VRAG-DFD: Verifiable Retrieval-Augmentation for MLLM-based Deepfake Detection

Hui Han[1,2*], Shunli Wang[2*], Yandan Zhao[2], Taiping Yao[2], Shouhong Ding[2†]
[1]Shanghai Jiao Tong University [2]Tencent Youtu Lab

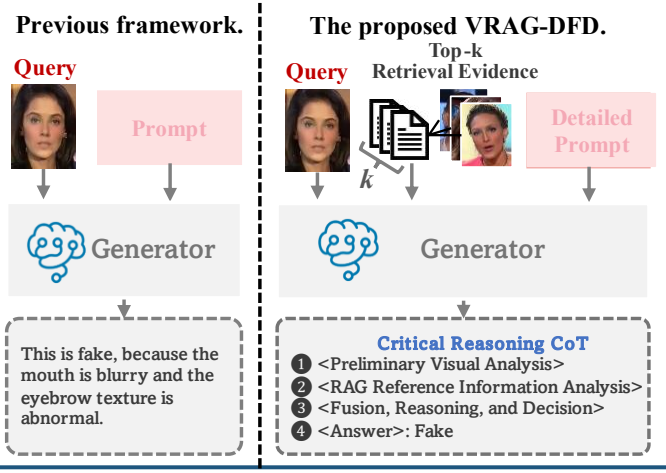
https://github.com/abigcatcat/VRAG-DFD

CVPR JUNE 3-7, 2026 DENVER COLORADO

Findings Track

Motivation

- **Traditional DFD Models**: Binary outputs with poor interpretability and generalization.
- **MLLM-based DFD Methods**: Architectural redundancy without high-level reasoning.
- **Knowledge Injection**: Coarse-grained, static knowledge fails to provide specific references.

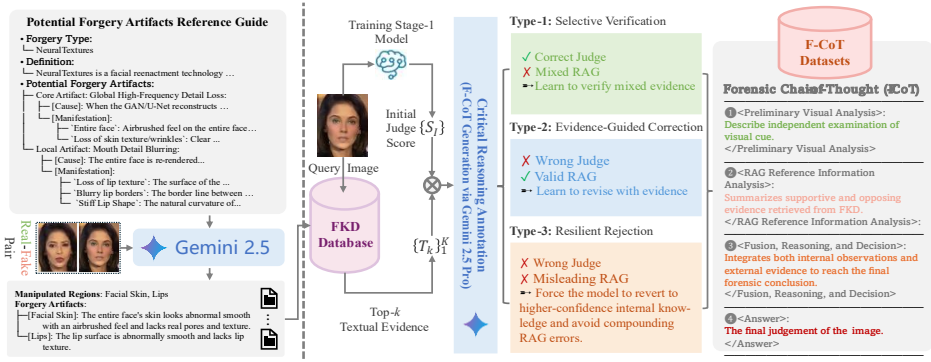


Contributions

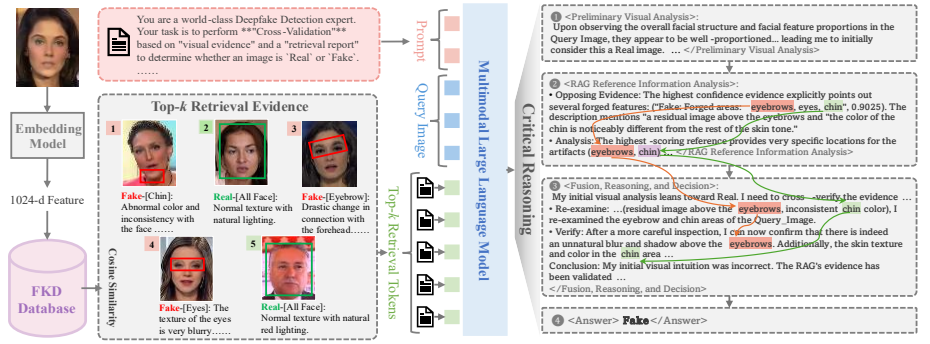
- **FKD & F-CoT:** A detailed forensic knowledge database and an innovative critical-thinking instruction dataset.
- **VRAG-DFD**:A novel MLLM-based critical reasoning framework with an efficient optimization strategy.
- **SOTA Performance:** Achieved state-of-the-art and highly competitive generalization across multiple DFD benchmarks.

Method

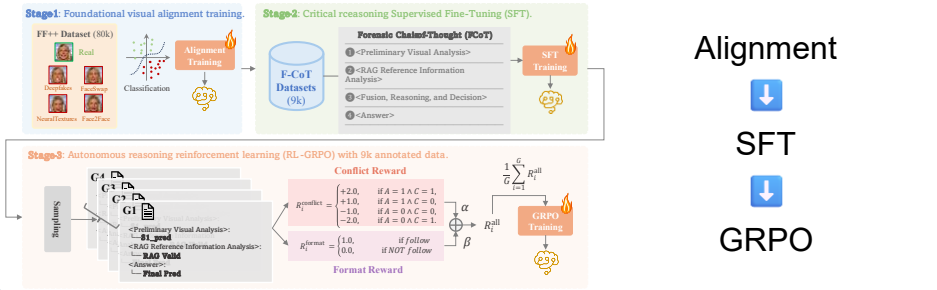
Dataset Construction: FKD & F-CoT



Architecture: Dynamic Forensic Retriever & Critical Reasoning MLLM



Three Stage Training Pipeline



Experiments

Cross-dataset Generalization Performance

Method	Venue	Explainable	Research Topic	CDF-v1	CDF-v2	DFDC	FFIW	WDF
F^3-Net [36]	ECCV 2020	✗	Frequency + CNN	81.11	77.92	67.35	70.11	72.80
LTW [43]	AAAI 2021	✗	Meta Training	-	77.14	69.00	76.63	-
SPSL [28]	CVPR 2021	✗	Phase Spectrum	-	79.90	77.00	79.40	70.20
SRM [29]	CVPR 2021	✗	Attention + FPN	-	84.00	69.50	80.60	72.20
PCL+I2G [54]	ICCV 2021	✗	Blending	-	90.03	67.52	-	-
SBI [41]	CVPR 2022	✗	Blending	93.44	93.18	72.42	84.83	-
CORE [33]	CVPR 2022	✗	Augmentation	-	80.90	72.10	71.00	72.40
DCL [44]	AAAI 2022	✗	Contrastive Learning	-	82.30	76.71	71.14	71.14
SeeABLE [21]	ICCV 2023	✗	Contrastive Learning	-	87.30	75.90	-	-
F^2Trans [31]	TIFS 2023	✗	Attention + Wavelet	86.29	89.87	-	-	-
AltFreezing [48]	CVPR 2023	✗	Spatial + Temporal	88.48	89.50	64.75	-	-
AUNet [5]	CVPR 2023	✗	Facial AU	-	92.77	73.82	81.45	-
UCF [49]	ICCV 2023	✗	Disentanglement	86.08	83.73	75.11	69.70	77.42
LAA-Net [32]	CVPR 2024	✗	Attention + FPN	-	95.40	-	-	-
FreqBlender [56]	NeurIPS 2024	✗	Frequency + Blending	-	94.59	74.59	86.14	-
UDD [10]	AAAI 2025	✗	Token Shuffle + Mixing	-	93.10	81.20	-	-
LESB [42]	WACVW 2025	✗	Blending	-	93.13	71.98	83.01	-
Effort [50]	ICML 2025	✗	Fine-tune CLIP-ViT	96.05	95.60	84.30	92.10	84.80
χ^2-DFD [7]	NeurIPS 2025	✓	MLLM + Detector	-	95.50	**85.30**	86.70	86.40
KFD [52]	ICML 2025	✓	MLLM + Static Knowledge	97.62	94.71	79.12	-	-
M2F2-Det [13]	CVPR 2025	✓	MLLM + Static Knowledge	-	95.10	-	88.70	-
Ours	CVPRF 2026	✓	MLLM + **Dynamic RAG**	**99.60**	**95.97**	81.82	**93.49**	**88.96**

Ablation Study of RAG & Retriever

Metrics	Test Set AUC CDF1	CDF2	DFDC	FFIW	WDF
w/o RAG Module	91.13	87.85	72.42	78.62	71.62
VRAG-DFD	**99.60**	**95.97**	**81.82**	**93.49**	**88.96**

Retrievers	Test Set AUC CDF1	CDF2	DFDC	FFIW	WDF	Avg.
Effort [50]	99.60	**95.97**	81.82	93.49	**88.96**	91.62
CLIP-LoRA [17]	**100.00**	95.77	**82.47**	**94.00**	88.42	**92.13**

Ablation Study of Three-stage Training Pipeline

Explanation Quality Evaluation

Metrics	Test Set AUC CDF1	CDF2	DFDC	FFIW	WDF
Stage-1	96.77	94.03	78.54	90.89	87.63
Stage-1&2	91.53	**95.8**	**81.58**	**92.37**	**89.13**
VRAG-DFD	**99.60**	**95.97**	**81.82**	**93.49**	**88.96**

Model	GPT-4o	Gemini2.5 Pro	Avg.
GPT-4o	4.60	3.25	3.93
Gemini2.5 Pro	7.31	7.02	7.16
VRAG-DFD	**7.55**	**7.78**	**7.66**

Contact

*Equal Contribution:
hanhui99@sjtu.edu.cn shunliwang@tencent.com
†Corresponding author:
ericshding@tencent.com