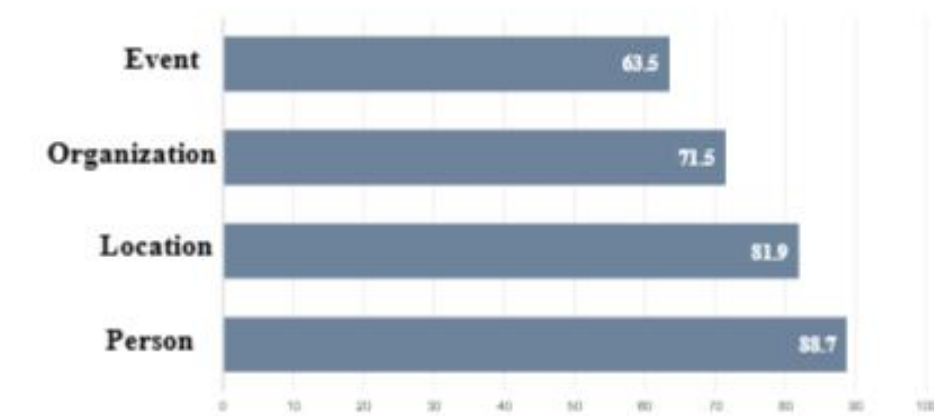


Figure 5: The interpretability accuracy of the outputs from the six models.

These results demonstrate that Fact-R1’s superior performance arises from its tailored misinformation reasoning design, combining long-CoT instruction tuning, DPO-based preference alignment, and GRPO-driven policy optimization.

Fact-R1 demonstrates strong reasoning ability by consistently describing the correct fake entities rather than overfitting to specific patterns.

Figure 6: Interpretability score distribution across forgery types.



FakeVV Dataset Construction

Table 1: Summary of datasets of video detection. Metadata refers to basic statistics such as # of likes/stars/edit time. “-” represents the exact time range is not found in the paper.

Dataset	Video	Title	Metadata	Comment	#Fake	#Real	Time Range	Interpretability	Construction Mode
FVC [22]	✓	✓	✓	✓	2,916	2090	-	✗	Web collection
VAVD [21]	✓	✓	✓	✓	123	423	2013/09-2016/10	✗	Web collection
YouTube-Covid [29]	✓	✓	✗	✓	113	67	2019/10-2020/04	✗	Web collection
TikTok-Covid [30]	✓	✓	✗	✗	226	665	-	✗	Web collection
TSC [43]	✓	✓	✗	✗	262	383	-	✗	Web collection
MYVC [9]	✓	✓	✗	✗	902	903	-	✗	Web collection
FakeSV [24]	✓	✓	✓	✓	1,827	1,827	2017/10-2022/02	✗	Web collection
FakeTT [6]	✓	✓	✓	✓	1,172	819	2019/05-2024/03	✗	Web collection
FakeVV (ours)	✓	✓	✓	✓	51,000	51,000	2006/11-2025/02	✓	Autotectonics