PA-FAS: Towards Interpretable and Generalizable Multimodal Face Anti-Spoofing via Path-Augmented Reinforcement Learning

Yingjie Ma[*1,2], **Xun Lin**[*2], **Yong Xu**[5], **Weicheng Xie**[1,3], **Zitong Yu**[2,3,4†]

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]School of Computing and Information Technology, Great Bay University
[3]Guangdong Provincial Key Laboratory of Intelligent Information Processing & Shenzhen Key Laboratory of Media Security, Shenzhen University
[4]Dongguan Key Laboratory for Intelligence and Information Technology
[5]Harbin Institute of Technology, Shenzhen

Abstract

In recent years, face anti-spoofing (FAS) has made notable progress in multimodal fusion, cross-domain generalization, and interpretability. With the development of large language models and reinforcement learning (RL), strategy-based training paradigms offer new opportunities for jointly modeling multimodality, generalization, and interpretability. However, compared to unimodal reasoning, multimodal reasoning introduces more complex logic, such as accurate feature representation and cross-modal verification, which significantly increases reasoning complexity and labeling difficulty. Due to the lack of high-quality annotations in existing multimodal FAS datasets, directly applying RL strategies is sub-optimal, hindering robust multimodal reasoning. In this paper, we find two key issues of supervised fine-tuning combined with reinforcement learning (SFT+RL) paradigms in multimodal FAS reasoning: 1) limited multimodal reasoning paths not only hinder the full utilization of multimodal information but also constrain the model's exploration space after SFT, thereby affecting the effectiveness of subsequent RL; and 2) the mismatch between single-task supervision and the diversity of multimodal reasoning paths leads to reasoning confusion, where models may exploit shortcuts by directly mapping input images to answers, bypassing the intended reasoning process. These issues further increase the complexity of multimodal reasoning and hinder the effective application of RL strategies. To address these challenges, we propose the PA-FAS framework with a reasoning path enhancement strategy for high-quality extended reasoning sequences construction based on limited annotated data to enrich the reasoning paths and alleviate exploration constraints. Additionally, we introduce an answer shuffling mechanism during SFT for comprehensive multimodal analysis rather than mining superficial cues, thus encouraging deeper reasoning and avoiding shortcut learning. Our method significantly improves multimodal reasoning accuracy and generalization, and successfully unifies multimodal fusion, cross-domain generalization, and interpretability towards trustworthy multimodal FAS.

Code — https://github.com/murInJ/PA-FAS

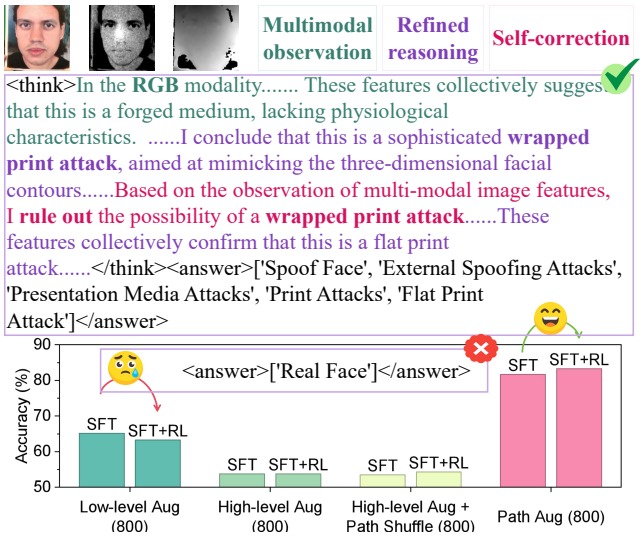

Figure 1: Accuracy of SFT and SFT+RL methods on different augmented datasets. With a fixed data size of 800, datasets with a single reasoning path fail to achieve higher accuracy in the subsequent RL stage after SFT training, and may even experience a decline in performance. In contrast, datasets with diverse reasoning paths demonstrate significantly better performance, achieving higher accuracy under both SFT and SFT+RL methods.

Introduction

Face recognition (FR) systems have been widely adopted in scenarios such as payment authentication, identity verification, and surveillance. However, due to their heavy reliance on visual information, they are highly vulnerable to Presentation Attacks (PAs), including printed photos, replayed videos, and 3D masks, posing significant security risks. To enhance system robustness, Face Anti-Spoofing (FAS) techniques have emerged, aiming to distinguish between genuine and spoofed facial presentations. Traditional FAS methods (Yu et al. 2022; Jiang et al. 2023; Yue et al. 2023; Liu et al. 2023b; Cai et al. 2023, 2024; Liu 2024; Liao et al. 2023; Zhou et al. 2023; Liu et al. 2024), primar-

[*]These authors contributed equally.
[†]Corresponding author

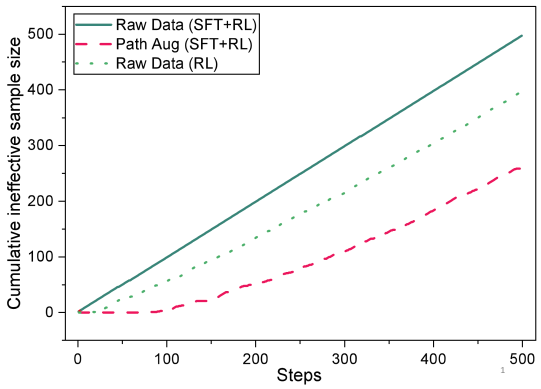

Figure 2: Cumulative effective sample size versus training steps in the RL and SFT+RL stages for models trained with 800 data.

ily based on RGB modality, struggle to cope with increasingly diverse and sophisticated attack modalities due to limited information. Recently, multimodal FAS approaches (Yu et al. 2023) incorporating DEPTH and infrared modalities alongside RGB have demonstrated significant improvements in both accuracy and robustness. However, these advances bring new challenges in integrating and interpreting heterogeneous modalities in real-world FAS applications where generalization ability and interpretability are crucial.

FAS research faces research gaps: 1) insufficient domain generalization (DG); 2) lack of interpretability in multimodal approaches; and 3) limited extensibility of Multimodal Large Language Models (MLLMs) to domain generalization scenarios. While existing methods have made progress in improving model robustness across domains and attack types (Zhou et al. 2023; Sun et al. 2023; Srivatsan, Naseer, and Nandakumar 2023; Liu et al. 2024; Fang et al. 2024), they are mostly developed for unimodal settings and offer limited insight into the model's decision process. On the other hand, although recent multimodal methods (Liu et al. 2023a; Cai et al. 2025; Xie et al. 2024; Yu et al. 2023; Han et al. 2023; Kong et al. 2024; Yu et al. 2024b,a; Li et al. 2023, 2021) show superior performance, they lack explicit interpretability mechanisms for identifying spoofing cues in DEPTH and infrared modalities. Meanwhile, recent MLLM-based FAS approaches (Zhang et al. 2025a; Shi et al. 2025; Zhang et al. 2025b) have demonstrated strong language-level reasoning capabilities in unimodal scenarios. However, they neglect the generalization issue and fail to address cross-modal cue integration and reasoning for real-world multimodal spoofing detection. Although some recent works have attempted to unify domain generalization and multimodal FAS (Lin et al. 2024, 2025a; Yang et al. 2025; Ma et al. 2025b; Lin et al. 2025b), and others have explored the integration of interpretability and domain generalization in FAS tasks (Zhang et al. 2025a), as well as leveraging multimodal datasets in MLLMs (Shi et al. 2025) to bridge multimodality and interpretability. However, a comprehensive solution that simultaneously addresses domain generalization, multimodal fusion, and interpretability remains unexplored.

Recent studies (Huang et al. 2025; Ma et al. 2025a; Zhao, Wei, and Bo 2025; Liu et al. 2025; Pan et al. 2025; Achiam et al. 2023; Bai et al. 2025; Team et al. 2023) have explored enhancing LLM reasoning capabilities via Chain-of-Thought (CoT) and Reinforcement Learning (RL) paradigms to improve logical reasoning and domain adaptability. Notably, the Group Relative Policy Optimization (GRPO) algorithm (Shao et al. 2024) introduces a rule-based group advantage strategy that avoids the need for expensive neural reward models, achieving impressive generalization with low training costs. This offers a new pathway toward constructing FAS systems with interpretable DG capabilities. Furthermore, the integration of Supervised Fine-Tuning (SFT) memory mechanisms with RL-based generalization strategies (Chu et al. 2025) provides theoretical grounding for staged training: building stable knowledge during the SFT phase, followed by policy exploration and self-improvement in the RL phase. However, unlike unimodal settings, multimodal FAS requires accurate representation of RGB, DEPTH, and infrared modalities, as well as complex reasoning logic such as modality corroboration, conflict resolution, and modality assistance, significantly increasing the difficulty of learning. This intrinsic complexity makes it costly and difficult to collect high-quality, fine-grained annotations. As a result, existing datasets typically contain only simple binary labels with limited modality coverage, lacking supervision for cross-modal relationships. As shown in Fig. 1, such weak supervision often causes models to overfit rigid patterns during SFT, while the RL phase suffers from insufficient feedback and exploration space, ultimately limiting the generalization and interpretability of SFT+RL, and even leading to worse performance than using SFT alone.

Our analysis identifies two key issues in applying the SFT+RL paradigm to current multimodal FAS settings: 1) The SFT phase lacks multimodal reasoning diversity, involving simple tasks and limited-scale data, weakening the full utilization of multimodal information and severely narrowing the RL exploration space; and 2) Even with conventional data augmentation, models often exploit shortcuts by relying solely on image inputs, ignoring intermediate reasoning processes and leading to fragile decision strategies with poor generalization and interpretability. To address these issues, we propose a Reasoning Path Augmentation (PA) strategy, which explicitly expands the original reasoning space at minimal cost with positive–negative random path sampling method, enabling the full utilization of multimodal information for reasoning, as well as promoting greater exploration and policy diversity during the RL stage under limited label settings. Building upon high-level augmentation methods, PA introduces diverse multimodal reasoning chains associated with each input, thereby enhancing both generalization and interpretability. Our main contributions include:

- We provide an in-depth analysis of the failure mechanisms of existing multimodal FAS datasets under the SFT+RL paradigm, offering both valuable insights and empirical evidence for designing explainable and generalizable training frameworks.
- We propose the PA-FAS framework with a novel Rea-

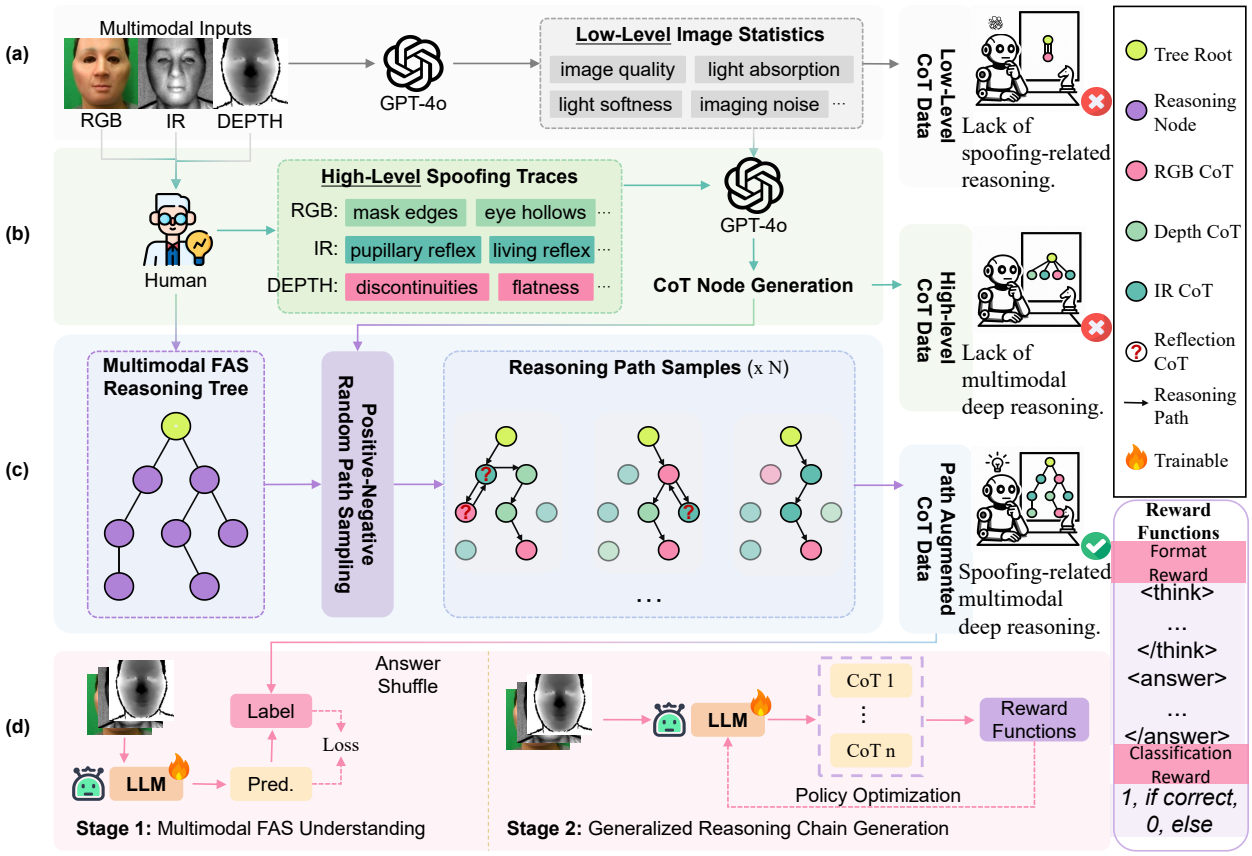

Figure 3: Schematic diagram of the PA-FAS framework. Raw data undergo (a) low-level and (b) high-level data annotation to obtain corresponding CoT. Subsequently, (c) Positive–Negative Random Path Sampling is employed to sample a specified number of reasoning paths from a human-constructed multimodal reasoning tree and integrate them into CoT. During the (d) SFT+RL training paradigm, answers are randomly shuffled in the SFT stage to prevent the policy model from forming shortcuts, thereby learning diverse reasoning paths and rich multimodal domain-specific knowledge. In the RL stage, the policy model achieves generalization through classification and format rewards.

soning Path Augmentation strategy that fully utilization of multimodal information and effectively expands the reasoning space during the RL phase with limited supervision. Additionally, we introduce an answer shuffling mechanism during the SFT phase to focus the model on learning reasoning paths and avoid shortcuts, thereby enhancing the model's generalization and interpretability.

- To the best of our knowledge, this is the first work that systematically integrates multimodal feature fusion, domain generalization, and reasoning interpretability within a unified FAS framework, paving the way for robust and trustworthy multimodal FAS systems.

Method

Analysis of Failure in Multimodal FAS Datasets

Previous studies demonstrate that combining SFT with RL can significantly improve model performance across various tasks (Chu et al. 2025), as the SFT stage allows the model to internalize domain-specific knowledge, reasoning patterns, and chain-of-thought structures, laying the groundwork for effective policy optimization during RL. However, when this paradigm is applied to multimodal FAS tasks, where datasets with only binary labels, lacking linguistic annotations of key visual cues, and exhibiting high task uniformity, the model often develops overconfident predictions during the SFT phase. Such overconfidence leads to extreme reward feedback during RL, where most samples receive either full (1) or zero rewards, and informative intermediate signals are largely absent. As shown in Fig. 2, the model fine-tuned on raw data accumulates ineffective samples at a nearly linear rate throughout RL, indicating a lack of exploration and highly polarized learning signals. In contrast, the model trained directly with RL, although not fine-tuned, demonstrates a more moderate cumulative trajectory, suggesting greater exploratory behavior. These findings highlight the limitations of the SFT+RL paradigm under current dataset settings and motivate the need for improved training strategies in the fine-tuning stage.

To mitigate this issue, we first adopt a cold-start data aug-

Method	CPS→W		CPW→S		CSW→P		PSW→C	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
SSDG [CVPR'20]	26.09	82.03	28.50	75.91	41.82	60.56	40.48	62.31
SSAN [CVPR'22]	17.73	91.69	27.94	79.04	34.49	68.85	36.43	69.29
SA-FAS [CVPR'23]	21.37	87.65	23.22	84.49	35.10	70.86	35.38	69.71
IADG [CVPR'23]	27.02	86.50	23.04	83.11	32.06	73.83	39.24	63.68
FLIP [IJCAI'22]	13.19	93.79	11.73	94.93	17.39	90.63	22.14	83.95
ViT [ICLR'20]	20.88	84.77	44.05	57.94	33.58	71.80	42.15	56.45
AMA [IJCV'24]	17.56	88.74	27.50	80.00	21.18	85.51	47.48	55.56
VP-FAS [TDSC'24]	16.26	91.22	24.42	81.07	21.76	85.46	39.35	66.55
ViTAF [ECCV'22]	20.58	85.82	29.16	77.80	30.75	73.03	39.75	63.44
MM-CDCN [CVPR'20]	38.92	65.39	42.93	59.79	41.38	61.51	48.14	53.71
CMFL [CVPR'21]	18.22	88.82	31.20	75.66	26.68	80.85	36.93	66.82
CLIP [ICML'21]	14.55	90.47	18.17	90.02	24.13	83.15	38.33	65.71
MMDG [CVPR'24]	12.79	93.83	15.32	92.86	18.95	88.64	29.93	76.52
DADM [ICCV'25]	11.71	94.89	**6.92**	**97.66**	19.03	88.22	**16.87**	**91.08**
Qwen2.5-VL-3B [arXiv'25]	30.86	75.01	49.56	44.35	19.72	88.10	33.72	70.01
Qwen2.5-VL-3B-SFT	5.12	97.65	44.84	52.71	15.15	90.73	27.91	76.20
Qwen2.5-VL-3B-SFT+GRPO	7.75	97.36	57.06	44.79	45.26	56.29	27.41	76.29
PA-FAS (Ours)	**2.39**	**99.73**	27.75	78.07	**9.48**	**94.50**	21.23	84.25

Table 1: Cross-dataset testing results under the fixed-modal scenarios (Protocol 1) among CASIA-CeFA (C), PADISI (P), CASIA-SURF (S), and WMCA (W). Best and second-best results are marked in **bold** and underline, respectively.

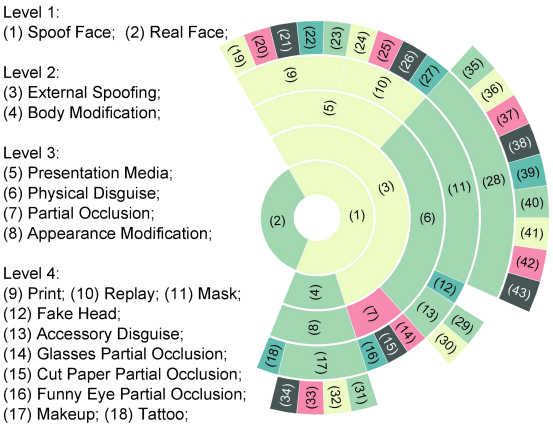

Figure 4: Sunburst diagram of the fine-grained hierarchical taxonomy for FAS. Every category is directly mapped to a node in the reasoning tree.

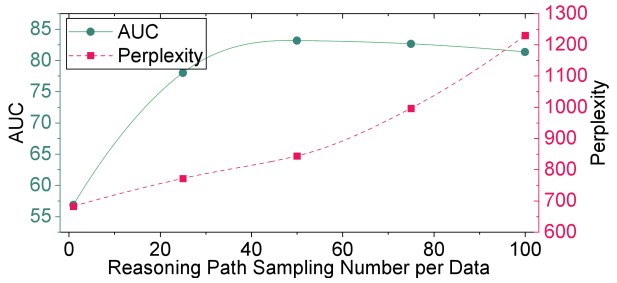

Figure 5: A diagram of AUC and Perplexity vs. reasoning path sampling number N. As sampling paths increase, perplexity rises sharply, causing AUC to peak and then decline slowly.

mentation strategy. Considering the high cost of manually annotating key visual clues in multimodal images, we pursue a more practical approach: as shown in Fig. 3 (a) and (b), for each of the few annotated samples, we generate multiple distinct versions of low-level and high-level CoT reasoning chains, aiming to introduce diversity in reasoning paths and enhance the model's generalization capability. However, as illustrated in Fig. 1 High-level Aug (800), this augmentation fails to alleviate the problem of extreme reward distribution. To further investigate, we conduct a reasoning content replacement experiment, wherein the reasoning text enclosed by $< think >$ tags during SFT is randomly substituted with CoT sequences from other samples. As shown in Fig. 1 'High-level Aug + Path Shuffle (800)', the performance of the model remains nearly unchanged regardless of whether the reasoning content is replaced, indicating that reasoning heavily relies on the visual predictions but neglects the CoT, a phenomenon we refer to as 'reasoning shortcut'. Therefore, to enable effective training under the SFT+RL paradigm for multimodal FAS tasks, we identify two key requirements: 1) the SFT stage must provide structurally diverse and semantically rich reasoning paths to construct a meaningful exploration space for RL; and 2) shortcut learning based on direct image-to-answer mapping should be avoided to ensure the model learns to reason explicitly through the CoT process. To satisfy the first condition, we propose the PA-FAS framework with a Reasoning Path Augmentation (PA) strategy. Unlike high-level data augmentation, PA focuses on diversifying the structure of reasoning

Method	Missing D		Missing I		Missing D & I	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
SSDG [CVPR'20]	38.92	65.45	37.64	66.57	39.18	65.22
SSAN [CVPR'22]	36.77	69.21	41.20	61.92	33.52	73.38
SA-FAS [CVPR'23]	36.30	69.07	39.80	62.69	33.08	74.29
IADG [CVPR'23]	40.72	58.72	42.17	61.83	37.50	66.90
FLIP [IJCAI'22]	23.66	83.90	24.06	84.04	27.07	79.79
ViT [ICLR'20]	40.04	64.69	36.77	68.19	36.20	69.02
AMA [IJCV'24]	29.25	77.70	32.30	74.06	31.48	75.82
VP-FAS [TDSC'24]	29.13	78.27	29.63	77.51	30.47	76.31
ViTAF [ECCV'22]	34.99	73.22	35.88	69.40	35.89	69.61
MM-CDCN [CVPR'20]	44.90	55.35	43.60	58.38	44.54	55.08
CMFL [CVPR'21]	31.37	74.62	30.55	75.42	31.89	74.29
CLIP [ICML'21]	28.07	77.00	29.10	77.04	32.58	73.36
MMDG [CVPR'24]	24.89	82.39	23.39	83.82	25.26	81.86
DADM [ICCV'25]	<u>21.56</u>	<u>85.17</u>	<u>20.82</u>	<u>85.28</u>	<u>22.61</u>	<u>84.04</u>
Qwen2.5-VL-3B [arXiv'25]	33.46	69.36	33.46	69.36	33.46	69.36
Qwen2.5-VL-3B-SFT	23.25	79.32	23.25	79.32	23.25	79.32
Qwen2.5-VL-3B-SFT+GRPO	34.37	68.68	34.37	68.68	34.37	68.68
PA-FAS (Ours)	**15.68**	**89.07**	**17.32**	**88.23**	**14.67**	**89.73**

Table 2: Cross-dataset testing results under the missing modalities scenarios (Protocol 2) among CASIA-CeFA (C), PADISI (P), CASIA-SURF (S), and WMCA (W). Best and second-best results are marked in **bold** and <u>underline</u>, respectively.

chains. By constructing multiple CoT paths for each sample that are semantically consistent but logically varied, as shown in Fig. 2, PA significantly expands the model's exploration space during RL. This enables effective reasoning path generalization with only limited annotated data. To address the second issue, we introduce an answer-shuffling mechanism during SFT, compelling the model to master diverse reasoning paths and every possible answer, which effectively blocks reasoning shortcuts while reserving ample exploration space for the RL phase.

Reasoning Path Augmentation

As shown in Fig. 3, we propose a novel PA strategy that constructs a structured and diverse set of reasoning paths to expand the output space during reinforcement learning, thereby significantly enhancing both exploration efficiency and training effectiveness. As illustrated in Fig. 1, the effectiveness of High-level augmentation shows a significant decline compared to Low-level augmentation, because under limited data availability, data augmentation by existing MLLM can introduce some degree of semantic diversity but remains extremely sparse in expanding valid reasoning trajectories. Moreover, in multimodal FAS tasks, token-level operations are often prohibitively costly and fail to effectively filter out semantically erroneous or logically inconsistent pseudo-augmented texts, causing the augmented data to suffer from substantial noise and errors that markedly degrade data quality and thus hinder large-scale deployment.

To address these challenges, we propose a reasoning-path-based data augmentation method. Leveraging the fine-grained label hierarchy illustrated in Fig. 4 for multimodal FAS tasks, we build a formal reasoning tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of reasoning nodes and $\mathcal{E}$ the set of directed edges. Each node $v \in \mathcal{V}$ represents a semantic class or logical decision unit, and every path $\mathcal{P} = (v_1, v_2, \ldots, v_n)$ encodes a complete reasoning chain from the root $v_1 = r$ to the target leaf node v_n.

Built upon this structure, we propose a Positive–Negative Random Path Sampling (PNRPS) strategy to systematically sample a diverse and logically rich set of reasoning paths $\mathcal{P}_i = \{\mathcal{P}_i^{(1)}, \ldots, \mathcal{P}_i^{(N)}\}$ for each original data x_i, thereby enhancing its reasoning ability. Assume a dataset consisting of M instances $\{(x_i, \ell_i)\}_{i=1}^{M}$, where each x_i is a data sample and ℓ_i is its corresponding ground-truth label. The core mechanisms are as follows: (1) Single-node operation constraint: Each node $v \in \mathcal{V}$ is allowed at most one forward exploration step $(+, v)$ and one backward reflection step $(-, v)$ to avoid redundant walks; (2) Path length constraint: Given a maximum taxonomy depth D, we set an upper bound $L_{\max} = \alpha(D-1)$ on the reasoning path length to control complexity, where $\alpha > 1$ is a tunable scaling factor that determines the maximum allowable reasoning steps relative to the taxonomy depth; (3) Semantic consistency enforcement: Each node is associated with a predefined Chain-of-Thought (CoT) clause template, and the final reasoning text is constructed by sequentially composing these templates along the path; (4) Structural sampling strategy: We perform rule-guided depth-first traversal from the root, randomly sampling N valid paths that utilize information from any of the RGB, IR, and DEPTH modalities and mapping them into logically coherent CoT descriptions. The PA process can be formalized as a mapping:

$$\{(x_i, \ell_i)\}_{i=1}^{M} \rightarrow \bigcup_{i=1}^{M} \left\{ \left(x_i,\, \mathrm{CoT}(\mathcal{P}_i^{(j)}) \right) \,\middle|\, j = 1, \ldots, N \right\} \quad (1)$$

where $\mathrm{CoT}(\mathcal{P})$ represents the reasoning description composed by chaining CoT sub-clauses corresponding to path $\mathcal{P}$.

Method	CW→PS		PS→CW		Average	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
SSDG [CVPR'20]	25.34	80.17	46.98	54.29	35.66	67.23
SSAN [CVPR'22]	26.55	80.06	39.10	67.19	32.82	73.62
SA-FAS [CVPR'23]	25.20	81.06	36.59	70.03	61.79	30.89
IADG [CVPR'23]	22.82	83.85	39.70	63.46	31.26	73.65
FLIP [IJCAI'22]	15.92	92.38	<u>23.85</u>	<u>83.46</u>	19.88	87.92
ViT [ICLR'20]	42.66	57.80	42.75	60.41	42.70	59.10
AMA [IJCV'24]	29.25	76.89	38.06	67.64	33.65	72.26
VP-FAS [TDSC'24]	25.90	81.79	44.37	60.83	35.13	71.31
ViTAF [ECCV'22]	29.64	77.36	39.93	61.31	34.78	69.33
MM-CDCN [CVPR'20]	29.28	76.88	47.00	51.94	38.14	64.41
CMFL [CVPR'21]	31.86	72.75	39.43	63.17	35.64	67.96
CLIP [ICML'21]	19.36	90.57	29.98	79.22	24.67	84.89
MMDG [CVPR'24]	20.12	88.24	36.60	70.35	28.36	79.30
DADM [ICCV'25]	12.61	93.81	**20.40**	**89.51**	<u>16.50</u>	**91.66**
Qwen2.5-VL-3B [arXiv'25]	34.26	68.92	45.52	51.34	39.80	60.13
Qwen2.5-VL-3B-SFT	<u>0.60</u>	<u>99.94</u>	33.76	69.94	17.18	84.94
Qwen2.5-VL-3B-SFT+GRPO	**0.60**	**99.94**	33.78	69.71	17.19	84.82
PA-FAS (Ours)	2.53	99.54	28.75	77.13	**15.64**	<u>88.33</u>

Table 3: Cross-dataset testing under limited source domain scenarios (Protocol 3) among CASIA-CeFA (C), PADISI USC (P), CASIA-SURF (S), and WMCA (W).

Data	HTER(%)↓	AUC(%)↑
Low-level Augmentation Data	34.37	68.68
High-level Augmentation Data	44.49	52.95
Reasoning Path Augmentation Data	**24.45**	**83.17**

Table 4: Ablation on augmentation under the SFT+RL paradigm.

Method	HTER(%)↓	AUC(%)↑
w/o Shuffle	24.45	83.17
w/ Shuffle Path	24.12	84.23
w/ Shuffle Answer	**15.21**	**89.13**

Table 5: Ablation on Path-Augmented data under shuffling.

In our implementation, we start with only 800 labeled instances and generate $N = 50$ reasoning paths per data, yielding a total of approximately 4×10^4 structurally diverse and semantically coherent augmented samples. This strategy significantly improves the coverage diversity and structural controllability of the training data, providing a robust foundation for reinforcement learning with improved generalization and reasoning capability.

Answer Shuffling for SFT+RL Paradigm

Although reasoning-path augmentation enlarges the training set, the coupling of a single task with overly rich multimodal reasoning paths can confuse the model during SFT: the final answer is often produced by directly looking at the image, bypassing the reasoning chain and creating a shortcut. This shortcut makes the model overconfident and drastically shrinks the exploration space in the RL phase. To sever such shortcuts, we introduce answer shuffling in Fig. 3(d). During SFT, the final answer in each chain-of-thought is randomly swapped with the answer from another sample, forcing the model to focus on learning the diverse reasoning paths instead of memorizing the answer and preserving room to explore every possible answer.

After the answer-shuffled SFT phase, we move to the RL stage shown in Fig. 3(d) and adopt Group Relative Policy Optimization (GRPO). For every question–answer pair (q, a) the old policy $\pi_{\theta_{\text{old}}}$ samples a group of G responses $\{o_i\}_{i=1}^{G}$. The reward for each response is defined as

$$\mathcal{R} = \mathcal{R}_{\text{format}} + \mathcal{R}_{\text{classification}}, \tag{2}$$

where $\mathcal{R}_{\text{classification}} = 1$ if the predicted class matches the ground-truth label and 0 otherwise. Given the corresponding rewards $\{\mathcal{R}_i\}_{i=1}^{G}$, GRPO computes the relative advantage

$$\hat{A}_{i,t} = \frac{\mathcal{R}_i - \text{mean}(\{\mathcal{R}_i\})}{\text{std}(\{\mathcal{R}_i\})}. \tag{3}$$

The optimization objective is

$$\mathcal{J}_{\text{GRPO}}(\theta) = E_{(q,a)\sim\mathcal{D},\{o_i^G\}\sim\pi_{\theta_{\text{old}}}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_{i,t}\Big) \Bigg], \tag{4}$$

with

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{5}$$

While the original GRPO adds a KL term $D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ to curb large policy updates, we drop it under data-scarce conditions to avoid suppressing exploration.

Experiments

We conduct our evaluation and training using four widely adopted multimodal FAS datasets: WMCA (George et al. 2019), SURF (Zhang et al. 2020), CeFA (Liu et al. 2021), and PADISI (Rostami et al. 2021). To assess the domain generalization capability of our model under multimodal settings, we follow the cross-domain evaluation protocol proposed in MMDG (Lin et al. 2024), which includes several sub-protocols covering scenarios such as fixed modality, missing modality, and limited source domains. Detailed configurations are provided in supplementary material.

Implementation Details

We adopt Qwen2.5VL (Bai et al. 2025) as our base multimodal large language model. All supervised fine-tuning and reinforcement learning stages are trained for a fixed 500 steps with a constant learning rate of 1e-6. For fair comparison, we categorize competing approaches into four groups: (1) Uni-modal DG methods that extend the input to multi-modal, (2) Multi-modal FAS methods, (3) Multi-modal DG FAS methods, and our proposed (4) interpretable Multi-modal DG FAS methods. This classification enables a more structured and meaningful evaluation for the FAS task.

Cross-Dataset Testing

Complete Modality Scenario. Protocol 1 is designed to evaluate model performance across unseen domains using multimodal data from varied scenarios. For example, the sub-protocol **CPS** $\rightarrow$ **W** represents that we take **C**, **P**, and **S** as training sets, while **W** is testing set. As shown in Table 1, on the interpretable domain generalization multimodal FAS task, the zero-shot baseline Qwen2.5VL-3B yields an average HTER of 33.46%. After SFT, this figure falls to 23.25%, yet it rebounds sharply to 34.37% once the RL stage is added, revealing the risk of SFT+RL collapse under data-scarce conditions. By contrast, our PA-FAS drives the average HTER down to 15.21%, setting a new best-in-class record. This demonstrates that with only $\approx$800 high-quality structured reasoning paths, PA-FAS surpasses the generalization performance achieved by $\approx$35,000 raw data lacking reliable annotations, effectively mitigating the dual challenges of data scarcity and label noise.

Missing Modality Scenario During Testing. In Protocol 2, for each LOO sub-protocol of Protocol 1, we design three test-time missing-modal scenarios to validate the model's performance when modalities are missing. Table 2 shows that under various modality-missing conditions the performance curves of Qwen2.5VL-3B in its zero-shot and raw-data-trained states almost overlap, revealing that without high-quality chain-of-thought annotations the model relies solely on RGB images and fails to leverage multi-modal cues. Our proposed approach alleviates this limitation, driving the average HTER down from 33.46% to 15.85%. Interestingly, retaining only the RGB modality even yields a slight improvement over cases where DEPTH or infrared is individually missing, suggesting that the other modalities currently introduce a mild interference to the RGB signal.

Limited Source Domain Scenario. In Protocol 3, we limit the number of source domains by proposing two sub-protocols, namely $CW \rightarrow PS$ and $PS \rightarrow CW$. As shown in Table 3, our method slashes HTER by 31.73% and 16.77% compared with the zero-shot baseline, demonstrating robust and superior performance even when source-domain data are severely limited. Especially under scenarios of limited labeled data, our path augmentation method still achieves performance comparable to that obtained with large amounts of unlabeled data, even in the context of restricted source domains, indicating our advantage of data efficiency.

Ablation Study

Effectiveness of Reasoning Path Augmentation. As shown in Table 4, training on approximately 800 reasoning-path-augmented data under the SFT+RL paradigm reduces HTER by 9.92% compared with using roughly 35,000 raw data, demonstrating a substantial edge in both data efficiency and performance. A further comparison with token-level augmentation confirms that this gain stems from the expansion of valid reasoning paths rather than superficial diversity.

Effectiveness of Answer Shuffling Mechanism. As shown in Table 5, although reasoning-path augmentation expands the pool of valid data, it still falls victim to reasoning shortcuts: path-level shuffling even yields a slightly higher HTER than the unshuffled baseline, highlighting the persistence of such shortcuts. In contrast, introducing answer shuffling decreases HTER by 9.24%, conclusively demonstrating its ability to sever these shortcuts and markedly improve model learning.

Impact of Sampling Numbers per Data. We conduct studies on reasoning path enhancement under different sampling number N, as shown in Fig. 5. The results indicate that as the sampling quantity increases, the model's perplexity rises sharply, suggesting that an excessive number of reasoning paths hinders model convergence. When $N \approx 50$, the model achieves the best performance. However, when it exceeds 50, the increase in perplexity leads to a gradual decline in AUC, indicating that an appropriate sampling quantity enables the model to fully utilize reasoning path information to achieve higher accuracy, while excessive samples result in redundant path information, causing model confusion and limited performance.

Conclusion

In this paper, we introduce a reasoning-path augmentation strategy together with an answer-shuffling mechanism, enabling the SFT+RL paradigm to be effectively applied to multimodal FAS under scarce annotations. It provides initial evidence that interpretability, multimodal fusion, and cross-domain generalization can be jointly modeled in a unified and trustworthy FAS training framework. However, the limited labeled data renders the training process less stable and efficient, and the utilization of multimodal information remains sub-optimal. Future works focus on refining the reward function design, optimizing RL strategies, and exploring low-cost, high-efficiency data-annotation schemes to further enhance performance and practical deployability.

Acknowledgments

This work was supported by CCF-Tencent Rhino-Bird Open Research Fund, National Natural Science Foundation of China (Grant No. 62576076), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037), Guangdong Provincial Key Laboratory (Grant 2023B1212060076), Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047). The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cai, R.; Cui, Y.; Li, Z.; Yu, Z.; Li, H.; Hu, Y.; and Kot, A. 2023. Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. In *ICCV*, 8037–8048.

Cai, R.; Cui, Y.; Yu, Z.; Lin, X.; Chen, C.; and Kot, A. 2025. Rehearsal-free and efficient continual learning for cross-domain face anti-spoofing. *IEEE Trans. Pattern Anal. Mach. Intell.*

Cai, R.; Yu, Z.; Kong, C.; Li, H.; Chen, C.; Hu, Y.; and Kot, A. C. 2024. S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens. *IEEE Trans. Inf. Forensics Secur.*

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Fang, H.; Liu, A.; Jiang, N.; Lu, Q.; Zhao, G.; and Wan, J. 2024. Vl-fas: Domain generalization via vision-language model for face anti-spoofing. In *ICASSP*, 4770–4774. IEEE.

George, A.; Mostaani, Z.; Geissenbuhler, D.; Nikisins, O.; Anjos, A.; and Marcel, S. 2019. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Trans. Inf. Forensics Secur.*, 15: 42–55.

Han, S.; Cai, R.; Cui, Y.; Yu, Z.; Hu, Y.; and Kot, A. 2023. Hyperbolic face anti-spoofing. *arXiv preprint arXiv:2308.09107*.

Huang, J.; Xu, Z.; Zhou, J.; Liu, T.; Xiao, Y.; Ou, M.; Ji, B.; Li, X.; and Yuan, K. 2025. SAM-R1: Leveraging SAM for Reward Feedback in Multimodal Segmentation via Reinforcement Learning. *arXiv preprint arXiv:2505.22596*.

Jiang, F.; Li, Q.; Liu, P.; Zhou, X.-D.; and Sun, Z. 2023. Adversarial learning domain-invariant conditional features for robust face anti-spoofing. *Int. J. Comput. Vis.*, 131(7): 1680–1703.

Kong, C.; Zheng, K.; Liu, Y.; Wang, S.; Rocha, A.; and Li, H. 2024. M3FAS: An Accurate and Robust MultiModal Mobile Face Anti-Spoofing System. *IEEE Trans. Dependable Secure Comput.*

Li, K.; Yang, H.; Chen, B.; Li, P.; Wang, B.; and Huang, D. 2023. Learning polysemantic spoof trace: A multi-modal disentanglement network for face anti-spoofing. In *AAAI*, volume 37, 1351–1359.

Li, Z.; Li, H.; Luo, X.; Hu, Y.; Lam, K.-Y.; and Kot, A. C. 2021. Asymmetric modality translation for face presentation attack detection. *IEEE Trans. Multimedia*, 25: 62–76.

Liao, C.-H.; Chen, W.-C.; Liu, H.-T.; Yeh, Y.-R.; Hu, M.-C.; and Chen, C.-S. 2023. Domain invariant vision transformer learning for face anti-spoofing. In *WACV*, 6098–6107.

Lin, X.; Liu, A.; Yu, Z.; Cai, R.; Wang, S.; Yu, Y.; Wan, J.; Lei, Z.; Cao, X.; and Kot, A. 2025a. Reliable and Balanced Transfer Learning for Generalized Multimodal Face Anti-Spoofing. *IEEE Trans. Pattern Anal. Mach. Intell.*

Lin, X.; Liu, A.; Yu, Z.; Cai, R.; Wang, S.; Yu, Y.; Wan, J.; Lei, Z.; Cao, X.; and Kot, A. 2025b. Reliable and Balanced Transfer Learning for Generalized Multimodal Face Anti-Spoofing. *IEEE Trans. Pattern Anal. Mach. Intell.*

Lin, X.; Wang, S.; Cai, R.; Liu, Y.; Fu, Y.; Tang, W.; Yu, Z.; and Kot, A. 2024. Suppress and rebalance: Towards generalized multi-modal face anti-spoofing. In *CVPR*, 211–221.

Liu, A. 2024. Ca-moeit: Generalizable face anti-spoofing via dual cross-attention and semi-fixed mixture-of-expert. *Int. J. Comput. Vis.*, 132(11): 5439–5452.

Liu, A.; Tan, Z.; Wan, J.; Escalera, S.; Guo, G.; and Li, S. Z. 2021. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *WACV*, 1179–1187.

Liu, A.; Tan, Z.; Yu, Z.; Zhao, C.; Wan, J.; Liang, Y.; Lei, Z.; Zhang, D.; Li, S. Z.; and Guo, G. 2023a. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.*, 18: 4775–4786.

Liu, A.; Xue, S.; Gan, J.; Wan, J.; Liang, Y.; Deng, J.; Escalera, S.; and Lei, Z. 2024. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *CVPR*, 222–232.

Liu, Y.; Chen, Y.; Gou, M.; Huang, C.-T.; Wang, Y.; Dai, W.; and Xiong, H. 2023b. Towards unsupervised domain generalization for face anti-spoofing. In *ICCV*, 20654–20664.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Ma, X.; Ding, Z.; Luo, Z.; Chen, C.; Guo, Z.; Wong, D. F.; Feng, X.; and Sun, M. 2025a. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*.

Ma, Y.; Yu, Z.; Lin, X.; Xie, W.; and Shen, L. 2025b. Big-Moe: Bypassing Isolated Gating For Generalized Multimodal Face Anti-Spoofing. In *ICASSP*, 1–5. IEEE.

Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Rostami, M.; Spinoulas, L.; Hussein, M.; Mathai, J.; and Abd-Almageed, W. 2021. Detection and continual learning of novel face presentation attacks. In *ICCV*, 14851–14860.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shi, Y.; Gao, Y.; Lai, Y.; Wang, H.; Feng, J.; He, L.; Wan, J.; Chen, C.; Yu, Z.; and Cao, X. 2025. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *Vis. Intell.*, 3(1): 1–25.

Srivatsan, K.; Naseer, M.; and Nandakumar, K. 2023. Flip: Cross-domain face anti-spoofing with language guidance. In *ICCV*, 19685–19696.

Sun, Y.; Liu, Y.; Liu, X.; Li, Y.; and Chu, W.-S. 2023. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *CVPR*, 24563–24574.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Xie, X.; Cui, Y.; Tan, T.; Zheng, X.; and Yu, Z. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Vis. Intell.*, 2(1): 37.

Yang, J.; Lin, X.; Yu, Z.; Zhang, L.; Liu, X.; Li, H.; Yuan, X.; and Cao, X. 2025. Dadm: Dual alignment of domain and modality for face anti-spoofing. In *ICCV*.

Yu, Z.; Cai, R.; Cui, Y.; Liu, A.; and Chen, C. 2024a. Visual prompt flexible-modal face anti-spoofing. *IEEE Trans. Dependable Secure Comput.*

Yu, Z.; Cai, R.; Cui, Y.; Liu, X.; Hu, Y.; and Kot, A. C. 2024b. Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. *Int. J. Comput. Vis.*, 132(11): 5217–5238.

Yu, Z.; Liu, A.; Zhao, C.; Cheng, K. H.; Cheng, X.; and Zhao, G. 2023. Flexible-modal face anti-spoofing: A benchmark. In *CVPR*, 6346–6351.

Yu, Z.; Qin, Y.; Li, X.; Zhao, C.; Lei, Z.; and Zhao, G. 2022. Deep learning for face anti-spoofing: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5): 5609–5631.

Yue, H.; Wang, K.; Zhang, G.; Feng, H.; Han, J.; Ding, E.; and Wang, J. 2023. Cyclically disentangled feature translation for face anti-spoofing. In *AAAI*, volume 37, 3358–3366.

Zhang, G.; Wang, K.; Yue, H.; Liu, A.; Zhang, G.; Yao, K.; Ding, E.; and Wang, J. 2025a. Interpretable Face Anti-Spoofing: Enhancing Generalization with Multimodal Large Language Models. *arXiv preprint arXiv:2501.01720*.

Zhang, H.; Fang, Z.; Zhao, N.; Hou, S.; Ma, L.; Pei, R.; and He, Z. 2025b. FaceCoT: A Benchmark Dataset for Face Anti-Spoofing with Chain-of-Thought Reasoning. *arXiv preprint arXiv:2506.01783*.

Zhang, S.; Liu, A.; Wan, J.; Liang, Y.; Guo, G.; Escalera, S.; Escalante, H. J.; and Li, S. Z. 2020. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Trans. Biom. Behav. Identity Sci.*, 2(2): 182–193.

Zhao, J.; Wei, X.; and Bo, L. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*.

Zhou, Q.; Zhang, K.-Y.; Yao, T.; Lu, X.; Yi, R.; Ding, S.; and Ma, L. 2023. Instance-aware domain generalization for face anti-spoofing. In *CVPR*, 20453–20463.