From Intuition to Investigation: A Tool-Augmented Reasoning MLLM Framework for Generalizable Face Anti-Spoofing (Supplementary Material)

Haoyuan Zhang[1,2,3 †], Keyao Wang[3 †], Guosheng Zhang[3], Haixiao Yue[3], Zhiwen Tan[3], Siran Peng[1,2], Tianshuo Zhang[1,2], Xiao Tan[3], Kunbin Chen[3], Wei He[3], Jingdong Wang[3], Ajian Liu[1,2], Xiangyu Zhu[1,2], Zhen Lei[1,2,4,5 *]

[1]SAI, UCAS; [2]MAIS, CASIA; [3]Baidu Inc; [4]CAIR, HKISI, CAS; [5]M.U.S.T

{zhanghaoyuan2023,pengsiran2023,ajian.liu,xiangyu.zhu,zhen.lei}@ia.ac.cn
{wangkeyao,zhangguosheng,yuehaixiao,tanzhiwen,tanxiao01,chenkunbin,hewei06}@baidu.com
tianshuo.zhang@nlpr.ia.ac.cn,wangjingdong@outlook.com

A. Data Annotation Pipeline

A.1. Data Selection

In the ToolFAS-16K dataset, we cover the ten spoof types in CelebA-Spoof [15]. The sample of each spoof type is shown in Figure S1.

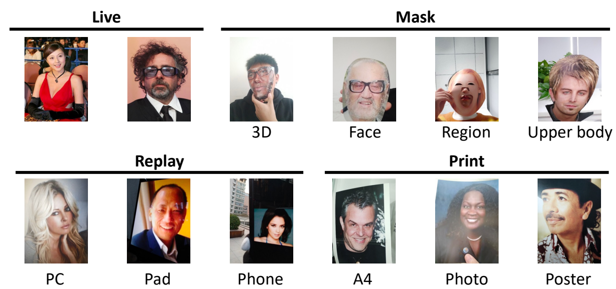

Figure S1. The detailed spoof types in ToolFAS-16K.

A.2. Annotation System Prompt Design

To ensure consistent reasoning behavior during annotation, we design a structured system prompt $\mathcal{S}$ that explicitly defines the model's role, objectives, and interaction rules. Formally, the system prompt is composed of five components:

$$\mathcal{S} = \{\mathcal{P}_{\text{role}}, \mathcal{P}_{\text{principle}}, \mathcal{P}_{\text{workflow}}, \mathcal{P}_{\text{tools}}, \mathcal{P}_{\text{conclusion}}\},$$

where $\mathcal{P}_{\text{role}}$ specifies the model's *Role & Mission*, $\mathcal{P}_{\text{principle}}$ defines the *Core Principles* for reasoning, $\mathcal{P}_{\text{workflow}}$ encodes the *Behavioral Rules* for multi-turn interactions, $\mathcal{P}_{\text{tools}}$ enumerates the *Available Tools*, and $\mathcal{P}_{\text{conclusion}}$ provides instructions for generating the *Final Conclusion*. This structured design ensures that the annotator model follows a stable reasoning pattern alternating between internal thought and tool invocation. The complete system prompt is shown in Figure S3.

A.3. Expert Model

In our experiments, these expert models achieve around 90% accuracy on the CelebA-Spoof domain without using RGB inputs, and each shows distinctive sensitivity to certain spoof types, validating their reliability.

Architecture. We use a simple CNN architecture to train expert model $\{\mathcal{E}_1, \cdots, \mathcal{E}_K\}$ for each tool with only the tool result as input and the binary classification as output. The detailed architecture is illustrated in Figure S4.

Implementation. We use the CelebA-Spoof dataset [15] to train the expert model for each tool. All images are aligned, cropped, and resized to 224×224. We randomly select 5,000 identities from CelebA-Spoof and split them into 4,000 for training and 1,000 for testing. All samples belonging to each identity are included. The models are trained for 10 epochs using the Adam optimizer with a learning rate of 0.001.

Accuracy. We evaluate each expert model using a fixed threshold of 0.5, and the results are presented in Figure S2. As shown, the average accuracy of all tools exceeds 80%, demonstrating that even when only tool results are used as input, they possess considerable discriminative and generalization capabilities. Moreover, different visual tools show varying effectiveness across spoof types, highlighting the necessity of employing diverse visual tools.

[†]Equal contribution; [*]Corresponding author
Work done during Haoyuan Zhang's internship at Baidu Inc.

Table S1. Comparison with the closest and SOTA FAS methods in leave-one-out protocol on MSU-MFSD (M), CASIA-FASD (C), ReplayAttack (I), and OULU-NPU (O) datasets. Avg. indicates the average performance across four experimental scenarios. The scores presented in bold represent the best performance.

Methods	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O		Avg.
	HTER(%)	AUC	HTER(%)	AUC	HTER(%)	AUC	HTER(%)	AUC	HTER(%)
FGHV [7]	9.17	96.92	12.47	93.47	16.29	90.11	13.58	93.55	12.88
GDA [18]	9.20	98.00	12.20	93.00	10.00	96.00	14.40	92.60	11.45
PatchNet [11]	7.10	98.46	11.33	94.58	13.40	95.67	11.82	95.07	10.91
SSAN [12]	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63	9.82
IADG [19]	5.41	98.19	8.70	96.40	10.62	94.50	8.86	97.14	8.40
UDG-FAS [8]	5.95	98.47	9.82	96.76	5.86	98.62	10.97	95.36	8.15
TTDG [20]	4.16	98.48	7.59	98.18	9.62	98.18	10.00	96.15	7.84
SA-FAS [10]	5.95	96.55	8.78	95.37	6.58	97.54	10.00	96.23	7.83
DiVT-M [5]	2.86	99.14	8.67	96.92	3.71	99.29	13.06	94.04	7.08
GAC-FAS [4]	5.00	97.56	8.20	95.16	4.29	98.87	8.60	97.16	6.52
FLIP [9]	4.95	98.11	0.54	99.98	4.25	99.07	2.31	99.63	3.01
CFPL-FAS [6]	1.43	99.28	2.56	99.10	5.43	98.41	2.50	99.42	2.98
I-FAS [14]	**0.32**	**99.88**	0.04	99.99	3.22	98.48	**1.74**	99.66	1.33
TAR-FAS (Ours)	2.86	99.35	**0.00**	**100.00**	**0.43**	**99.86**	1.91	**99.72**	**1.30**

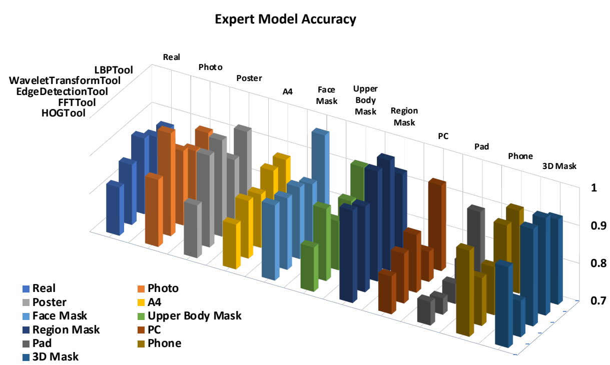

Figure S2. The detailed accuracy of expert model for each tool.

B. Tool-Aware FAS Training Pipeline

B.1. Training Prompt Design

System Prompt. The system prompt contains *Role & Mission* description, tool description in XML format. The *Role & Mission* description is shown as follows:

> **System Prompt**
>
> ## Role & Mission
> You are a face forensics expert. Your mission is to classify an image as either 'Real' or 'Spoof' by analyzing evidence strictly within the **facial region**, focusing only on physical presentation attack.

While, the tool description follows Json Schema and adopt the Hermes template used in Qwen2.5-VL [1].

First-Round Query. The first-round query ask the model to classify whether the given image is of a real person.

> **First-Round Query**
>
> Is this photo of a real person? (Do not use any tools)

Tool Guidance and Format Declaration. In the second round, user prompt conatains tool guidance and format declaration.

> **Tool Guidance**
>
> Wait, you should re-examine the image and give the final answer (use tools if needed).

> **Format Declaration**
>
> Think first, call tools if needed, then answer. Format strictly as: <think> ... </think> <tool_call> ... </tool_call> (if tools needed) <answer>(<Spoof>/<Real>)</answer>

After the second round, the user prompt will contain a tool result image and format declaration.

B.2. DT-GRPO Format Constrain

Fast answer format. A classification will be give in the first-turn with a strict format of <CLS><reason></reason>. A wrong format will cause to a penalty of -1.

Reasoning format. The reasoning format for each turn should be <think></think> + <tool_call></tool_call> or <think></think> + <answer><CLS></answer>. Wrong format or invalid tool call will get a format penalty of -1.

System Prompt

Role & Mission
You are a face forensics expert. Your mission is to classify an image as either 'real' or 'spoof' by analyzing evidence strictly within the **facial region**, focusing **only on physical presentation attacks** (e.g., printed photos, screen displays).
You will be given a hint early in the conversation. Do not mention the hint when making your decision. Your final classification must match the hint, and be supported by image evidence and tool-based analysis.

Core Principle
Concentrate solely on **physical attack artifacts**. These include: - **Semantic clues**: unnatural flatness, rigid expressions, lack of 3D structure, unnatural reflections.
- **Pixel-level clues**: print texture, Moiré patterns, screen glare, paper/screen surface noise.
Since input images are cropped and aligned, **do not consider black borders and compression artifacts as spoof clues**.

Workflow & Behavioral Rules
1. Begin with a brief visual scan of the **facial region**. **Ignore black borders and all context outside the face.**
2. If needed, call **ONE tool at a time** to test a specific hypothesis, either to look for signs of physical attack, or to confirm their absence.
- Each tool request must include a clear expectation (what you're testing for). 3. When you receive tool results, you may receive an **Expert Judgment** on the result.
- You may consider the expert's interpretation as a reference, but **must perform your own independent analysis**.
- Your reasoning should not blindly follow the expert; only adopt it when it aligns with your observations.
4. When you are confident, provide your conclusion.

Available Tools
- **ZoomInTool**: Inspect local details for physical (print/screen) or digital (blending) artifacts.
- **FFTTool (Fast Fourier Transform):** Visualizes the image's frequency domain. Used to detect periodic patterns like screen Moiré effects or subtle artifacts from digital generation.
- **EdgeDetectionTool**: Find inconsistent edges from cutouts or digital blending.
- **LBPTool**: Analyze skin texture for unnatural or synthetic patterns.
- **WaveletTransformTool**: Find subtle digital tampering or noise mismatches via multi-scale analysis.
- **HOGTool**: Check facial structure gradients, which are often disrupted in physical attacks.

Final Conclusion
Your conclusion must be 'Real' or 'Spoof'.

Figure S3. Complete System prompt.

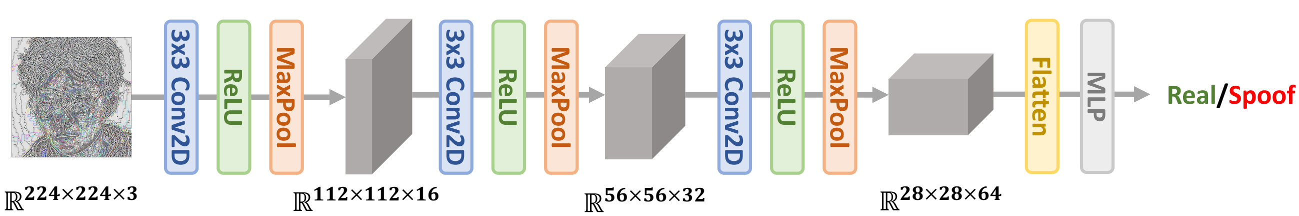

Figure S4. The detailed architecture of expert model for each tool. The input of expert model is only tool result (without RGB image). The 'MaxPool' perform double downsampling. The 'MLP' denotes two linear layer with a ReLU achitation function.

C. Experiments

To further demonstrate the effectiveness of TAR-FAS, we conduct several additional experiments. We first present quantitative results under the widely used leave-one-out evaluation protocol. We then analyze the behavior of our tool-augmented reasoning model, which is trained under the One-to-Eleven evaluation protocol used in the main paper. Finally, we provide additional reasoning examples across different datasets.

C.1. ICMO Protocol

C.1.1. Implementation Details

To further evaluate the cross-domain performance of TAR-FAS, we conduct experiments on the widely used four leave-one-out settings and compare its performance with the latest state-of-the-art approaches. This protocol includes the REPLAY-ATTACK (I) [3], CASIA-FASD (C) [16], MSU-MFSD (M) [13], and OULU-NPU (O) [2] datasets. In each setting, three datasets are used for training (source domains) and the remaining one for testing (target domain). We apply our proposed tool-aware FAS training strategy using ICMO datasets as source domain in both the FAS knowledge transfer stage and the DT-GRPO training stage. All other training hyperparameters remain the same as those described in the implementation details of the main paper.

C.1.2. Results

As shown in Table S1, our method achieves a significant performance advantage over all single-modal methods, outperforming them by a substantial margin. This advantage highlights the effectiveness of multimodal learning in enhancing model generalization. Furthermore, our approach also surpasses recent CLIP-based methods, such as FLIP [9] and CFPL-FAS [6], as evidenced by the average HTER reduction to 1.30%, compared to 2.98% and 3.01%, respectively. Compared with the previous MLLM-based method I-FAS [14], our model further improves performance by a small margin, achieving state-of-the-art (SOTA) results. This improvement suggests that incorporating external visual tools enables the model to capture fine-grained visual cues more effectively, thereby enhancing both the robustness and interpretability on FAS task.

C.2. One-to-Eleven Protocol

C.2.1. Fast Answer Performance

Table S2. Performance comparison of fast and reasoning answer.

	Results	
	HTER(%)	AUC
Fast Answer	10.54	95.06
Reasoning Answer	**7.54**	**96.67**

To emphasize the effectiveness of tool-augmented reasoning, we compare the performance of the fast answer and the reasoning answer. As shown in Table S2, the reasoning answer outperforms the fast answer with a margin of 1.61% HTER, demonstrating the effectiveness of incorporating external visual tools.

C.2.2. Reasoning Accuracy in Training Process

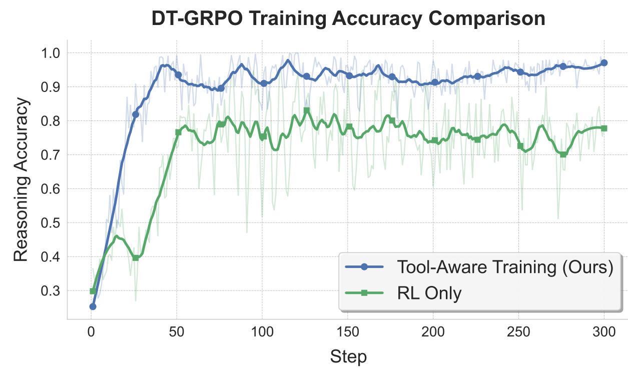

Figure S5. The reasoning accuracy comparison between our *tool-aware training* and *RL only training* during training process.

We compare our tool-aware FAS training strategy with a RL only training strategy using the same tool-diversity reward. As shown in Figure S5, the reasoning accuracy of our method overpass RL only strategy with a substantial margin. The result demonstrate the necessity of knowledge transfer and tool-call format inject process which layes a solid foundation for DT-GRPO.

C.2.3. Effectiveness of DT-GRPO

Table S3. Performance comparison of DT-GRPO and ST-GRPO (Single-Tool GRPO) in DeepEyes [17].

RL	Results	
	HTER(%)	AUC
ST-GRPO [17]	9.98	94.41
DT-GRPO (Ours)	**7.54**	**96.67**

To demonstrate the proposed DT-GRPO, we compare the performance against Single-Tool GRPO (ST-GRPO) used in Deepeyes [17] which enhance multimodal reasoning performance. The reward of ST-GRPO can be formulated as:

$$R_{ST\text{-}GRPO} = R^{fmt}_{rsn} + R^{acc}_{rsn} + \mathbb{I}_{tool} \cdot \mathbb{I}_{R^{acc}_{rsn}>0} \quad , \qquad \text{(S1)}$$

where $R^{fmt}_{rsn} \in \{-1, 0\}$ denotes reasoning format reward, $R^{acc}_{rsn} = \mathbb{I}_{CLS=label}$ denotes reasoning accuracy reward, $\mathbb{I}_{tool} \in \{0, 1\}$ indicates whether tool is called. Notably, only ZoomIn tool are allowed in ST-GRPO.

As shown in Table S3, our proposed DT-GRPO outperforms the ST-GRPO with a substantial margin, gaining generalization performance from diverse tool and tool-diversity reward. Thus demonstrate that our DT-GRPO enables the model to autonomously learn efficient and adaptive diverse tool-use.

C.2.4. Impact of Maximum Tool-Call Time

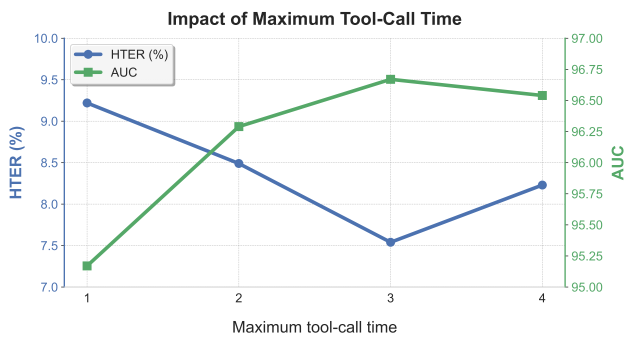

Figure S6. Impact of maximum tool-call time. Increasing the maximum tool-call time from 1 to 3 improves performance, while setting it to 4 leads to a slight drop due to longer rollouts and higher sample uncertainty.

We further study the impact of the maximum tool-call limit. During training, this limit is controlled by the maximum number of turns allowed in the rollout process. The relationship between the tool-call limit and model performance is shown in Figure S6. As illustrated, increasing the limit from 1 to 3 consistently improves performance. However, setting the limit to 4 results in a slight performance drop, likely due to longer rollout and increased sample uncertainty. Therefore, we set the maximum tool-call time to 3 for optimal performance. This experiment further demonstrates the importance of external visual tools in achieving robust FAS performance.

C.3. Additional Reasoning Samples

To further exhibit the tool-augmented reasoning process of TAR-FAS, we choose one sample from each of the eleven evaluated datasets. The results are shown in Figure S7-S17.

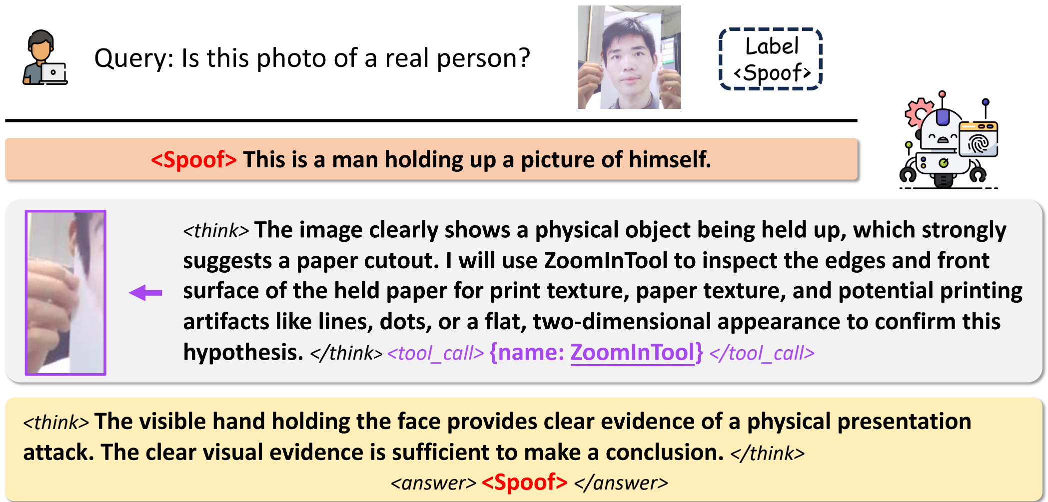

Figure S7. Sample of CASIA-MFSD.

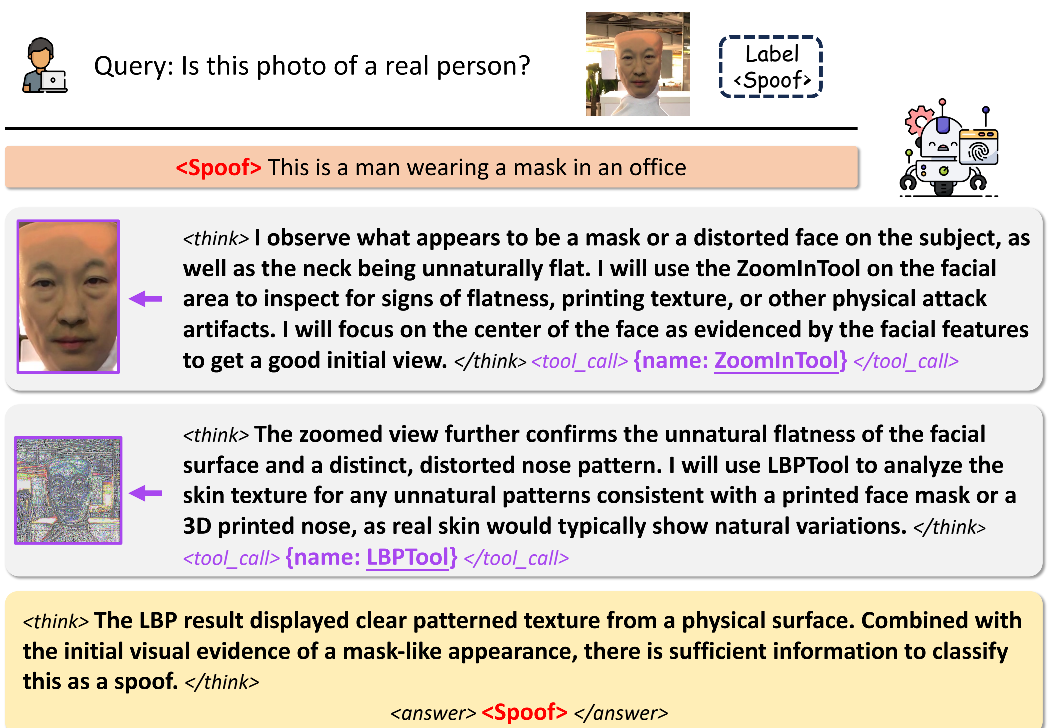

Figure S8. Sample of CASIA-SURF-3DMask.

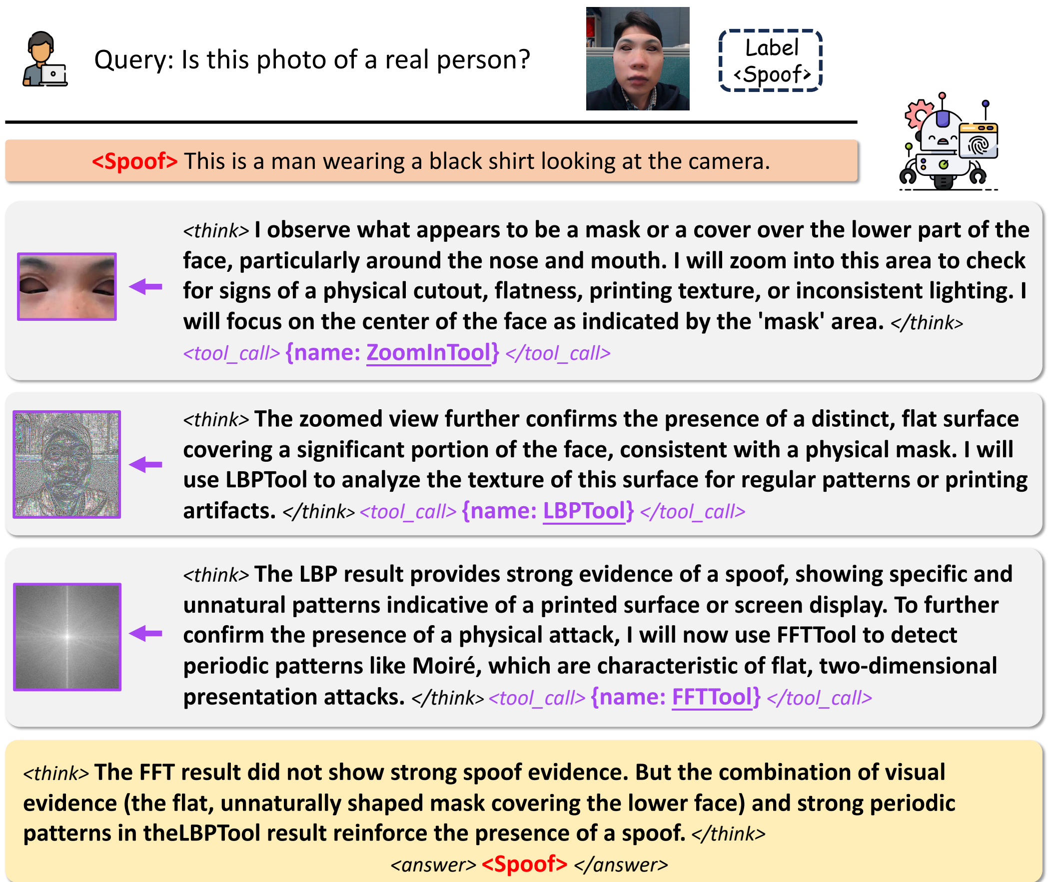

Figure S9. Sample of HKBU-MARs-V1+.

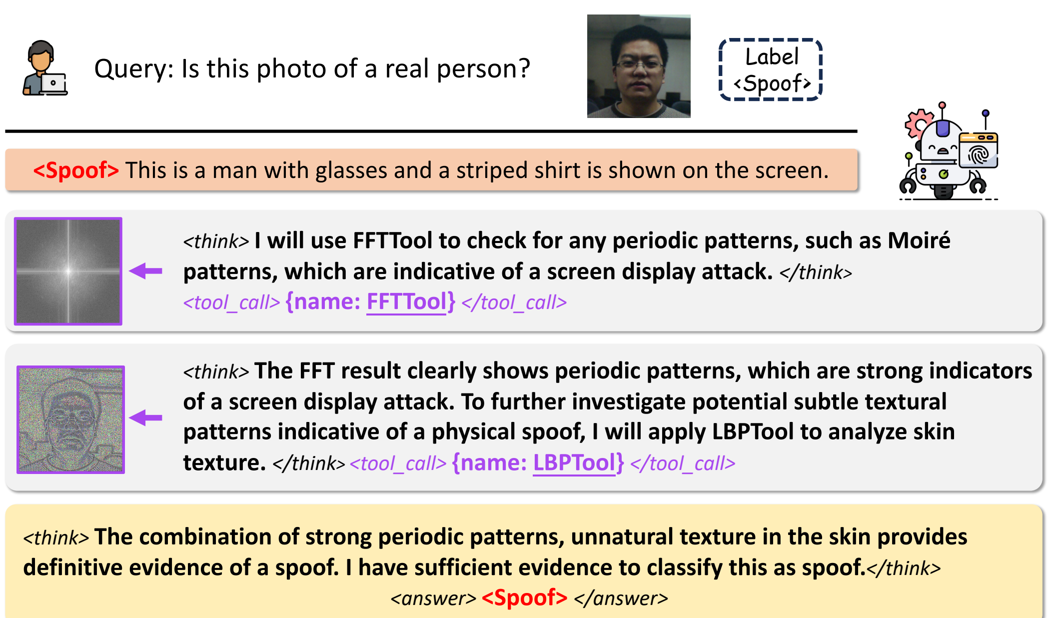

Figure S10. Sample of MSU-MFSD.

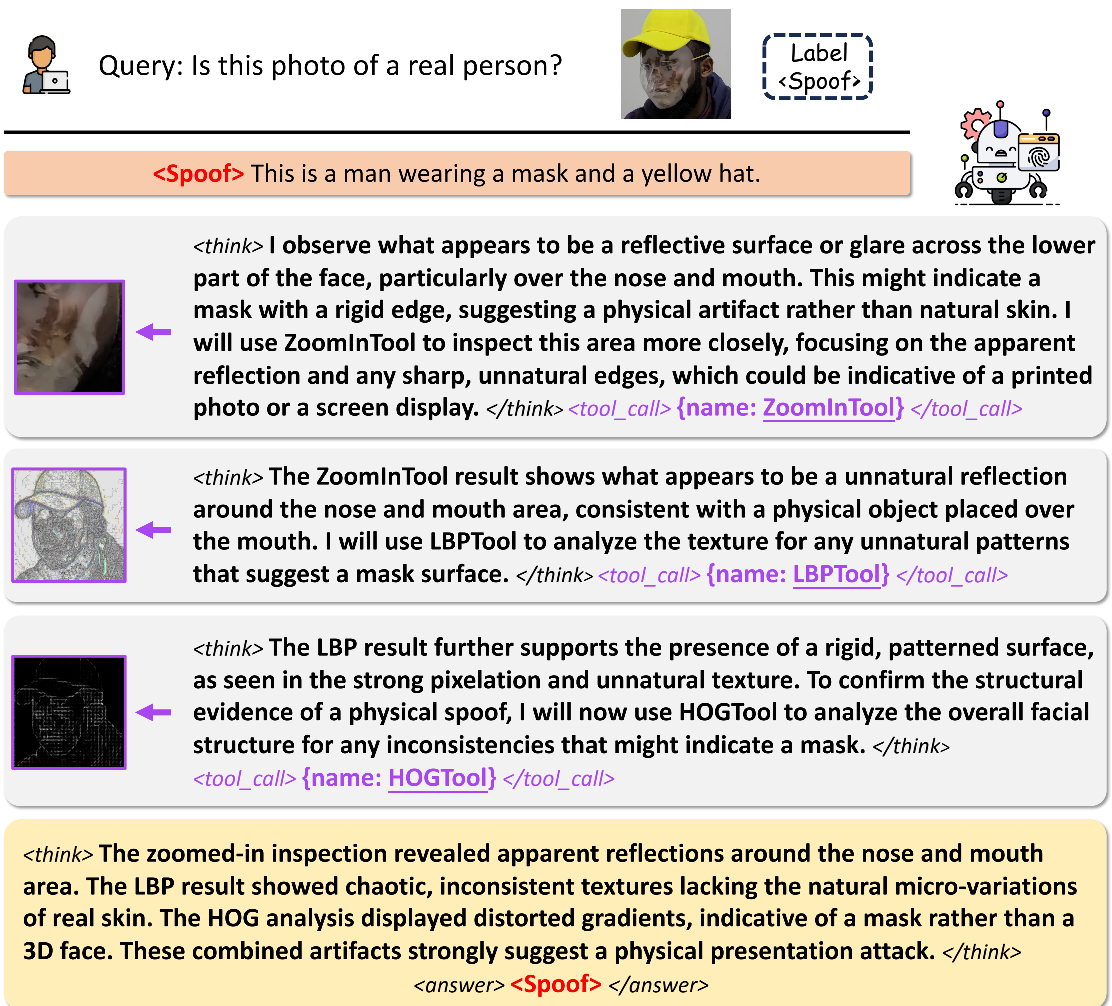

Figure S11. Sample of HiFiMask.

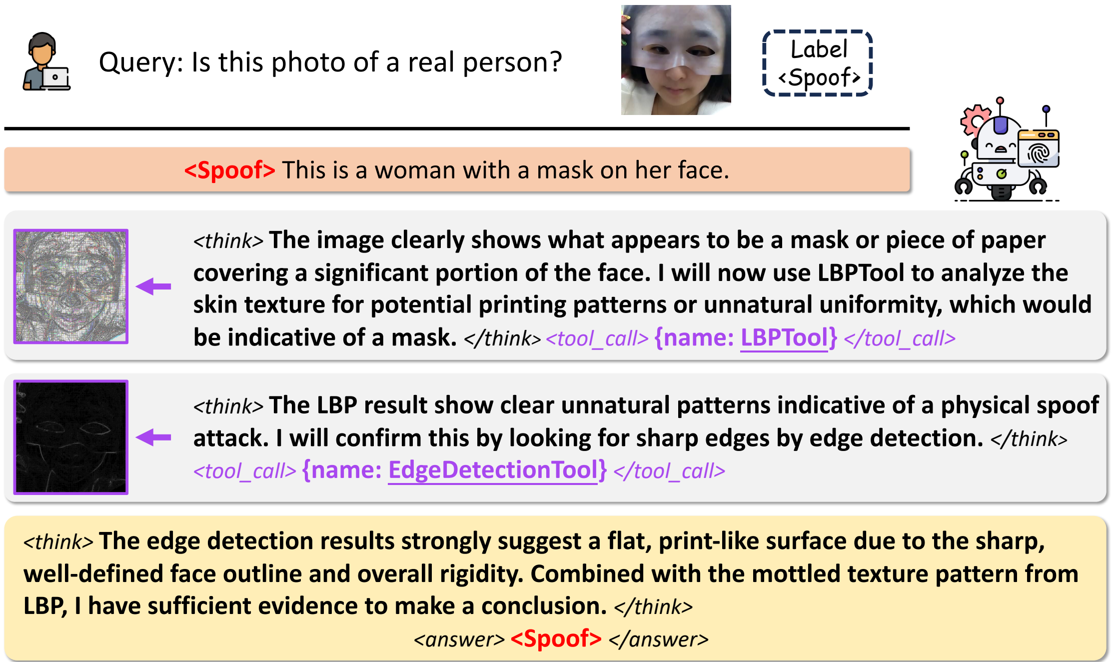

Figure S12. Sample of Rose-Youtu.

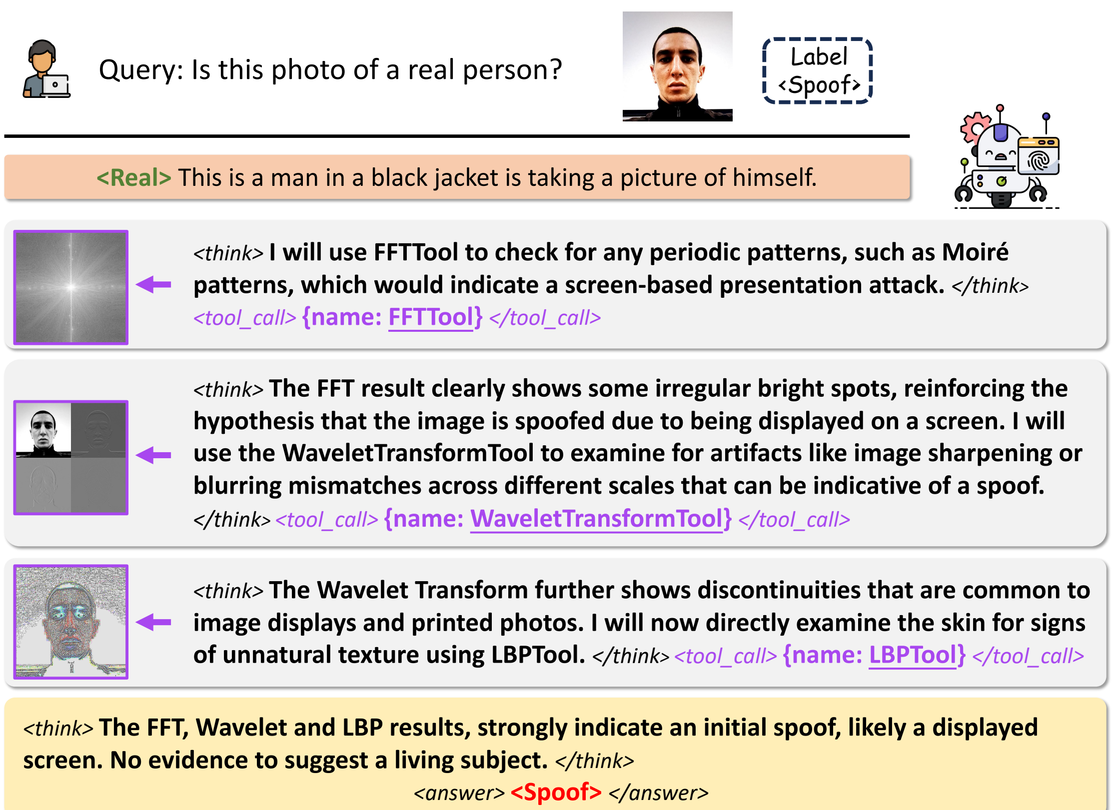

Figure S13. Sample of OULU-NPU.

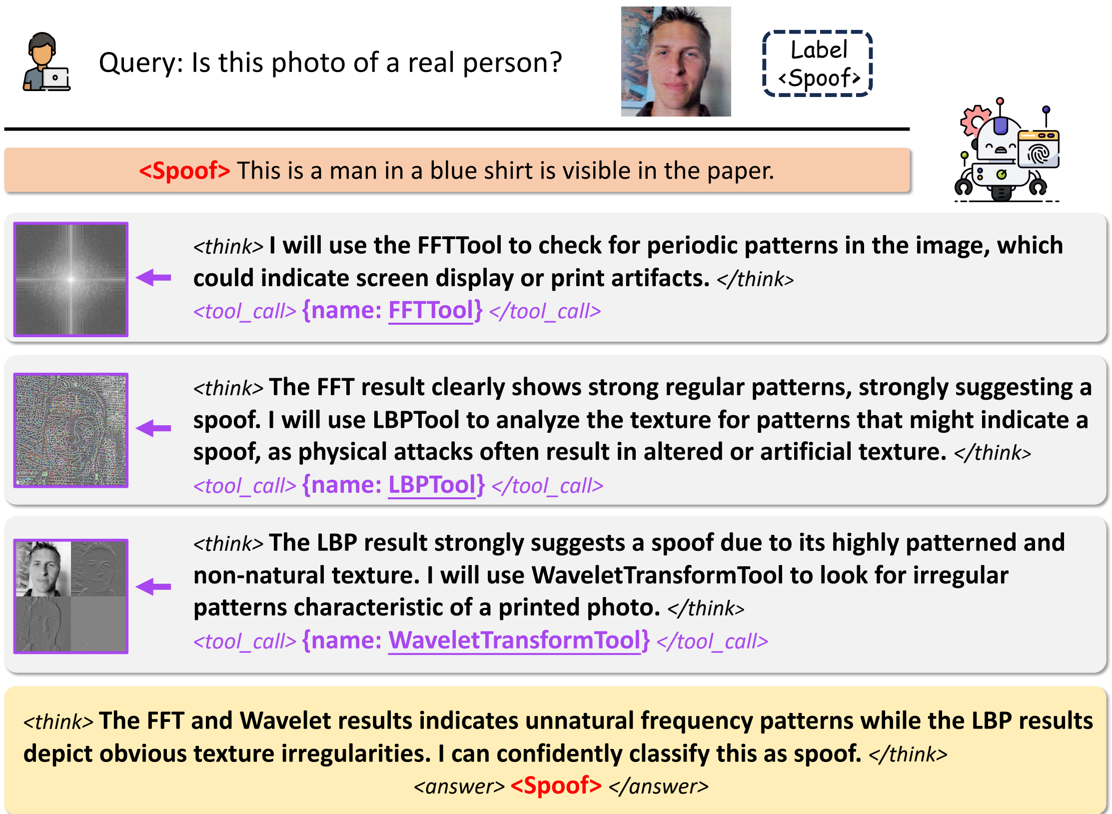

Figure S14. Sample of REPLAY-ATTACK.

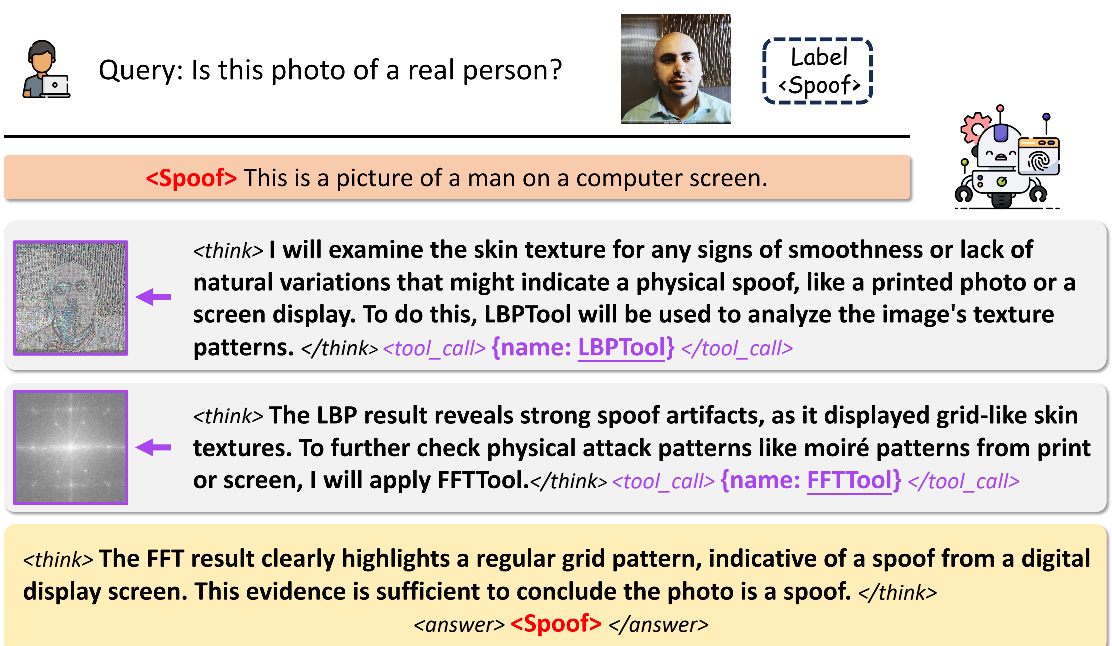

Figure S15. Sample of SIW.

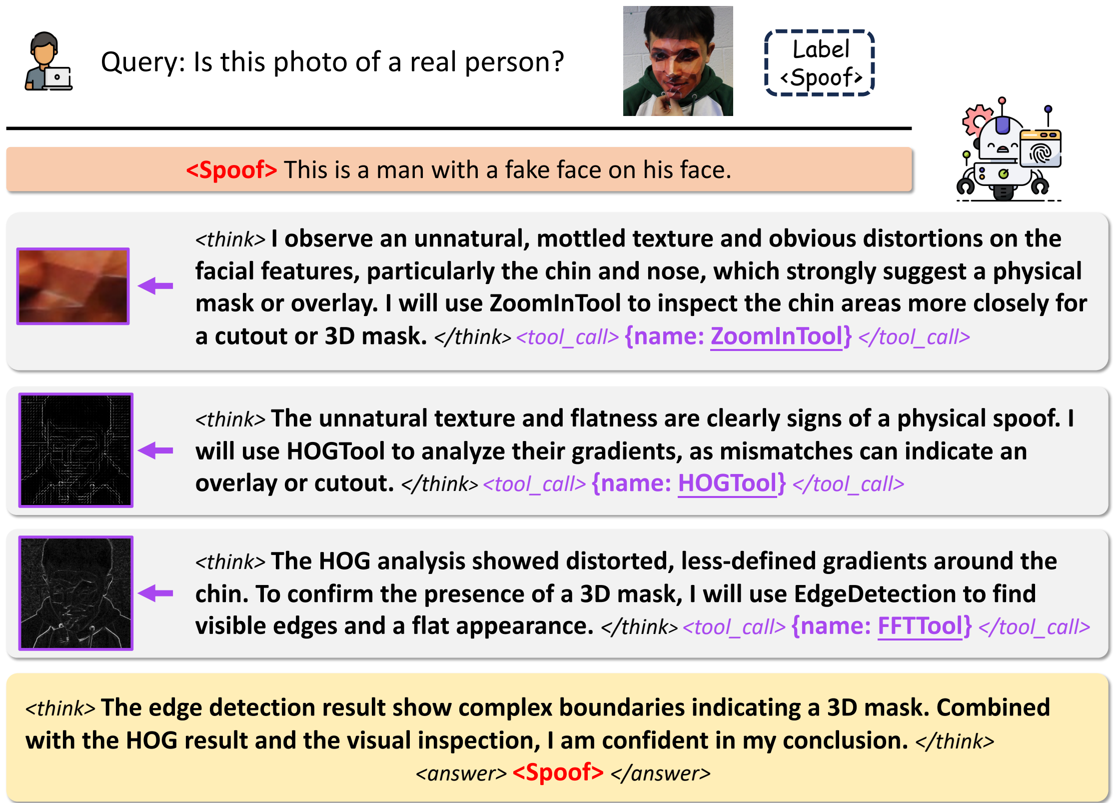

Figure S16. Sample of SIW-M-V2.

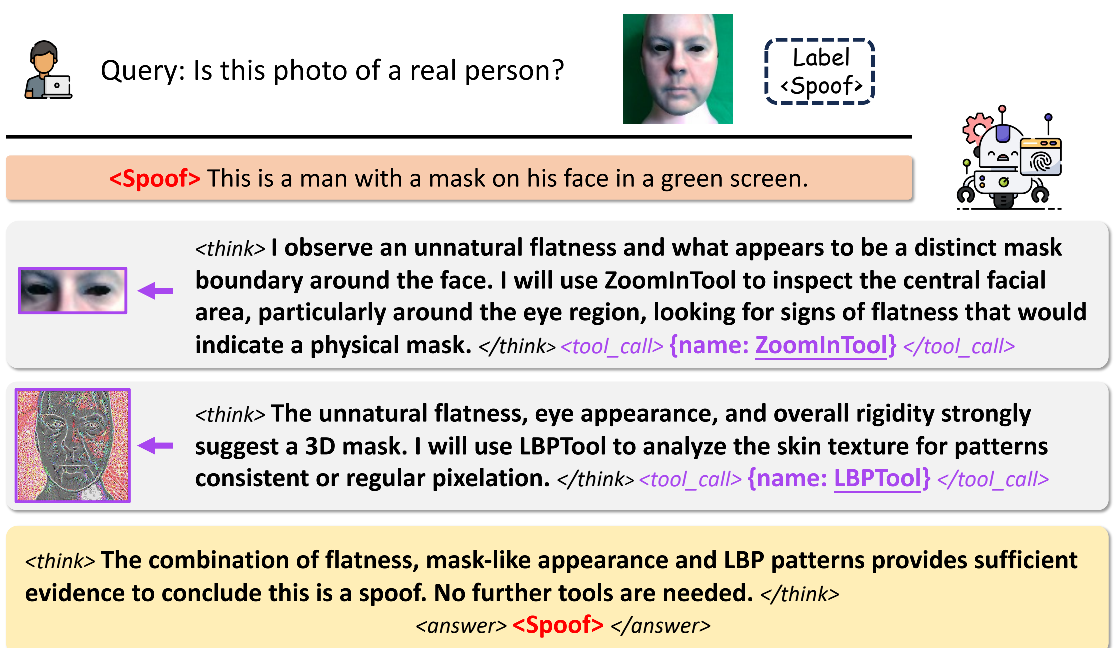

Figure S17. Sample of WMCA.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 612–618. IEEE, 2017. 4

[3] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012. 4

[4] Binh M Le and Simon S Woo. Gradient alignment for cross-domain face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 188–199, 2024. 2

[5] Chen-Hao Liao, Wen-Cheng Chen, Hsuan-Tung Liu, Yi-Ren Yeh, Min-Chun Hu, and Chu-Song Chen. Domain invariant vision transformer learning for face anti-spoofing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6098–6107, 2023. 2

[6] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 222–232, 2024. 2, 4

[7] Shice Liu, Shitao Lu, Hongyi Xu, Jing Yang, Shouhong Ding, and Lizhuang Ma. Feature generation and hypothesis verification for reliable face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1782–1791, 2022. 2

[8] Yuchen Liu, Yabo Chen, Mengran Gou, Chun-Ting Huang, Yaoming Wang, Wenrui Dai, and Hongkai Xiong. Towards unsupervised domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20654–20664, 2023. 2

[9] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19685–19696, 2023. 2, 4

[10] Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, and Wen-Sheng Chu. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24563–24574, 2023. 2

[11] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20281–20290, 2022. 2

[12] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4123–4133, 2022. 2

[13] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 4

[14] Guosheng Zhang, Keyao Wang, Haixiao Yue, Ajian Liu, Gang Zhang, Kun Yao, Errui Ding, and Jingdong Wang. Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models. *arXiv preprint arXiv:2501.01720*, 2025. 2, 4

[15] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 70–85. Springer, 2020. 1

[16] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012. 4

[17] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 4

[18] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Kekai Sheng, Shouhong Ding, and Lizhuang Ma. Generative domain adaptation for face anti-spoofing. In *European conference on computer vision*, pages 335–356. Springer, 2022. 2

[19] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma. Instance-aware domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20453–20463, 2023. 2

[20] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, and Lizhuang Ma. Test-time domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–187, 2024. 2