FaceShield: Explainable Face Anti-Spoofing with Multimodal Large Language Models

Hongyang Wang[1,2*], Yichen Shi[3,4*], Zhuofu Tao[4,5], Yuhao Gao[1,2], Liepiao Zhang[6], Xun Lin[7]
Jun Feng[1,2], Xiaochen Yuan[10], Zitong Yu[7,8,9†], Xiaochun Cao[11]

[1]Shijiazhuang Tiedao University
[2]Shijiazhuang Key Laboratory of Artificial Intelligence
[3]Shanghai Jiao Tong University
[4]Eastern Institute of Technology
[5]University of California, Los Angeles
[6]GRGBanking
[7]Great Bay University
[8]Guangdong Provincial Key Laboratory of Intelligent Information Processing & Shenzhen Key Laboratory of Media Security, Shenzhen University
[9]Dongguan Key Laboratory for Intelligence and Information Technology
[10]Macao Polytechnic University
[11]Shenzhen Campus of Sun Yat-sen University

Abstract

Face anti-spoofing (FAS) is crucial for protecting facial recognition systems from presentation attacks. Previous methods approached this task as a classification problem, lacking interpretability and reasoning behind the predicted results. Recently, multimodal large language models (MLLMs) have shown strong capabilities in perception, reasoning, and decision-making in visual tasks. However, there is currently no universal and comprehensive MLLM and dataset specifically designed for FAS task. To address this gap, we propose FaceShield, a MLLM for FAS, along with the corresponding pre-training and supervised fine-tuning (SFT) datasets, FaceShield-pre10K and FaceShield-sft45K. FaceShield is capable of determining the authenticity of faces, identifying types of spoofing attacks, providing reasoning for its judgments, and detecting attack areas. Specifically, we employ spoof-aware vision perception (SAVP) that incorporates both the original image and auxiliary information based on prior knowledge. We then use an prompt-guided vision token masking (PVTM) strategy to random mask vision tokens, thereby improving the model's generalization ability. We conducted extensive experiments on three benchmark datasets, demonstrating that FaceShield significantly outperforms previous deep learning models and general MLLMs on four FAS tasks, i.e., coarse-grained classification, fine-grained classification, reasoning, and attack localization.

Code — Codehttps://github.com/Why0912/FaceShield

Introduction

Face anti-spoofing (FAS) is essential in facial recognition systems, ensuring that presentation attacks (PAs), such as print, replay, and 3D wearable masks, are effectively prevented. It has attracted considerable interest in industry and academia in the past decade.

Existing deep learning FAS models can be categorized into two types: vision-based methods and vision-language-based methods. As shown in Fig. 2(a), vision-based methods rely solely on image data (e.g., RGB, Depth, Infrared(IR)) and binary labels to train CNNs (Yu et al. 2020, 2021; Wang et al. 2025) or ViTs (George and Marcel 2021; Cai et al. 2025) for FAS. While they can achieve satisfactory results against known attack types and environments, these methods are prone to overfitting on spurious correlations and lack strong extrapolation capabilities. As illustrated in Fig. 2(b), Vision-language-based methods do not use binary labels but instead train CLIPs with image-text pairs (Srivatsan, Naseer, and Nandakumar 2023; Liu et al. 2024; Mu et al. 2024; Lin et al. 2025). The text labels in these methods provide more domain-agnostic information, enhancing models' generalization capability. Although these FAS models demonstrate some recognition capabilities, they still face challenges such as limited generalization ability, poor interpretability, and a lack of capability for fine-grained localization of attack regions.

Recently, MLLMs have shown remarkable capabilities across various visual tasks (Ye et al. 2025), such as remote sensing (Zhang et al. 2024; Kuckreja et al. 2024; Muhtar et al. 2024), medical imaging (Li et al. 2023; Sun et al. 2024), and deepfake detection (Xu et al. 2024; Huang et al. 2024). By leveraging the general capabilities of language foundation models alongside the visual information extracted by vision towers, these specialized MLLMs integrate perception, reasoning, and decision-making within a single model. Regarding the FAS task, SHIELD (Shi et al. 2025) conducted extensive evaluations on existing general-

*These authors contributed equally.
†Corresponding author.

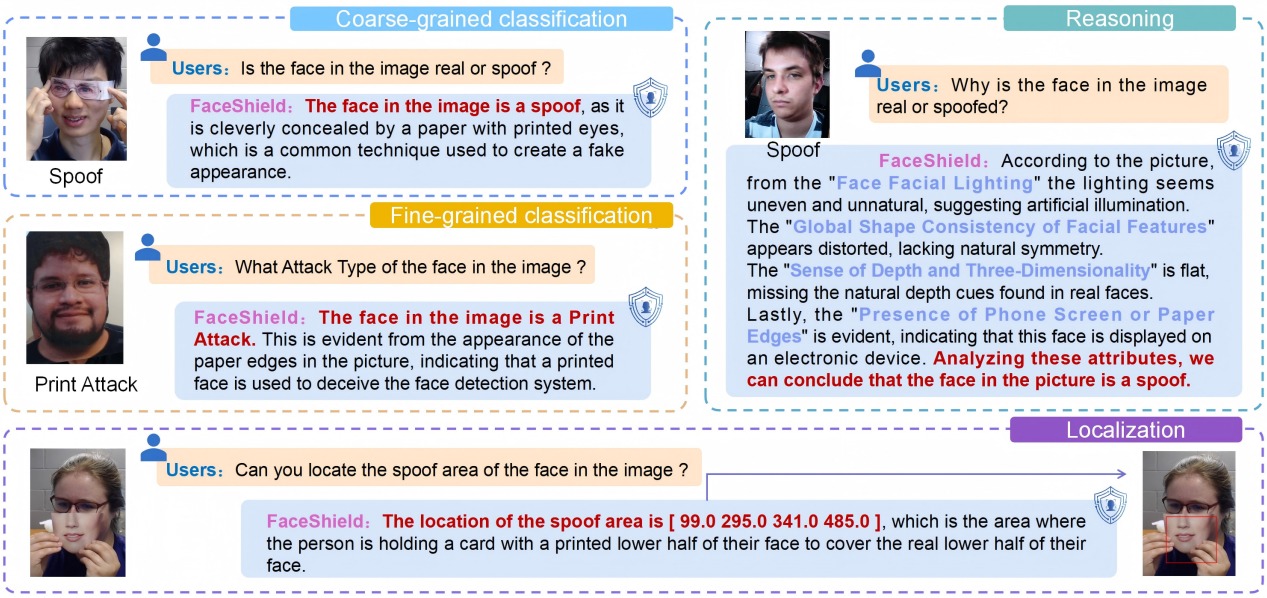

Figure 1: FaceShield Multi-task Response Demonstration. This figure shows the model's performance on four tasks: coarse-grained classification (real vs. spoofed faces), fine-grained classification (specific attack types like print attacks), reasoning (explaining spoofing using features such as lighting and symmetry), and localization (detecting spoofed regions). It highlights FaceShield's ability to handle diverse, complex questions accurately.

purpose MLLMs, revealing that their performance on FAS tasks still has room for improvement. (Zhang et al. 2025) introduced a model capable of performing classification and description attack type. However, the model is limited in its ability to handle more nuanced tasks, such as identifying specific attack types, reasoning, and localizing spoofed areas. These limitations underscore the broader challenges in training FAS MLLMs, including: (1) a lack of pretraining and SFT datasets specific to FAS tasks, (2) the need to extend traditional FAS tasks to fully exploit MLLM capabilities, and (3) the difficulty for general-purpose vision towers to capture the subtle distinctions between real faces and PAs, unlike with natural images.

Motivated by the above discussion, in this paper, we expand the traditional FAS task to include four sub-tasks (see Fig. 1 for examples): coarse-grained classification, fine-grained classification, reasoning, and attack localization. We then introduce FaceShield, an MLLM specifically designed for these tasks. As can be seen from Fig. 2(c), we propose a pretraining and SFT dataset generation pipeline. This pipeline constructs two multimodal FAS instruction datasets containing 50k dialogues for FaceShield training. To the best of our knowledge, FaceShield is the first FAS MLLM, equipped with multiple detection capabilities. Additionally, FaceShield-pre10K and FaceShield-sft45K are the first high-quality datasets that can be used to train a FAS-specific MLLM. Our main contributions include:

- We develop a novel data generation pipeline that utilizes a MLLM and predefined prompts, and construct two multimodal FAS instruction datasets (i.e., FaceShield-pre10K and FaceShield-sft45K) with 12 attack types. To our best knowledge, these are the first multitask instruction datasets for the FAS community.
- We propose FaceShield, the first multitask MLLM for FAS that is capable of coarse-grained classification, fine-grained classification, reasoning, and attack localization. FaceShield utilizes the Spoof-Aware Vision Perception (SAVP) and Prompt-guided Visual Token Masking (PVTM) strategies to enhance the discrimination of confusing attack areas.
- Extensive experiments demonstrate that FaceShield significantly outperforms previous specialized FAS models and general MLLMs across multiple datasets in various FAS evaluation tasks.

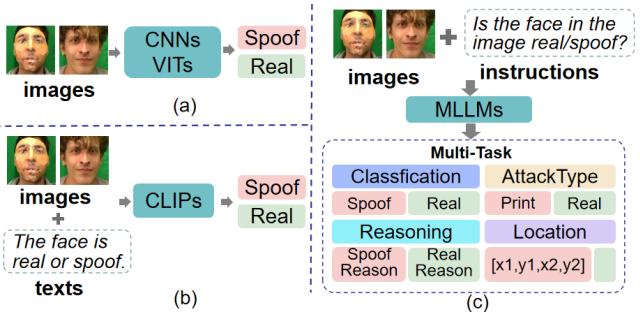

Figure 2: Pipelines of different FAS methods (a) traditional deep learning models, (b) multimodal models, and (c) MLLM

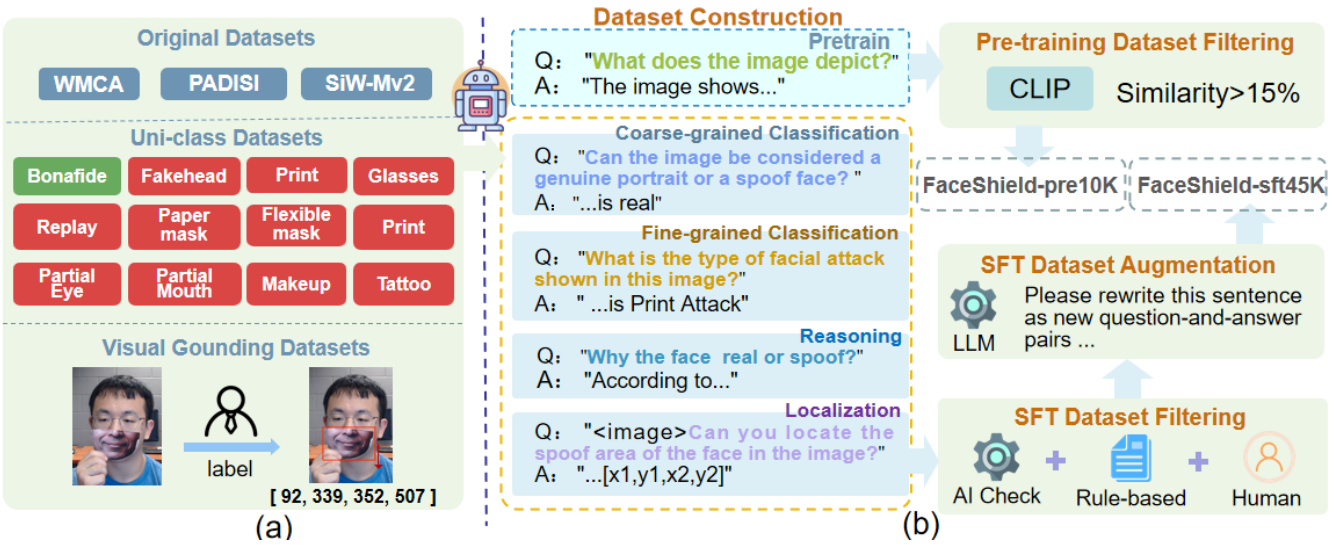

Figure 3: Construction pipeline of our proposed instruction datasets (i.e., FaceShield-pre10K and FaceShield-sft45K).The initial datasets (WMCA, PADISI, SiW-Mv2) are combined to form a uni-class dataset covering 12 spoofing types, with selected images annotated for visual grounding. Using MLLM with structured prompts, we generate two datasets: a pretraining dataset and an SFT dataset divided into four tasks (coarse-grained classification, fine-grained classification, reasoning, and localization). The pretraining data is filtered by CLIP for similarity, producing the FaceShield-pre10k dataset. SFT data undergoes multi-level filtering (LLM-based, keywords, and human reviews), followed by augmentation, resulting in the FaceShield-sft45k dataset. Additional details can be found in the appendix.

FaceShield-pre10K and FaceShield-sft45K

Dataset Collection

Fig. 3(a) illustrates the annotation process for existing FAS datasets. Based on the class types from WMCA (W) (George et al. 2020), PADSIS (P) (Rostami et al. 2021), and SiW-Mv2 (S) (Guo et al. 2022), we unify the annotation categories into 12 types: Bonafide, Fakehead, Print, Glasses, Replay, Paper mask, Flexible mask, Rigid mask, Partial Eye, Partial Mouth, Makeup, and Tattoo. We re-annotate all images at both image- and region-levels, resulting in 12,091 images with class labels and 3,139 images with bounding box annotations.

Instruction Construction

As shown in Fig. 3(b), we construct two instruction datasets using Bunny-Llama-3-8B-V (He et al. 2024). A system prompt containing class labels is used to guide the MLLM assistant in generating question-answer (QA) pairs, based on the task type and few-shot examples.

For the pretraining dataset FaceShield-pre10K, we generate image descriptions only, without task instructions. Pairs with a CLIP (Radford et al. 2021) similarity score below 15% are filtered out to ensure quality.

For the instruction-tuning dataset FaceShield-sft45K, MLLM-generated QA pairs are filtered using both manual and keyword-based strategies. The resulting high-quality seed set is then diversified using LLaMA3 (Dubey et al. 2024) to enhance linguistic richness and dialogue ability.

The dataset covers four tasks: (1) ***Coarse-grained classification***, which predicts whether a face is real or spoofed; (2) ***Fine-grained classification***, which identifies specific PA types beyond binary classification; (3) ***Reasoning***, which provides explanations and justifications before making a judgment; (4) ***Attack localization***, which outputs coordinates of attack regions if spoofing is detected.

FaceShield

Our goal is to train a FAS task-specific MLLM with two main objectives: 1) Enhance the visual encoder's ability to extract features from real faces and presentation attacks, and 2) Utilize the extensive knowledge stored in the LLM to improve the model's generalization capabilities when facing unknown domains.A naive training approach involves direct pre-training and SFT using RGB images and constructed QA data. However, the high similarity between real faces and PAs in RGB appearance poses significant challenges to this method. As shown in Fig. 4(a), Spoof-Aware Vision Perception (SAVP) combines images preprocessed based on prior knowledge, by extracting predefined local descriptor operators (Yu et al. 2024), with the original RGB image. Our complete model framework is shown in Fig. 4(b), RGB images and the extracted local descriptor images are fed into the vision encoder to extract vision token V_{RGB} and V_{SAV}, respectively. These features are then processed through the Prompt-Guided Vision Token Masking (PVTM) module, which extracts highly generalizable vision tokens. These tokens are sent to a projector to align with text prompt token P to produce V_{align}, which is then fed into the language foundation model for inference result $\mathcal{Y}$ as follows:

$$V_{align} = \text{Projection}(V_{RGB}, V_{SAV}) \quad (1)$$

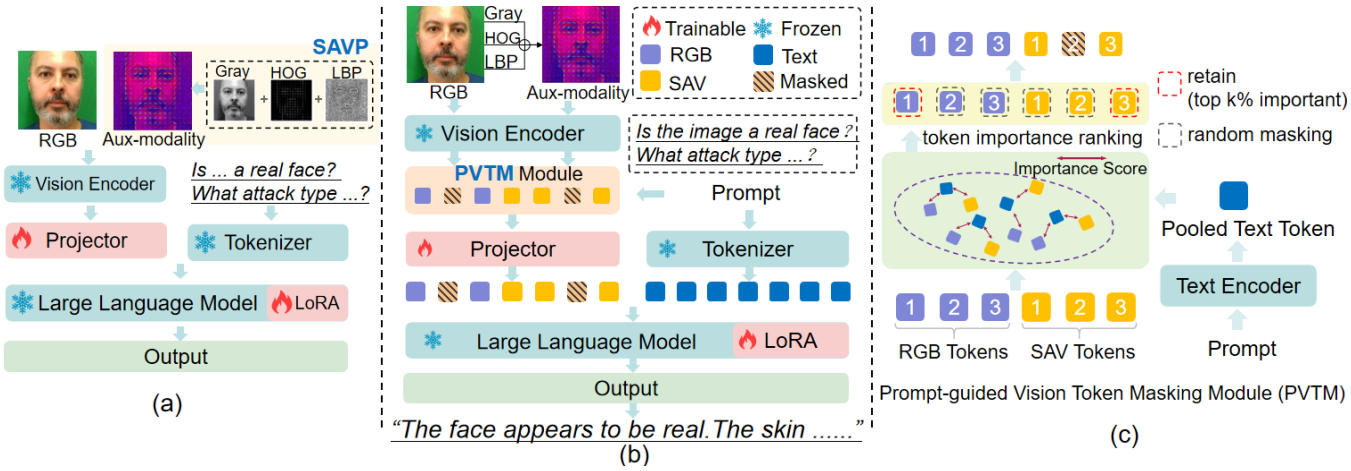

Figure 4: Proposed model architectures. (a) Proposed model with Spoof-Aware Vision Perception (SAVP). (b) Proposed model with SAVP and Prompt-Guided Vision Token Masking (PVTM). (c) Details about PVTM.

$$\mathcal{Y} = \text{MLLM}(V_{align}, P) \tag{2}$$

Spoof-Aware Vision Perception

Bonafide faces and PAs lack distinct discriminative features in RGB-based appearance space, whereas local descriptors (Yu et al. 2020, 2024; Xie et al. 2024) extracted through image preprocessing can enhance their subtle live/spoof clues. As shown in Fig. 4(a), we extract features from the original images using Local Binary Pattern (LBP) (Pietikäinen 2010), Gray, and Histogram of Oriented Gradients (HOG) (Dalal and Triggs 2005), and concatenate them. LBP and Gray-specific computations are as follows:

$$\begin{aligned} LBP &= \sum_{i=0}^{P-1} s(g_i - g_c) \cdot 2^i, \\ s(x) &= \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{3}$$

where g_c is the pixel value of the central pixel in the considered neighborhood, and g_i represents the pixel values of the P surrounding pixels.

$$\text{Gray}(I) = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B, \tag{4}$$

where R, G, and B are the red, green, and blue intensity values of the pixel, respectively.

The HOG calculates the gradient magnitude and direction at each pixel using edge detection operators and then divides the image into small, overlapping cells. Within each cell, gradients are binned according to their direction into histograms. Histograms from the cells within each block are concatenated and normalized based on the block's overall gradient energy. The final HOG descriptor is formed by the vector of these normalized histograms from all blocks.

We perform the above three steps of feature extraction on the original image, then concatenate these as three channels to form a complete image. This composite image is then fed into the vision encoder to extract features as complementary information. The spoof-aware vision token V_{SAV} and final vision input V for FaceShield as follows:

$$V_{SAV} = \text{Encoder}(\text{Concat}[LBP, Gray, HOG]) \tag{5}$$

$$V = \text{Concat}[V_{RGB}, V_{SAV}]) \tag{6}$$

Prompt-Guided Vision Token Masking

To further enhance the alignment between visual features and text prompts, and alleviate overfitting on spurious correlations, we leverage text prompts to guide token selection after the visual encoder. As shown in Fig. 4(c), the text tokens extracted from the prompt are pooled (Song et al. 2024), and then their similarity with all visual tokens is calculated. We assume that visual tokens with higher similarity are more relevant to subsequent tasks and even for the same image, the important visual tokens may not be consistent across different tasks. The calculation of similarity between visual tokens V_i and text prompt tokens P is as follows:

$$Sim(V_i, P) = \frac{V_i \cdot P}{\|V_i\| \|P\|} \tag{7}$$

Subsequently, we apply a softmax function to the similarities between all V_i and P, using the resulting values S^i_{rank} as an importance metric for each visual token. We then rank all tokens based on this importance calculated as follows:

$$S^i_{rank}(V_i, P) = \frac{e^{S(V_i,P)}}{\sum_j e^{S(V_j,P)}} \tag{8}$$

Afterward, we retain the top $k\%$ of vision tokens in importance. The remaining tokens are then randomly masked with a probability of $p\%$, reducing the influence of less important tokens while keeping acceptable information loss for the final decision-making process.

Training Details

We adopt a two-stage training strategy. In the first stage (pretraining), we perform continual pretraining to align visual

embeddings from a pretrained vision encoder with text embeddings using FaceShield-pre10K. In the second stage (supervised fine-tuning, SFT), we apply visual instruction tuning on FaceShield-sft45K to fully exploit the MLLM's capabilities across domain-specific multimodal tasks.

For LLM adaptation, we apply LoRA (Hu et al. 2021) with a rank of 128 and a scaling factor of 256 for each transformer block. Cross-entropy loss is used for next-token prediction in both stages. During pretraining, we update only the vision projector and the PVTM module for one epoch. In the SFT stage, we fine-tune the LoRA layers in the LLM along with the vision projector.

Experiments and Results

Protocols and Evaluation Metrics

We use 10% of each source dataset in FaceShield-sft45K to construct three test subsets: W, S, and P. For coarse-grained classification we perform both intra- and cross-dataset evaluations. For fine-grained classification, reasoning and attack localization, we conduct intra-dataset evaluation only. In intra-dataset testing, models are trained on all source data and evaluated on the combined test sets (W&S&P). For cross-dataset testing, two datasets are used for training (including pretraining and SFT), and the remaining one for testing (e.g., training on W and S, testing on P).For coarse-grained classification, fine-grained classification, and reasoning, we report Half Total Error Rate (HTER) (Yu et al. 2022) and Accuracy (ACC). For reasoning, we further evaluate reasoning quality using BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004), and METEOR (Banerjee and Lavie 2005). For attack localization, we report AP@40 and AP@50.

Implementation Details

We use Siglip (Zhai et al. 2023) as the visual encoder and Phi-3 (Abdin et al. 2024) as the language foundation model. PVTM retains the top 10% of the most important tokens and randomly masks 5% of the tokens in the remaining 90%. Adam optimizer is used in the pretrain stage with a learning rate of 5×10^{-4}. As for SFT stage, we decrease the learning rate to 2×10^{-4}. All experiments are conducted on a single NVIDIA A100 GPU. Each experiment is repeated 10 times on the model, and the final results are reported as the mean ± standard deviation.

Comparison with Existing Methods

Coarse-Grained Classification Task For the coarse-grained classification task, we compare FaceShield with state-of-the-art FAS methods (He et al. 2016; Wang et al. 2022; Zhou et al. 2022, 2023) and open-source MLLMs (Liu et al. 2023; Bai et al. 2023; Zhu et al. 2023; He et al. 2024).

Intra-dataset Testing. Table 1 demonstrates that Our FaceShield significantly outperforms three representative traditional FAS methods (He et al. 2016; Wang et al. 2022; Zhou et al. 2022). Moreover, our performance greatly exceeds the zero-shot capabilities of general MLLM. We also fine-tune the open-source MLLM (i.e., Bunny (He et al.

Method	ACC(%) ↑	HTER(%) ↓
Traditional		
ResNet (He et al. 2016)	97.55	2.32
PatchNet (Wang et al. 2022)	98.22	1.78
CoOp (Zhou et al. 2022)	98.73	1.27
MLLM		
LLaVA (Liu et al. 2023)	65.54	27.76
Qwen-VL (Bai et al. 2023)	51.94	38.70
Minigpt4 (Zhu et al. 2023)	26.86	65.50
Bunny (He et al. 2024)	81.20	17.87
Bunny (fine-tuned) (He et al. 2024)	98.23	1.52
FaceShield (Ours)	**99.41 ± 0.06**	**0.53 ± 0.06**

Table 1: Intra-dataset results on coarse-grained classification.

Method	ACC(%) ↑	HTER(%) ↓
W & S → P		
ResNet (He et al. 2016)	46.12	50.00
PatchNet (Wang et al. 2022)	77.18	22.87
IADG (Zhou et al. 2023)	72.96	27.01
FAS-AUG (Cai et al. 2024)	91.7	7.3
FaceShield (Ours)	**93.17 ± 0.22**	**6.37 ± 0.21**
W & P → S		
ResNet (He et al. 2016)	53.36	49.16
PatchNet (Wang et al. 2022)	56.16	45.37
IADG (Zhou et al. 2023)	57.20	42.81
FAS-AUG (Cai et al. 2024)	88.2	11.7
FaceShield (Ours)	**89.93 ± 0.15**	**10.3 ± 0.14**
S & P → W		
ResNet (He et al. 2016)	74.01	29.75
PatchNet (Wang et al. 2022)	78.15	41.50
IADG (Zhou et al. 2023)	78.55	26.27
FAS-AUG (Cai et al. 2024)	87.9	13.1
FaceShield (Ours)	**92.56 ± 0.08**	**5.71 ± 0.08**

Table 2: Cross-dataset results on coarse-grained classification. W, S, and P denote WMCA, SiW-Mv2, and PADISI, respectively.

Test Dataset	ACC(%) ↑	HTER(%) ↓
S & P & W → CASIA-MFSD	90.59	6.37
S & P & W → Replay-Attack	82.42	20.07

Table 3: Cross-dataset result on CASIA-MFSD and Replay-Attack.

Method	ACC(%) ↑
LLaVA (Liu et al. 2023)	16.39
Qwen-VL (Bai et al. 2023)	16.55
Minigpt4 (Zhu et al. 2023)	19.51
Bunny (He et al. 2024)	27.03
Bunny (Fine-tuned) (He et al. 2024)	94.43
FaceShield (Ours)	**95.81 ± 0.11**

Table 4: Results of fine-grained classification task.

Method	BLEU-1 (%) ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L (%) ↑	METEOR (%) ↑	ACC (%) ↑	HTER (%) ↓
LLaVA (Liu et al. 2023)	45.05	31.75	23.42	17.80	30.51	25.52	37.84	50.11
Minigpt4 (Zhu et al. 2023)	17.85	7.85	3.53	1.94	27.54	21.88	33.86	50.00
Qwen-VL (Bai et al. 2023)	20.92	14.53	11.01	8.77	21.45	12.49	47.64	41.49
Bunny (He et al. 2024)	33.64	27.12	22.65	19.33	36.74	19.12	50.68	39.73
Bunny (fine-tuned) (He et al. 2024)	89.57	86.96	84.91	81.29	80.15	51.64	98.56	1.16
FaceShield (Ours)	**90.89 ± 0.14**	**88.02 ± 0.15**	**85.75 ± 0.17**	**83.98 ± 0.19**	**82.98 ± 0.20**	**53.10 ± 0.16**	**99.29 ± 0.04**	**0.57 ± 0.04**

Table 5: Results of the reasoning task with metrics BLEU, ROUGE-L, METEOR, ACC, and HTER.

Method	AP@40 (%) ↑	AP@50 (%) ↑
Qwen-VL (Bai et al. 2023)	2.07	1.49
Lenna (Wei et al. 2023)	37.77	35.41
Sphinx (Lin et al. 2023)	47.86	46.30
Bunny (He et al. 2024)	73.50	71.65
Bunny (fine-tuned) (He et al. 2024)	92.30	89.71
FaceShield (Ours)	**97.78 ± 0.21**	**95.60 ± 0.19**

Table 6: Results of the attack localization task with metrics AP@40 and AP@50.

FaceShield-pre10K	Fine-grained Classification	Reasoning	
	ACC (%) ↑	ACC (%) ↑	HTER (%) ↓
×	94.78 ± 0.19	98.83 ± 0.06	0.94 ± 0.05
✓	**95.81 ± 0.11**	**99.29 ± 0.04**	**0.57 ± 0.04**

Table 7: Ablation study on pretraining w/ or w/o FaceShield-pre10K dataset.

2024)), selecting RGB images and language data from the dataset to conduct experiments on Bunny. We find that FaceShield also surpasses the well-tuned MLLM (Bunny) with 1% HTER decrease.

Cross-dataset Testing. Table 2 shows the performance of FaceShield in cross-domain scenarios, where we trained on two out of three selected datasets and tested on one. FaceShield demonstrates performance far exceeding traditional FAS models in cross-domain scenarios. Under the S&P→W protocol, it achieves the HTER of 5.72%, showcasing FaceShield's strong generalization capabilities compared to traditional methods. Further results in Table 3 confirm its robustness across unseen datasets like CASIA-MFSD and Replay-Attack.

Fine-Grained Classification Task Table 4 shows the results under the fine-grained classification task. For open-source MLLMs, we incorporated 12 types of attacks into the prompt, allowing it to respond with the correct type. For the fine-tuned MLLM and our FaceShield, we selected keywords from the responses for evaluation. It is evident that supervised fine-tuning can significantly improve the model's performance, with FaceShield achieving the best results.

Reasoning Task We also explore the models' reasoning capacity and Table 5 displays the results for the reasoning

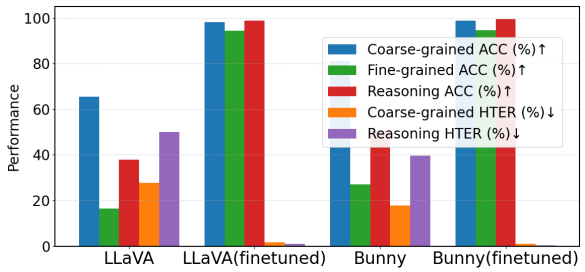

Figure 5: Comparison of performance after fine-tuning using our proposed dataset on LLaVA and Bunny models.

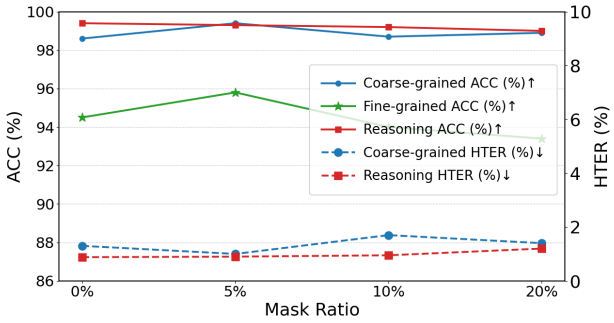

Figure 6: Ablation of visual token masking ratio p in PVTM on three tasks (i.e., coarse- & fine-grained classification, and reasoning)

task. General MLLMs perform poorly in both classification and reasoning. FaceShield significantly outperforms general MLLMs in both the reasoning process and judgment, and also exceeds open-source MLLMs (e.g., Bunny) fine-tuned on our dataset. FaceShield not only provides accurate results but also delivers detailed and correct reasoning, effectively enhancing the explainability of FAS methods.

Attack Localization Task It is vital to locate the spoof regions for explainable FAS, and Table 6 presents the results for the attack localization task. Due to the scarcity of attack localization annotations in general pre-trained datasets, general MLLMs perform poorly on this task. In contrast, FaceShield shows excellent results, achieving over 95% for both AP@40 and AP@50. It accurately locates attack areas, providing new insights into attack region detection for FAS tasks.

SAVP	PVTM	Coarse-grained classification		Fine-grained classification	Reasoning		Attack localization	
		ACC (%) ↑	HTER (%) ↓	ACC (%) ↑	ACC (%) ↑	HTER (%) ↓	AP@40 (%) ↑	AP@50 (%) ↑
×	×	98.23	1.52	94.43	98.56	1.16	92.30	89.71
×	✓	98.32	1.83	95.06	98.70	1.04	92.21	90.82
✓	×	98.73	1.06	94.59	**99.41**	**0.48**	97.09	95.21
✓	✓	**99.41 ± 0.06**	**0.53 ± 0.06**	**95.81 ± 0.11**	99.29 ± 0.04	0.57 ± 0.04	**97.78 ± 0.21**	**95.6 ± 0.19**

Table 8: Ablation results across different tasks.

Ablation Study

Effectiveness of the proposed datasets. We conduct pretraining and SFT with our proposed FaceShield-pre10K and FaceShield-sft45K on LLaVA (Liu et al. 2023) and Bunny (He et al. 2024) to evaluate the efficacy and generalization of our constructed datasets. As shown in Table 7, pretraining with the FaceShield-pre10K dataset significantly improves performance, with fine-grained classification accuracy and reasoning accuracy increasing, while HTER is reduced. This demonstrates that pretraining with FaceShield-pre10K enhances the model's capabilities in FAS-related tasks.

Additionally, the results in Fig. 5 show that fine-tuning on our dataset further boosts performance across three tasks for both LLaVA and Bunny. This validates the effectiveness of our dataset in enriching MLLMs with FAS-related knowledge and improving their overall performance in FAS tasks, supporting the efficacy of our advanced dataset construction pipeline.

Effectiveness of SAVP. Results from the first two rows in Table 8 show that leveraging local descriptors as complementary visual inputs significantly improves performance across four tasks, particularly in the attack localization task, where AP@40 and AP@50 increased by 5.6% and 5.94%, respectively. It indicates that prior knowledge-based auxiliary information significantly enhances the model's ability to distinguish easily confusable facial images. The local live/spoof details within the auxiliary data proves especially valuable for fine-grained attack region detection tasks.

Effectiveness of PVTM. It can be seen from the last two rows of Table 8 that the proposed PVTM provides reasonable improvements for FaceShield across three (coarse- and fine-grained classification, and attack localization) tasks. It indicates that masking less important tokens helps prevent the model from task-unrelated noises and spurious correlations. However, PVTM leads to a slight performance decrease on the reasoning task. This might be because masking partial visual tokens may compromise overall image perception and lose information for reasoning.

We further study PVTM with different visual token masking ratios p, as shown in Fig.6. Through experiments on three (coarse- and fine-grained classification, and reasoning) tasks, we find that masking 5% of the tokens strikes an optimal balance between reducing spurious correlations and preserving essential information. However, no improvement is found when masking more tokens due to severe information loss. Additionally, we investigate varying proportions of visual token retain ratio k. We preserve 0%, 10%, 20%, and 30% of the tokens, respectively, and then randomly mask p=5% of the remaining tokens. The results show that retaining k=10% visual tokens with strong importance achieves optimal performance. However, the more tokens preserved, the poorer the model performs, suggesting that as the importance of tokens decreases, the likelihood of spurious correlations increases.

Visualization and Analysis

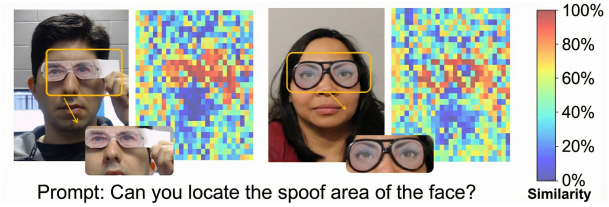

Figure 7: Importance visualization of SAV tokens for attack localization task.

Fig. 7 illustrates the visualization obtained after applying SAVP to the attack localization task, highlighting the effectiveness of SAV tokens. Compared to RGB tokens, which may suffer from dispersed attention, SAV tokens, once applied, can accurately focus on spoofed regions (e.g., eyeglass areas, hand-held masks). These tokens show strong alignment with the ground-truth annotations, producing more focused and interpretable activations around spoof artifacts. Additionally, visual tokens in the attack regions exhibit higher similarity scores with the task prompt. With PVTM applied, the model is further guided to focus on these deception-relevant areas, emphasizing the critical roles of both SAV tokens and PVTM in improving localization performance.

Conclusion

In this paper, we expand the FAS task into four subtasks: coarse-grained classification, fine-grained classification, reasoning, and attack localization, and propose the FaceShield, a specialized MLLM tailored for these FAS tasks. Considering the specific training data requirements of MLLM, we establish a pipeline for constructing datasets tailored for FAS task pre-training and supervised fine-tuning, resulting in the creation of the FaceShield-pre10K and FaceShield-sft45K datasets. This paper represents a preliminary exploration of MLLM for FAS task, and future works will focus on incorporating multiple visual modalities and refining the granularity of attack region localization.

Acknowledgments

This work was sponsored by the CCF-Tencent Rhino-Bird Open Research Fund, Guangdong Provincial Key Laboratory (Grant 2023B1212060076), the National Natural Science Foundation of China (Grant Nos. 62576076, 62025604, and 62411540034), the Science and Technology Development Fund of Macau (Project 0044/2024/AGJ), the Science and Technology Foundation of Guangdong Province (Project 2024A0505090002), the Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047), and the Hebei Provincial Natural Science Foundation (Grant No. F2024210005). The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University and the EIT High Performance Computing Platform.

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

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 Transactions on Pattern Analysis and Machine Intelligence*.

Cai, R.; Soh, C.; Yu, Z.; Li, H.; Yang, W.; and Kot, A. C. 2024. Towards Data-Centric Face Anti-Spoofing: Improving Cross-domain Generalization via Physics-based Data Synthesis. *IJCV*.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 886–893. Ieee.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

George, A.; and Marcel, S. 2021. On the Effectiveness of Vision Transformers for Zero-shot Face Anti-Spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 1–8.

George, A.; Mostaani, Z.; Geissenbuhler, D.; Nikisins, O.; Anjos, A.; and Marcel, S. 2020. Biometric Face Presentation Attack Detection With Multi-Channel Convolutional Neural Network. *IEEE Transactions on Information Forensics and Security*, 15: 42–55.

Guo, X.; Liu, Y.; Jain, A.; and Liu, X. 2022. Multi-domain Learning for Updating Face Anti-spoofing Models. In *ECCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, M.; Liu, Y.; Wu, B.; Yuan, J.; Wang, Y.; Huang, T.; and Zhao, B. 2024. Efficient Multimodal Learning from Data-centric Perspective. *arXiv preprint arXiv:2402.11530*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Huang, Z.; Xia, B.; Lin, Z.; Mou, Z.; and Yang, W. 2024. FFAA: Multimodal Large Language Model based Explainable Open-World Face Forgery Analysis Assistant. *arXiv preprint arXiv:2408.10072*.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, X.; Liu, A.; Yu, Z.; Cai, R.; Wang, S.; Yu, Y.; Wan, J.; Lei, Z.; Cao, X.; and Kot, A. 2025. Reliable and Balanced Transfer Learning for Generalized Multimodal Face Anti-Spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lin, Z.; Liu, C.; Zhang, R.; Gao, P.; Qiu, L.; Xiao, H.; Qiu, H.; Lin, C.; Shao, W.; Chen, K.; Han, J.; Huang, S.; Zhang, Y.; He, X.; Li, H.; and Qiao, Y. 2023. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. arXiv:2311.07575.

Liu, A.; Hui, M.; Zheng, J.; Yuan, H.; Yu, X.; Liang, Y.; Escalera, S.; Wan, J.; and Lei, Z. 2024. FM-CLIP: Flexible Modal CLIP for Face Anti-Spoofing. In *ACM Multimedia 2024*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.

Mu, L.; Bai, J.; He, X.; Ye, J.; Liang, X.; Yang, Y.; Zhuang, J.; and Hu, H. 2024. TeG-DG: Textually Guided Domain Generalization for Face Anti-Spoofing. arXiv:2311.18420.

Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; and Xiao, P. 2024. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. arXiv:arXiv:2402.02544.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pietikäinen, M. 2010. Local binary patterns. *Scholarpedia*, 5(3): 9775.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rostami, M.; Spinoulas, L.; Hussein, M.; Mathai, J.; and Abd-Almageed, W. 2021. Detection and Continual Learning of Novel Face Presentation Attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14851–14860.

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. *Visual Intelligence*, 3(1): 9.

Song, D.; Wang, W.; Chen, S.; Wang, X.; Guan, M.; and Wang, B. 2024. Less is More: A Simple yet Effective Token Reduction Method for Efficient Multi-modal LLMs. arXiv:2409.10994.

Srivatsan, K.; Naseer, M.; and Nandakumar, K. 2023. FLIP: Cross-domain Face Anti-spoofing with Language Guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 19685–19696.

Sun, G.; Qin, C.; Fu, H.; Wang, L.; and Tao, Z. 2024. STLLaVA-Med: Self-Training Large Language and Vision Assistant for Medical. In *EMNLP*.

Wang, C.-Y.; Lu, Y.-D.; Yang, S.-T.; and Lai, S.-H. 2022. 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 (CVPR)*, 20281–20290.

Wang, H.; Shi, Y.; Feng, J.; Yu, Z.; and Tao, Z. 2025. PNSS: Unknown Face Presentation Attack Detection with Pseudo Negative Sample Synthesis. *Computers, Materials & Continua*, 83(2).

Wei, F.; Zhang, X.; Zhang, A.; Zhang, B.; and Chu, X. 2023. Lenna: Language enhanced reasoning detection assistant. *arXiv preprint arXiv:2312.02433*.

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

Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; and Zhang, J. 2024. FakeShield: Explainable Image Forgery Detection and Localization via Multi-modal Large Language Models. ArXiv preprint arXiv:2410.02761.

Ye, Q.; Yu, Z.; Shao, R.; Cui, Y.; Kang, X.; Liu, X.; Torr, P.; and Cao, X. 2025. Cat+: Investigating and enhancing audio-visual understanding in large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yu, Z.; Cai, R.; Cui, Y.; Liu, X.; Hu, Y.; and Kot, A. C. 2024. Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. *International Journal of Computer Vision*, 1–22.

Yu, Z.; Qin, Y.; Li, X.; Zhao, C.; Lei, Z.; and Zhao, G. 2022. Deep learning for face anti-spoofing: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 5609–5631.

Yu, Z.; Qin, Y.; Zhao, H.; Li, X.; and Zhao, G. 2021. Dual-Cross Central Difference Network for Face Anti-Spoofing. In *IJCAI*.

Yu, Z.; Zhao, C.; Wang, Z.; Qin, Y.; Su, Z.; Li, X.; Zhou, F.; and Zhao, G. 2020. Searching Central Difference Convolutional Networks for Face Anti-Spoofing. In *CVPR*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

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

Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; and Mao, X. 2024. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20453–20463.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.