Scientific Reports

https://doi.org/10.1038/s41598-026-43800-5

Article in Press

Analyzing the effect of reasoning-based supervision on face anti-spoofing

Received: 30 December 2025

Accepted: 6 March 2026

Published online: 13 March 2026

Cite this article as: Min J., Lim K., Kim M. *et al.* Analyzing the effect of reasoning-based supervision on face anti-spoofing. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-43800-5

Jimin Min, Kyungtae Lim, Minjun Kim, Dongsu Kim, Seoyeon Oh, Eunkyung Kim & Haneol Jang

Analyzing the Effect of Reasoning-Based Supervision on Face Anti-Spoofing

Jimin Min[1,+,†], Kyungtae Lim[2,+], Minjun Kim[2], Dongsu Kim[1], Seoyeon Oh[1], Eunkyung Kim[3,*], and Haneol Jang[1,*]

[1]Hanbat National University, Department of Computer Engineering, 125 Dongseo-daero, Yuseong-gu, Daejeon 34158, Republic of Korea
[2]KAIST, Graduate School of Culture Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
[3]Hanbat National University, Department of Artificial Intelligence Software, 109 Jiphyeonbuk-ro, Sejong 30139, Republic of Korea
[*]Corresponding authors: ekim@hanbat.ac.kr; hejang@hanbat.ac.kr
[+]These authors contributed equally to this work
[†]Present address: Datamaker Inc., 871 Yuseong-daero, Yuseong-gu, Daejeon 34127, Republic of Korea.

ABSTRACT

Face anti-spoofing (FAS) has become a crucial component in securing face recognition systems against presentation attacks, such as printed photos, replay videos, and 3D masks. While recent advances have improved generalization to unseen spoofing attempts, many existing methods remain black-box models that provide binary decisions without interpretable reasoning. In this paper, we investigate explainable face anti-spoofing from a supervision-centric perspective, using a vision–language model (VLM) to analyze how natural language explanations influence model behavior. To enable this study under controlled conditions, we construct an explanation-augmented benchmark by enriching four standard FAS datasets—MSU-MFSD, CASIA-FASD, Replay-Attack, and OULU-NPU—with both vanilla and reasoning-structured captions generated via the GPT-4o API. We further adopt a dual-objective training strategy that combines spoof classification loss with explanation generation loss, allowing us to examine the effect of explanation-based supervision while keeping the backbone architecture fixed. Through extensive cross-dataset evaluations, we show that reasoning-style captions can enhance detection performance and domain generalization in many settings, while also introducing inductive biases that may degrade performance when emphasized cues are misaligned with unseen attack types. These findings suggest that explanations in FAS should be viewed not only as interpretable outputs, but also as controllable training signals that shape generalization behavior. To support reproducibility, we publicly release the explanation annotations and associated metadata—excluding all face images—via a Hugging Face repository at https://huggingface.co/datasets/DescriptiveFAS/MCIO_public.

Introduction

Face recognition systems[1] have become ubiquitous in various applications, from smartphone authentication to secure access control. However, these systems are vulnerable to presentation attacks or spoofing attempts, where attackers try to deceive the system using printed photos[2], digital screens[3], or 3D masks[4]. Face Anti-Spoofing (FAS) has thus emerged as a critical security component that determines whether the captured face is from a genuine live person or a spoofing medium[5]. With the widespread adoption of face recognition in security-critical domains, developing robust FAS systems has become an imperative challenge in biometric security research.

While significant progress has been made in FAS research, most existing approaches rely heavily on discriminative methods that produce binary decisions without providing interpretable reasoning for their determinations. This "black box" nature of current FAS systems poses challenges in real-world deployments, where understanding the basis of decisions is crucial for security personnel and system administrators. In practical operating environments, explanations are not merely user-facing justifications, but serve as analytical tools for diagnosing failure cases, identifying spurious visual cues learned during training, and understanding why models fail to generalize across domains. As a result, binary predictions alone are often insufficient for reliable deployment and maintenance of FAS systems under evolving attack scenarios.

Rather than designing a fully interpretable discriminative model, we adopt a vision–language model (VLM) as a means to impose explicit and controllable supervision during training. The motivation for using a VLM is not explanation generation per se, but the ability to leverage language as a supervisory signal that constrains which visual cues the model should attend to when learning spoof-related representations. By expressing spoofing rationales in natural language, the training process can explicitly guide the model toward semantically meaningful and interpretable decision cues, which is difficult to achieve using

conventional discriminative objectives alone.

Existing research on explainable face anti-spoofing has largely focused on two directions: post-hoc visualization methods that highlight salient regions after prediction, and task reformulation approaches that jointly generate predictions and explanations. However, these methods typically treat explanations as outputs of the model, rather than as controllable supervision signals during training. Consequently, how the structure and content of explanations influence representation learning and domain generalization in FAS models remains underexplored.

Although post-hoc explainable FAS methods can retrospectively visualize salient regions or features, they provide limited insight into how different forms of supervision shape model behavior during training. In particular, existing approaches lack a controlled setting in which the influence of explicitly guided decision rationales can be isolated and systematically examined under identical architectures and data distributions. In this work, we define the performance gap as the absence of a controlled framework for analyzing how explicitly guided decision rationales—imposed at training time—affect spoof detection performance and domain generalization. By treating explanations as supervision signals rather than interpretative outputs, our approach enables principled comparison between models trained to rely on different types of visual evidence, which is not achievable with post-hoc explainability alone.

To facilitate a controlled investigation of explanation-based supervision, we introduce a new benchmark that augments widely-used FAS datasets (MSU-MFSD **(M)**[6], CASIA-FASD **(C)**[7], Idiap Replay-Attack **(I)**[8] and OULU-NPU **(O)**[9]) with detailed explanations of spoofing detection reasoning, created using the GPT-4o API. This benchmark enables systematic analysis of how different explanation formats affect representation learning and generalization, rather than serving solely as a resource for explanation generation.

We hypothesize that training models with captions that emulate human reasoning patterns will lead to improved performance in both spoofing detection and reasoning capabilities. Each reasoning-style annotation explicitly describes the visual cues involved in spoofing detection, intermediate observations, and the causal relationship between observed cues and the final decision. This approach enables our model to learn both discriminative features for spoofing detection and structured reasoning patterns for explaining its decisions.

However, training a model to simultaneously perform spoofing detection and generate explanations presents unique challenges. We observed that models trained solely on explanation generation often show decreased performance in the fundamental binary classification task of spoofing detection. To address this issue, we adopt a dual-objective training strategy that treats spoofing detection as the primary task and explanation generation as auxiliary supervision. Specifically, we add a linear classification layer that operates on the vision encoder's features, allowing us to directly optimize for spoofing detection accuracy while maintaining the model's ability to generate meaningful explanations.

In summary, our main contributions are as follows.

- We investigate how treating explanations as explicit supervision signals, rather than post-hoc outputs, affects representation learning in face anti-spoofing models.
- We introduce an explanation-augmented FAS benchmark that enables controlled comparison between different supervision formats.
- We analyze how reasoning-structured supervision influences cross-domain generalization, highlighting both its benefits and potential brittleness under unseen attack distributions.

Related Work

The task of face anti-spoofing (FAS) has seen significant advancements with the adoption of deep learning-based methods, primarily aimed at enhancing generalization performance on unseen datasets or unknown attack types.

Domain generalization through visual representation learning. Early CNN-based approaches focused on extracting spoof-discriminative features that remain stable across domains. Methods such as[10,11] leveraged meta-learning and metric learning strategies, while subsequent works[12–14] emphasized isolating liveness-related cues or modeling fine-grained local patterns. More recent studies[15,16] revisited optimization dynamics and domain-gap assumptions to improve cross-dataset robustness. Transformer-based architectures further extended this direction by enhancing representational capacity and adaptation efficiency.[17] introduced adapter-based few-shot transfer for vision transformers. Along this line, S-adapter[18] proposes to generalize vision transformers for face anti-spoofing by introducing statistical tokens into the transformer architecture. By keeping the backbone fixed and injecting lightweight adapter modules, S-adapter demonstrates that parameter-efficient adaptation can substantially improve generalization without increasing model capacity. Similarly, DiVT[19] and the hybrid CNN–Transformer framework in[20] improve cross-domain robustness through domain-invariant objectives and cross-stage feature fusion. These studies collectively suggest that while larger or more expressive transformer backbones can improve detection accuracy, competitive performance can also be achieved through architectural adaptation and lightweight design

choices, rather than scaling model size alone. Overall, these methods primarily operate at the level of visual feature learning and architectural refinement, aiming to improve generalization without altering the nature of supervision.

Expanding modalities and supervision signals. To enhance robustness beyond pure RGB appearance, several works incorporate complementary modalities. Beyond purely visual cues, Beyond the Pixel World[21] demonstrates that integrating acoustic signals with facial imagery can significantly improve spoof detection performance on mobile devices, highlighting the value of complementary modalities in real-world scenarios. Similarly, M3FAS[22] proposes a multimodal mobile face anti-spoofing system that combines multiple sensory inputs to achieve robust performance under challenging acquisition conditions. In parallel, multimodal supervision has also been explored through vision–language alignment. Works such as[23,24] align visual representations with textual descriptions or prompts to facilitate cross-domain adaptation. These approaches improve robustness by expanding input modalities or leveraging language guidance, but do not explicitly analyze how decision rationales are shaped during training. In most cases, language serves as auxiliary semantic guidance rather than as an explicit structural constraint on reasoning processes during learning.

Explainable face anti-spoofing. Explainable FAS (X-FAS) aims to improve interpretability while maintaining detection performance. Concept-based approaches such as SPED[25] localize spoof-related evidence but primarily address *where* cues appear rather than *how* they inform decisions. More recent work has emphasized explainability as a means for system-level analysis rather than mere visualization. Unveiling explainability in face anti-spoofing[26] combines hybrid feature extraction with XAI-guided aggregation to analyze how different feature sources contribute to spoof detection, positioning explainability as a diagnostic tool for understanding model behavior. Task-reformulation approaches further integrate prediction and explanation. I-FAS[27] casts FAS as a VQA problem, coupling classification and explanation at inference time. In contrast, FaceShield[28] employs a multimodal large language model to perform explainable face anti-spoofing, generating free-form textual explanations alongside spoof predictions. While effective in demonstrating the feasibility of MLLM-based explainable FAS, such approaches primarily focus on explanation generation at inference time.

Beyond the FAS domain, explainability has also been explored as a tool for system analysis and robustness evaluation. For instance, ADMM-based adversarial false data injection attacks[29] and EVADE[30] leverage model interpretability to analyze failure modes and vulnerabilities in safety-critical systems. Although these studies address different application domains, they share a common perspective that explanations can serve as analytical instruments for understanding and diagnosing model behavior. In the context of face anti-spoofing, the diversity of physical and digital attack types, as summarized in Digital and physical face attacks[31], further underscores the need for analysis-oriented explainability.

Our work differs from these methods by treating explanations as an explicit and controllable supervision signal during training, rather than as post-hoc outputs or task-coupled responses. By fixing the backbone architecture and varying only the explanation format, we enable a controlled analysis of how reasoning-structured supervision influences detection performance and domain generalization. By reframing explanations as training-time supervisory signals rather than inference-time artifacts, our approach shifts the focus from interpretability as output to interpretability as a mechanism for shaping representation learning under domain shifts.

Explainable FAS

Proposed Datasets

To facilitate the development of explainable face anti-spoofing (FAS) models, we created an enriched dataset by utilizing the most widely used benchmark datasets in the FAS domain: MSU-MFSD **(M)**[6], CASIA-FASD **(C)**[7], Idiap Replay-Attack **(I)**[8], OULU-NPU **(O)**[9]. In this work, we define *human-like reasoning* in the context of face anti-spoofing as a structured explanatory process that (i) explicitly references spoof-relevant visual cues, (ii) decomposes the decision into intermediate analytical steps, and (iii) establishes a causal link between observed evidence and the final live/spoof judgment. These datasets are well-regarded for their comprehensive coverage of common attack types, specifically print and replay attacks, and include video recordings of 35 to 55 subjects per dataset, capturing diverse facial appearances and environmental conditions. As the original datasets consist of video data, we preprocessed the videos to generate image-based datasets, selecting frames under specific criteria. To construct the image-based dataset, we grouped frames by their original video directories and selected two representative frames per video: one from the early segment and another from the middle. This strategy ensures that each spoofing or live sample is represented with diverse temporal cues while minimizing redundancy.

To improve the quality of the generated dataset, we incorporated meta-information, including live/spoof labels and spoofing attack types (e.g., print, replay), during the data generation process. For each sample, we generated detailed captions that provide reasoning for the spoofing decision, grounded in both visual and contextual cues. To ensure consistency across generated explanations, the caption generation process was tightly constrained through fixed system messages, structured instruction prompts, and a predefined reasoning schema. Specifically, all reasoning-based captions followed an identical six-stage analytical structure and were generated in a label-aware manner, which significantly reduced variability in reasoning depth and format across samples. We leveraged the GPT-4o API to generate these reasoning-style captions, as it was one of the

Prompting GPT-4o to generate vanilla data

```
messages = [ {"role":"system", "content": """You are a face anti-spoofing analyst specializing in holistic pattern analysis.

Your task is to analyze and explain the characteristics that make the given image appear to be a <live/spoof> image.
You are given a <live / spoof> image, <'' / its attack type>, and resolution.
However, respond as if you don't know whether the given image is live or spoof. Assume that the conclusion that it's a <live / spoof> image is a result of your analysis.
Also, the resolution of the given image will be referenced but not mentioned directly in the response.
Be as detailed as possible in your response.

Note: Reflections in eyeglasses, if present, are not indicators of spoofing. Do not analyze or reference reflections in your response. Focus on other critical characteristics,
such as texture, lighting consistency, depth map and other features that clearly indicate <'liveness' / 'spoofing'>.

<'' / (Additional Information)
* printed_photo:
This type of spoofing attack involves taking a photograph of a real person, printing that photo, and then photographing the printed photo again.
* replay:
This type of spoofing attack involves taking a photograph of a real person, displaying that photo on a device's screen, and then photographing the displayed image again.>
"""}]

for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample['context']})
    messages.append({"role": "assistant", "content": sample['response']})
messages.append({"role": "user", "content": Focus on the face in the image and explain the characteristics that indicate the given image represents a <live / spoof> image.

Disregard any reflections in glasses if present, as they are not indicative of spoofing. Do not mention or analyze any aspect of eyeglasses reflections in your response.

Analyze other aspects like texture, lighting, depth and other features that suggest <liveness / spoofing>.
It is imperative to analyze exclusively and rigorously from the perspective of the face within the image. Disregard all other aspects of the image and focus solely on facial
features and characteristics. Exclude information related Background and Resolution.})
```

Figure 1. The system message and instruction used to generate vanilla-style captions. The prompt is designed to elicit general, concise descriptions for spoof classification without structured reasoning, forming the baseline in our comparative analysis.

most reliable models available at the time of dataset construction for producing consistent and semantically stable reasoning outputs at scale. Caption generation was performed under fixed and structured instruction prompts.

To enable the model to generate contextually accurate and unbiased explanations, we designed the process to guide GPT-4o to respond as if independently deducing the details, while still being aware of the provided information. To achieve high-quality data generation, we employed a two-step process. First, GPT-4-turbo was utilized to create initial drafts of few-shot examples, outlining structured and coherent responses. These drafts were then manually refined to ensure alignment with the intended reasoning format and inclusion of nuanced visual and contextual insights. The refined few-shot examples were subsequently used to guide GPT-4o in generating detailed, context-rich captions for the dataset.

We hypothesized that guiding the model to generate explicit reasoning-based captions that mimic human reasoning processes would enhance its ability to identify and reason about spoofing attempts. To enable comparative analysis of the impact of reasoning, we generated both a dataset with reasoning components and a vanilla dataset without them.

The prompts used to generate vanilla data are shown in Figure 1, and an example output is illustrated in Figure 2. Additional caption examples from MSU-MFSD, CASIA-FASD, Replay-Attack, and OULU-NPU are provided in the Supplementary Information. Due to dataset licensing restrictions, these examples are presented as text-only annotations without releasing the corresponding facial images.

Similarly, the prompts used to generate reasoning data are shown in Figure 3, and an example output is provided in Figure 4. Further dataset-specific examples and label structures are included in the Supplementary Information to clarify how reasoning captions are aligned with spoof labels across different benchmarks.

The reasoning captions were designed following the step-by-step analytical process that humans employ when determining image authenticity. Specifically, we structured the analysis into six stages: initial observation, artifact detection, feature analysis, examine lighting and shadows, consider context, and final judgment. This structured formulation serves not only as an explanation format, but also as an explicit constraint on the model's learning process, encouraging consistent alignment between visual evidence and decision rationale across the dataset.

The prompts used to generate reasoning data are shown in Figure 3. It is meticulously designed to elicit reasoning-based responses, enabling the creation of captions that provide robust justifications for spoofing classifications. In Figure 4, we present an example of the generated data, illustrating the interpretability and depth of the enriched dataset. Similarly here, due to issues with the model incorrectly generating descriptions about glasses, we explicitly instructed it to disregard mentions of glasses in its analysis.

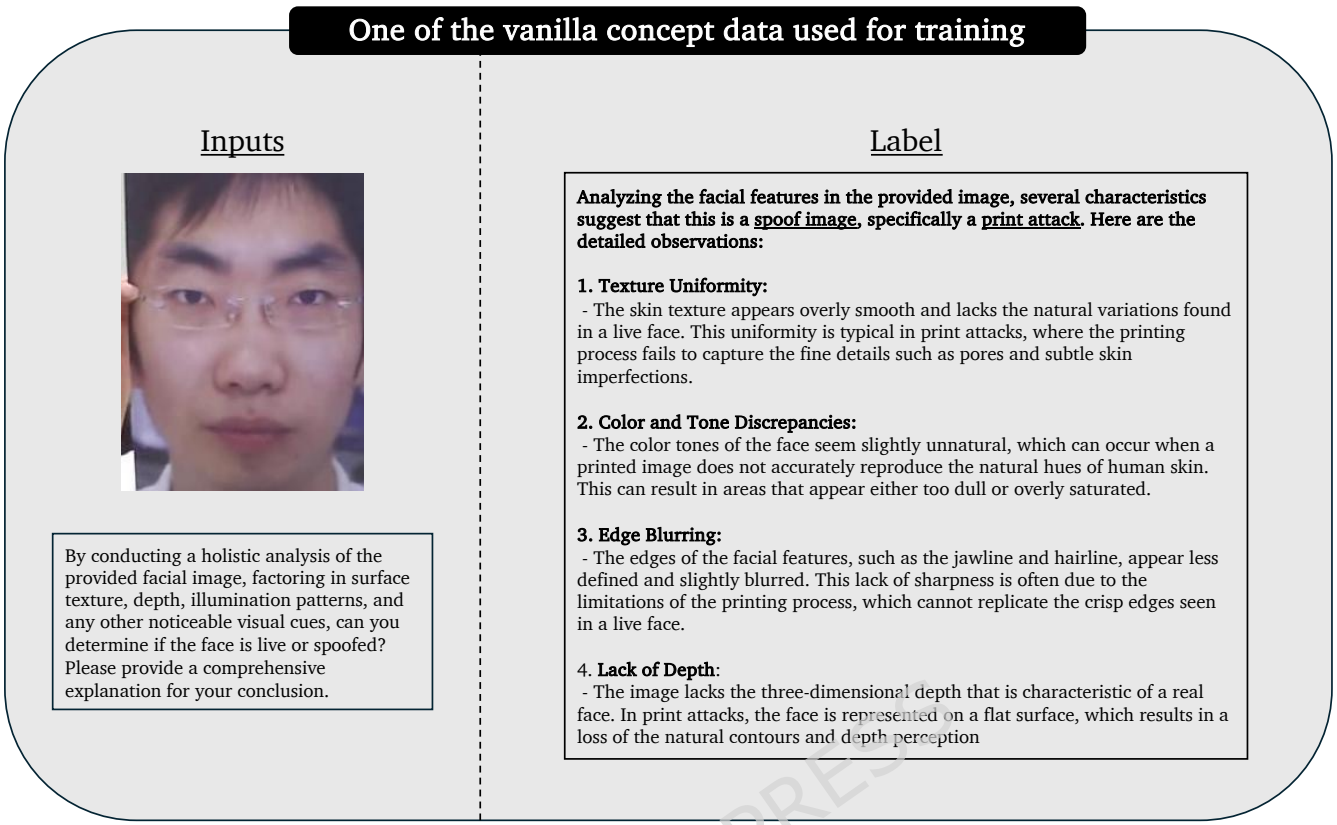

Figure 2. An example of a vanilla caption generated using the prompt in Figure 1. The explanation provides a high-level justification for the spoofing decision, demonstrating interpretability under minimal guidance.

Ethical Statement

All experiments and methods in this study were carried out in accordance with relevant guidelines and regulations. The use of publicly available benchmark datasets (MSU-MFSD, CASIA-FASD, Replay-Attack, OULU-NPU, and SiW-Mv2) for research purposes was conducted following the protocols approved by the respective institutional review boards or licensing committees of the original data collectors. Furthermore, we confirm that informed consent was obtained from all participants (and/or their legal guardians) for the collection of their data and for the publication of identifying information/images in an online open-access publication by the original dataset providers. Specifically, the face images presented in Figures 2, 4, 5, and 6 are from the CASIA-FASD dataset, for which explicit permission for publication has been granted.

Training Strategy

Figure 5 provides an overview of the proposed explanation-based face anti-spoofing framework. Based on this architecture, we describe the training strategy in detail, focusing on the dual-objective learning formulation and the parameter-efficient adaptation of the vision–language backbone.

To enable the model to simultaneously perform spoofing classification and generate explanatory captions, we adopt a dual-objective learning strategy. We introduce a linear classification head that takes the visual features from the vision encoder and predicts whether the input face is live or spoofed. The model is trained with two objectives: (1) a cross-entropy loss for generating natural language explanations, and (2) another cross-entropy loss for binary spoof classification.

These losses are combined using a weighted summation:

$$\mathscr{L}_{\text{total}} = \alpha \cdot \mathscr{L}_{\text{cls}} + \beta \cdot \mathscr{L}_{\text{cap}}, \tag{1}$$

The weighting scheme reflects our design principle that spoof detection is the primary task, while explanation generation serves as auxiliary supervision. Accordingly, we set $\alpha = 0.7$ and $\beta = 0.3$ to prevent the captioning objective from dominating the optimization, which we observed could degrade classification performance under limited FAS data. Our goal is not to exhaustively tune these hyperparameters, but to adopt a stable and interpretable balance that enables analysis of explanation-based supervision effects.

Prompting GPT-4o to generate reasoning data

```
messages = [ {"role":"system", "content": """You are a seasoned forensic expert working in a high-tech AI lab specializing in facial forgery detection.

You analyze facial images with precision, identifying subtle signs of manipulation and determining whether an image is real or a forgery. Your expertise is essential for training real-world AI systems used in high-value environments like security and digital authentication.

However, for this task, assume that you are only given a <unspoofed / spoofed> image and that the conclusion that it is a <live / spoof> image is the result of your analysis.

However, write your inference results as if you don't know whether the given image is live or spoofed.
It is essential to base your analysis on the actual visible facial features in as much detail as possible.

Note: Reflections involving the eyes should not be considered a factor in determining whether an image is real or spoofed. Do not mention or analyze any relevant aspects in your response. And the resolution of the given image will be referenced but not mentioned directly in the response.
"""}
]

for sample in fewshot_samples:
    messages.append({"role":"user“, "content":sample[‘context’]})
    messages.append({"role":"assistant", "content":sample[‘response’]})
messages.append({"role":"user", "content": Imagine you're a forensic investigator performing real-time facial image analysis. Find features that indicate a given image is a <real / spoofed> image. Then follow the steps below to make detailed and reliable inferences as if you didn't know it was a <live / spoof> image:

1. Initial observation
- Start with an overall assessment of the image, looking for irregularities or anomalies in facial features.
2. Artifact detection
- Look for visual artifacts such as blurriness, pixelation, or mismatched edges.
3. Feature analysis
- Examine specific facial features, such as eyes, nose, mouth, and skin texture, for signs of unnatural changes or inconsistencies.
4. Examine lighting and shadows
- Analyze the consistency of lighting, shadows, and reflections across the image to detect inconsistencies that could indicate manipulation.
5. Consider context
- Consider background information or context that may affect the authenticity of the image.
6. Make a final judgment
-  Decide whether the image is real or spoofed.

Note: Do not mention the features involved in the reflection and the resolution of the image directly in your response.

(Additional Information)
* printed_photo:
This type of spoofing attack involves taking a photograph of a real person, printing that photo, and then photographing the printed photo again.
* replay:
This type of spoofing attack involves taking a photograph of a real person, displaying that photo on a device's screen, and then photographing the displayed image again.

})
```


Figure 3. The system message and instruction used to generate reasoning-based captions. The prompt guides the model to produce structured, step-by-step reasoning aligned with six analytical stages, enabling detailed and interpretable spoofing analysis.

In the vanilla+reasoning setting, only half of the training samples retain reasoning-based captions, while the remaining samples use vanilla captions. This hybrid design is motivated by the observation that enforcing strong reasoning constraints on all samples can lead to over-reliance on specific explanation patterns, potentially harming generalization to unseen attack types. By mixing vanilla and reasoning supervision, we treat reasoning captions as a form of regularization rather than hard constraints, balancing learning stability with interpretability.

We employ LoRA for parameter-efficient adaptation of the large pretrained vision–language backbone. We emphasize that our objective is not to identify the optimal LoRA configuration for face anti-spoofing. Instead, we adopt widely used LoRA insertion points (attention projection layers) and moderate rank settings to minimally intervene in the backbone while enabling task-specific adaptation. This design choice allows us to isolate and analyze the effect of explanation-based supervision without confounding factors introduced by extensive architectural or capacity changes.

Concretely, LoRA adapters are inserted into the attention projection layers of the transformer blocks, namely `q_proj`, `k_proj`, and `v_proj`, for both the language model and the multimodal fusion/decoder modules. In addition, we train lightweight output-related modules used in our dual objectives, including `lm_head` (caption generation), `vision_cls_head` (binary spoof classification), and `multi_modal_projector` (vision–language feature projection). Thus, the backbone remains frozen, and the model is *partially* fine-tuned via LoRA and task heads rather than fully fine-tuned.

This dual-objective setup jointly optimizes classification accuracy and explanation generation while requiring only minimal

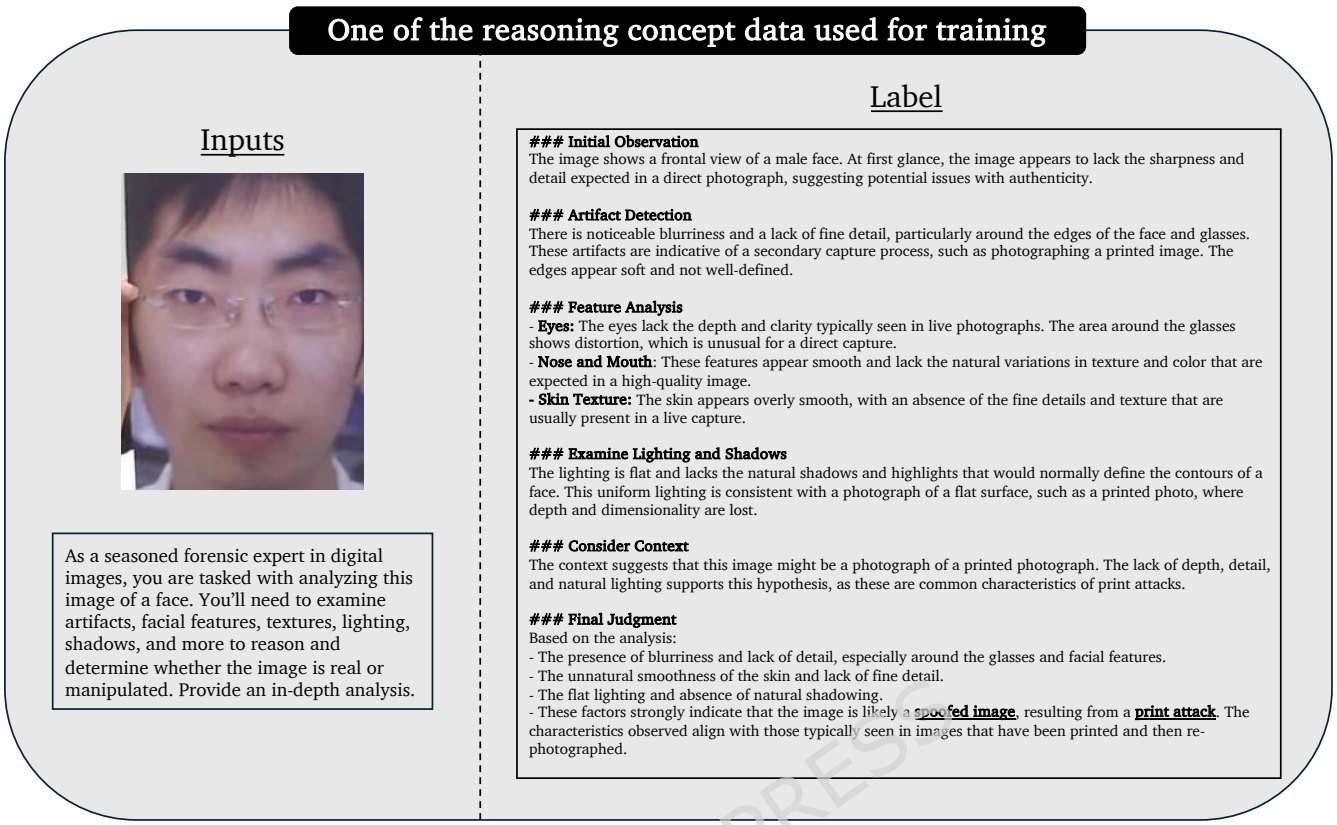

Figure 4. An example of a reasoning-based caption generated using the prompt in Figure 3. The output demonstrates the model's capacity to produce comprehensive, stage-wise justifications grounded in spoof-relevant visual cues.

parameter updates for task adaptation.

Experiments

Experimental Settings

Datasets and Protocols

We conduct our experiments using four widely adopted face anti-spoofing datasets: MSU-MFSD **(M)**[6], CASIA-FASD**(C)**[7], Replay-Attack**(I)**[8], and OULU-NPU**(O)**[9], which include various spoof types and recording environments. To further evaluate performance on unseen attack types, we additionally use the SiW-Mv2[32] dataset.

To evaluate the domain generalization performance of our model, we adopt three experimental protocols. First, the MCIO Leave-One-Out protocol is a widely used setting in face anti-spoofing research, where the model is trained on three datasets and tested on the remaining one. For example, a model trained on M, C, and I is evaluated on O. This setup reflects realistic deployment scenarios where the test distribution differs from the training distribution.

Second, we define the MCIO-Subset-to-SiW-Mv2 protocol, where four separate models are trained on different three-dataset combinations from MSU-MFSD, CASIA-FASD, Replay-Attack, and OULU-NPU, and all are evaluated on the SiW-Mv2 dataset[32]. Unlike MCIO datasets which mainly consist of print and replay attacks, SiW-Mv2 includes 13 diverse spoof types. This protocol enables us to assess how well models trained on limited spoof types generalize to a broader range of unseen attacks.

Finally, the MCIO-Full-to-SiW-Mv2 protocol involves training a single model on the full MCIO dataset (i.e., M, C, I, O) and evaluating it on SiW-Mv2. This setting allows us to examine whether increasing the diversity of training data improves generalization to unfamiliar spoofing scenarios.

Evaluation Metrics

We evaluate our model using commonly adopted metric in recent face anti-spoofing studies, Half Total Error Rate (HTER). HTER is defined as the average of the False Acceptance Rate (FAR) and False Rejection Rate (FRR), capturing the trade-off

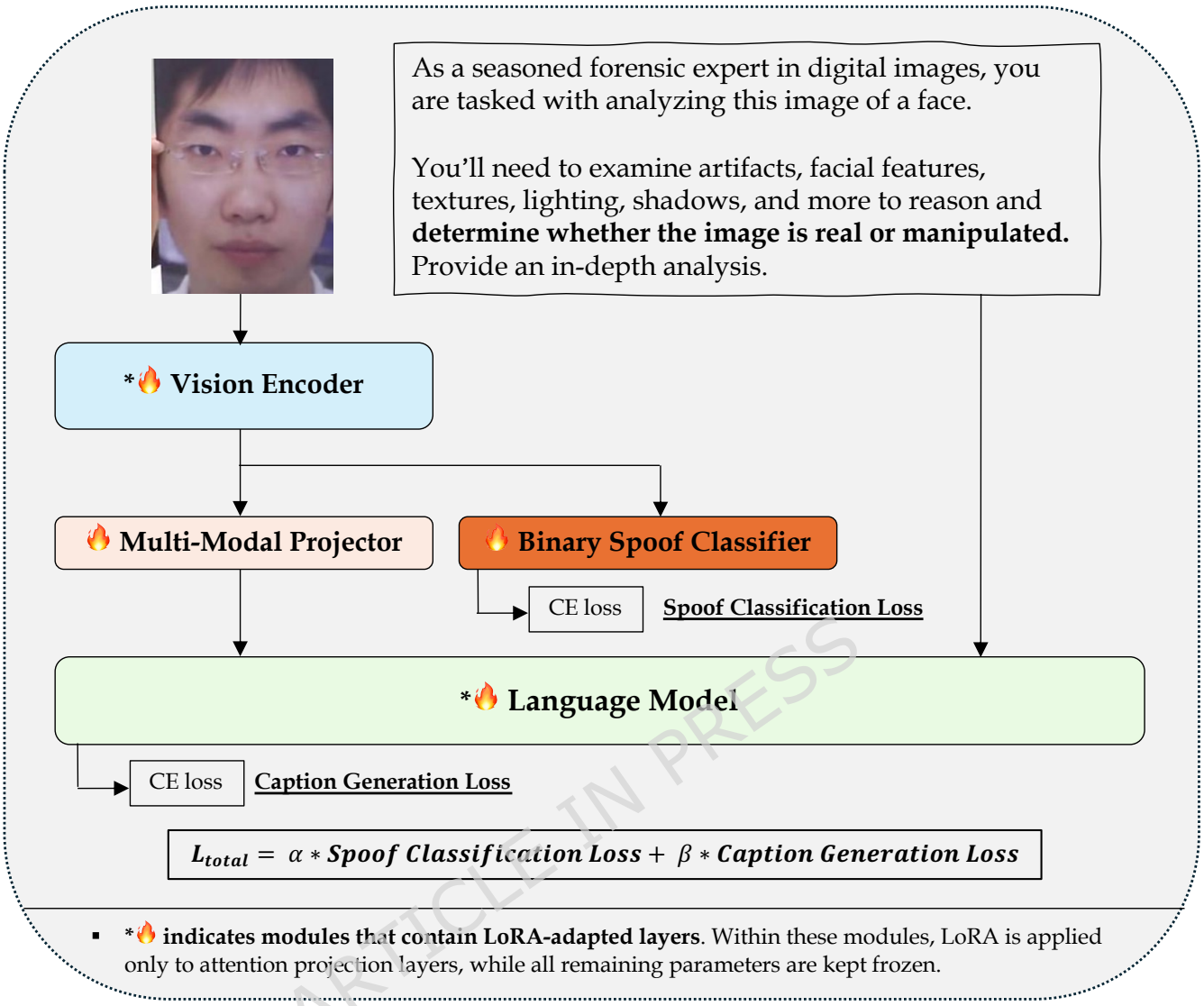

Figure 5. Architecture of the proposed explanation-based face anti-spoofing framework. The vision encoder extracts visual representations from the input image, which are fused with textual instructions through a multi-modal projector. The model is trained using a dual-objective loss that combines binary spoof classification and explanation generation.

between accepting spoof images as genuine and rejecting real images as fake. This metric is particularly suitable for inter-dataset evaluations, where the model is tested on datasets it has not seen during training, and is used to assess generalization performance across domains. In addition to HTER, we also report binary classification accuracy.

Since our model outputs textual explanations rather than direct class labels, we additionally utilize GPT-4o-mini to infer the predicted class from each generated response. Specifically, we prompt GPT-4o-mini to determine whether the explanation implies a "live" or "spoof" judgment, and use this as the model's predicted class. This indirect classification procedure allows us to compute HTER and accuracy based on textual outputs, enabling fair comparison with prior methods.

In addition, we report confusion matrices for representative cross-dataset settings to provide a more detailed breakdown of false acceptance and false rejection behaviors beyond aggregate HTER. We note that ROC/AUC analysis typically requires a continuous decision score (e.g., logits or calibrated probabilities), whereas our evaluation pipeline yields discrete binary predictions inferred from textual explanations via GPT-4o-mini. Therefore, we focus on confusion-matrix-based characterization and HTER/accuracy, which are well-defined under binary decisions and commonly adopted in cross-dataset FAS evaluation. The corresponding confusion matrices are provided in the Supplementary Information for completeness and detailed error analysis.

Metric	Method	OCI-M	MOI-C	MOC-I	MIC-O	avg.
Accuracy ↑	vanilla	0.87	0.90	0.89	0.69	0.84
	vanilla+reasoning	**0.92**	**0.94**	**0.90**	**0.71**	**0.87**
HTER ↓	vanilla	0.11	0.19	0.26	0.20	0.19
	vanilla+reasoning	0.11	**0.12**	**0.22**	**0.18**	**0.16**

Table 1. Comparison of binary classification accuracy and between vanilla and vanilla+reasoning models across MCIO leave-one-out protocols. Each protocol trains on three datasets and tests on the remaining one; for example, OCI-M trains on OULU[9], CASIA[7], and Idiap[8], and tests on MSU[6]. The best result for each column is shown in **bold**, and identical values are indicated with underlines.

Implementation Details

We fine-tune the model using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 0.01. Training is performed with a per-device batch size of 4 and a gradient accumulation step of 4, resulting in an effective batch size of 64, for a total of 2,500 training steps. To improve training stability and memory efficiency, we apply gradient checkpointing and use BFloat16 precision. LoRA-based parameter-efficient fine-tuning is applied to selected attention and output modules, with a LoRA rank of 32.

The training objective is defined as a weighted combination of a spoof classification loss and a captioning loss, with weights $\alpha = 0.7$ and $\beta = 0.3$. The classification loss $\mathscr{L}_{\text{cls}}$ is implemented as the standard cross-entropy loss over binary live/spoof labels, while the captioning loss $\mathscr{L}_{\text{cap}}$ corresponds to the standard cross-entropy language modeling loss computed over the generated explanation tokens.

The choice of loss weighting reflects the nature of face anti-spoofing as a classification-driven task. Accordingly, a larger relative weight is assigned to the spoof detection objective to prioritize classification stability, while the captioning loss is used as an auxiliary supervisory signal. In practice, we explored a small set of candidate (α, β) configurations under this principle and selected $\alpha = 0.7$ and $\beta = 0.3$ as a balanced setting.

All experiments are conducted on four NVIDIA RTX A6000 GPUs, each equipped with 48GB of VRAM. For clarity, GPT-4o was used to generate the caption datasets during benchmark construction, whereas GPT-4o-mini was used only at evaluation time to infer binary live/spoof predictions from the model-generated textual explanations.

Comparison to the State-of-the-art Methods

MCIO leave-one-out

To evaluate the domain generalization capability of our model, we adopt the widely-used MCIO leave-one-out protocol, where the model is trained on three out of four datasets (MSU-MFSD, CASIA-FASD, Replay-Attack, and OULU-NPU) and tested on the remaining one. This protocol is considered a standard setting for assessing a model's robustness in cross-dataset scenarios, which closely resemble real-world deployment conditions.

We compare two training configurations: the vanilla setting and the vanilla+reasoning setting. The vanilla training samples are constructed as image-text pairs that provide binary classification labels (live/spoof) and short descriptive phrases explaining the classification at a high level. In contrast, the vanilla+reasoning samples include not only binary judgments but also detailed, step-by-step reasoning that explicitly refers to spoofing cues observed in the image. Crucially, the number of training samples remains the same in both settings; half of the captions in the vanilla+reasoning set are replaced with reasoning-augmented prompts, while the rest remain in vanilla format.

As shown in Table 1, the vanilla+reasoning model consistently outperforms the vanilla baseline across all MCIO test splits, achieving both higher accuracy and lower HTER. These improvements demonstrate that incorporating explicit reasoning facilitates more robust decision-making, allowing the model to better generalize to unseen domains without additional data. The results clearly demonstrate that incorporating reasoning-based data into training improves the model's ability to detect spoof-specific visual cues. Furthermore, this approach enhances domain generalization performance across unseen test distributions, validating the effectiveness of reasoning supervision in both task accuracy and robustness.

To further illustrate the qualitative benefits of reasoning-based supervision, Figure6 shows an example generated by the model trained on MSU[6], OULU[9], and Idiap[8] (MOI) and tested on CASIA[7] (C). The model systematically analyzes the input image through six structured stages, identifying spoof-indicative patterns—such as pixelation, blurred facial features, and lighting inconsistencies—and concludes that the sample is the result of a replay attack. This demonstrates the model's ability to provide transparent and forensically aligned justifications even in cross-dataset evaluation settings.

To provide a clearer structural comparison, Figure 7 presents side-by-side examples for both a live and a spoof sample under the same MCIO leave-one-out setting. For each sample, the input image is displayed at the top, followed by the vanilla-based

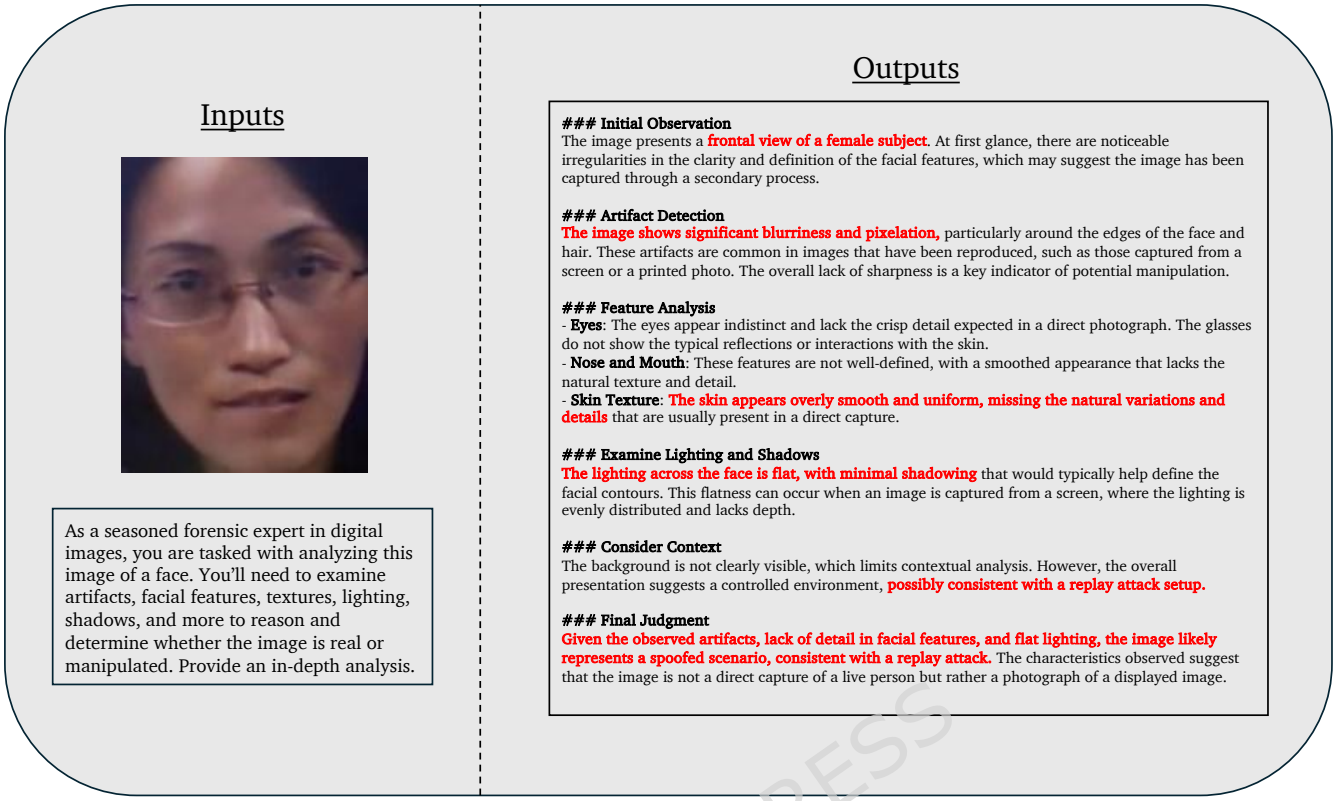

Figure 6. A reasoning-based explanation generated by the model trained under the MCIO leave-one-out protocol, specifically using the MOI (MSU[6], OULU[9], Idiap[8]) training set and evaluated on C(CASIA[7]). The input image (left) is analyzed through six structured reasoning stages (right)—initial observation, artifact detection, feature analysis, lighting/shadow inspection, contextual assessment, and final judgment. The explanation highlights spoof-specific cues such as pixelation, uneven lighting, and lack of skin detail, leading to the conclusion of a replay attack.

output in the middle and the reasoning-based output at the bottom, allowing direct comparison of explanation structure under identical visual input.

In the live case (Figure 7(a)), the vanilla model produces a descriptive judgment focusing on surface-level properties such as skin texture and lighting consistency. In contrast, the reasoning model explicitly organizes its assessment into analytical stages, including artifact detection and feature analysis, and justifies the live decision by emphasizing the absence of spoof-indicative irregularities. This structured reasoning provides a more explicit justification of authenticity.

In the spoof case (Figure 7(b)), both models identify suspicious characteristics. However, the vanilla output presents them in a relatively compact descriptive format, whereas the reasoning model decomposes the analysis into sequential stages and explicitly connects observed artifacts—such as blurriness, texture uniformity, and lighting flatness—to the final judgment of a print attack. This explicit decomposition clarifies how intermediate evidence contributes to the decision.

These examples suggest that reasoning-based supervision primarily influences the internal organization and evidential articulation of explanations. Even when the final binary outcome is identical, the reasoning model exhibits more structured cue-to-judgment mapping, which may contribute to the improved cross-dataset robustness observed in Table 1.

MCIO-Subset-to-SiW-Mv2

To further evaluate the generalization ability of our method to unseen attack types, we conduct experiments by testing on the SiW-Mv2[32] dataset, which includes 14 diverse spoofing attack types. In contrast, the MCIO datasets only contain two types of attacks (print and replay), making this a challenging cross-domain evaluation setting.

For this experiment, we reuse the four models trained in the MCIO leave-one-out protocol (Table 1), each trained on a different combination of three MCIO datasets. These models are evaluated on SiW-Mv2 without any additional fine-tuning. This setting allows us to assess how well the models can generalize to attack types that were never seen during training. We compare our models with two state-of-the-art FAS generalization methods: SA-FAS[15] and DiVT-M[19], both of which are designed specifically for domain generalization in FAS.

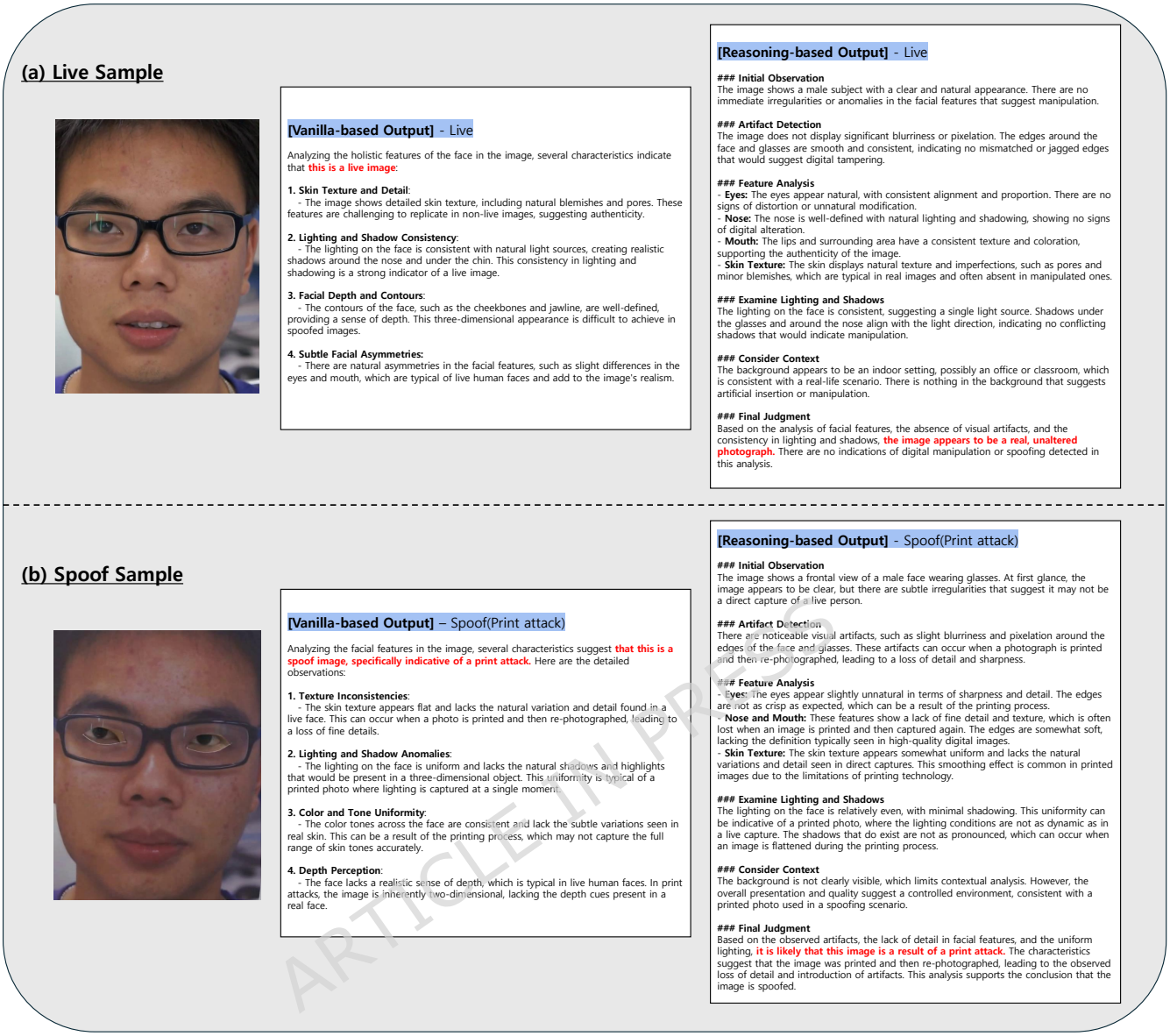

Figure 7. Qualitative comparison between the vanilla and vanilla+reasoning models under the MCIO leave-one-out protocol. Each column corresponds to a single input sample, with the input image shown at the top, the vanilla-based output in the middle, and the reasoning-based output at the bottom. (a) Live sample. The reasoning model performs stage-wise analysis and explicitly justifies the authenticity of the image by examining the absence of spoof-related artifacts. (b) Spoof sample. The reasoning model decomposes the decision into structured analytical steps, identifying visual artifacts and linking them to the final spoof judgment. Compared to the vanilla model, which provides a relatively concise descriptive assessment, the reasoning model exhibits more explicit cue-to-decision grounding.

Table 2 shows the results when the model is trained on OULU, CASIA, and Idiap (OCI) and tested on SiW-Mv2. Our vanilla+reasoning model achieves the highest accuracy of 0.82 and lowest HTER of 0.18, outperforming both the vanilla baseline (Accuracy: 0.76, HTER: 0.25) and the other comparison methods. This result suggests that the added reasoning supervision enables the model to better generalize to spoof types it has never seen, such as partial covering attacks. By guiding the model to focus on spoof-related visual evidence during training, reasoning captions compensate for the limited diversity in attack types, leading to improved robustness in cross-domain settings.

Table 3 presents the evaluation results when the model is trained on MSU, OULU, and Idiap (MOI) and tested on SiW-Mv2. In this setting, the vanilla model achieves the highest overall detection performance among all configurations, recording an average accuracy of 0.78 and HTER of 0.21. The vanilla+reasoning model shows slightly lower accuracy of 0.75 but maintains

Method	Metric	Covering (Partial)				Makeup			3D Attack (Mask)					2D Attack		Average
		Fun.	Eye.	Mou.	Pap.	Ob.	Im.	Cos.	Imp.(Full)	Sil.	Tra.	Pap.	Man.	Rep.	Print	Avg.
Vanilla	Accuracy ↑	0.67	0.81	**0.94**	0.71	0.50	0.76	0.52	0.90	0.65	0.74	0.90	0.66	0.86	0.95	0.76
	HTER ↓	0.33	0.19	**0.06**	0.29	0.50	0.24	0.48	0.10	0.35	0.26	0.10	0.34	0.14	0.05	0.25
Vanilla+Reasoning	Accuracy ↑	**0.79**	**0.91**	0.91	**0.83**	0.65	**0.82**	0.60	**0.91**	0.72	0.86	0.95	0.64	**0.89**	0.95	**0.82**
	HTER ↓	**0.21**	**0.09**	0.09	**0.17**	0.35	**0.18**	0.40	**0.09**	0.28	0.14	0.05	0.36	**0.11**	0.05	**0.18**
SA-FAS[15]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	0.43	0.50	0.34	0.38	0.35	0.21	0.36	0.15	0.33	0.27	**0.04**	0.25	0.19	0.12	0.28
DiVT-M[19]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	0.33	0.36	0.28	0.24	0.46	0.28	**0.35**	0.35	**0.19**	**0.11**	0.06	0.25	0.20	0.12	0.24

Table 2. Performance comparison across spoof types for the OCI-to-SiWMv2 protocol, where the model is trained on OULU[9], CASIA[7], and Idiap[8], and evaluated on SiW-Mv2[32] to assess generalization to unseen spoof types. The table reports binary classification accuracy and HTER across various attack categories. The best result for each column is shown in **bold**, and identical values are indicated with underlines.

Method	Metric	Covering (Partial)				Makeup			3D Attack (Mask)					2D Attack		Average
		Fun.	Eye.	Mou.	Pap.	Ob.	Im.	Cos.	Imp.(Full)	Sil.	Tra.	Pap.	Man.	Rep.	Print	Avg.
Vanilla	Accuracy ↑	0.68	0.83	**0.93**	0.78	0.52	0.87	0.57	**0.90**	0.62	0.81	0.91	0.76	0.82	0.95	**0.78**
	HTER ↓	0.32	0.17	**0.07**	0.22	0.48	0.13	0.43	**0.10**	0.38	0.19	0.09	0.24	0.18	0.05	**0.21**
Vanilla+Reasoning	Accuracy ↑	0.69	**0.88**	0.92	0.70	0.55	0.84	0.55	0.80	0.56	0.63	**0.94**	0.55	**0.90**	0.95	0.75
	HTER ↓	0.31	**0.12**	0.08	0.30	0.45	0.16	0.45	0.20	0.44	0.37	**0.06**	0.45	**0.10**	0.05	0.25
SA-FAS[15]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	0.35	0.36	0.30	**0.19**	**0.42**	**0.11**	**0.33**	0.19	0.33	0.24	0.04	0.28	0.22	0.15	0.25
DiVT-M[19]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	**0.29**	0.44	0.31	0.21	0.48	0.15	0.44	0.18	**0.23**	**0.18**	0.07	0.28	0.19	0.14	0.26

Table 3. Performance comparison across spoof types for the MOI-to-SiWMv2 protocol, where the model is trained on MSU[6], OULU[9], and Idiap[8], and evaluated on SiW-Mv2[32] to assess generalization to unseen spoof types. The table reports binary classification accuracy and HTER across various attack categories. The best result for each column is shown in **bold**, and identical values are indicated with underlines.

a competitive HTER of 0.25. This suggests that while the reasoning captions in the training data guided the model to focus on specific spoofing cues (e.g., print texture), these cues were absent in entirely different attack types like the '3D Attack (MASK)' in SiW-Mv2, potentially leading the model to make incorrect judgments. Both SA-FAS and DiVT-M record higher HTERs (0.25 and 0.26, respectively), despite being dedicated domain generalization methods.

We interpret this case as evidence that reasoning-structured supervision can introduce an inductive bias toward a particular set of spoof cues emphasized in the training distribution. When the target domain contains attack types whose discriminative evidence differs substantially from those cues, the imposed reasoning pattern may over-prioritize mismatched evidence and reduce detection accuracy. This result suggests that reasoning supervision is not universally beneficial, but rather acts as a controllable training signal whose effectiveness depends on the overlap between training cues and test-time attack characteristics.

Table 4 presents the results when the model is trained on MSU, CASIA, and OULU (MOC) and evaluated on SiW-Mv2. The vanilla and vanilla+reasoning models demonstrated strong, comparable performance, achieving accuracies of 0.81 and 0.80 with HTERs of 0.19 and 0.20, respectively. Notably, both models substantially surpassed the HTERs of SA-FAS (0.26) and DiVT-M (0.23). This suggests that our caption-based training enables stronger generalization than existing state-of-the-art methods, even without explicit reasoning supervision.

Table 5 shows the results when the model is trained on MSU, Idiap, and CASIA (MIC) and tested on SiW-Mv2. In this configuration, the vanilla+reasoning model achieves the best overall performance, recording the highest average accuracy (0.88) and the lowest HTER (0.12) among all compared methods. The improvements are especially consistent across covering, makeup, and 3D mask attacks, where the model demonstrates robust detection even on attack types unseen during training. This result confirms that reasoning supervision not only enhances spoof detection but also contributes significantly to domain generalization across diverse attack types.

These results collectively demonstrate that encouraging the model to generate grounds for its decisions—whether in a simple or reasoning-based format—can implicitly enhance generalization in cross-domain face anti-spoofing scenarios.

MCIO to SiW-Mv2

Table 6 presents the results when the model is trained on the full MCIO set—MSU, CASIA, Idiap, and OULU—and evaluated on SiW-Mv2. While both the vanilla and vanilla+reasoning models attained an identical average accuracy of 0.80, the vanilla+reasoning model delivered a better HTER of 0.20, compared to the vanilla model's 0.22. This superior performance places our vanilla+reasoning model ahead of strong competitors like SA-FAS (HTER: 0.27) and DiVT-M (HTER: 0.21). This

Method	Metric	Covering (Partial)				Makeup			3D Attack (Mask)					2D Attack		Average
		Fun.	Eye.	Mou.	Pap.	Ob.	Im.	Cos.	Imp.(Full)	Sil.	Tra.	Pap.	Man.	Rep.	Print	Avg.
Vanilla	Accuracy ↑	0.67	0.79	0.95	0.74	0.52	0.84	0.58	**0.95**	0.76	0.88	0.97	**0.86**	0.90	0.95	**0.81**
	HTER ↓	0.33	0.21	0.05	0.26	0.48	0.16	0.42	**0.05**	0.24	0.12	0.03	**0.14**	0.10	0.05	**0.19**
Vanilla+Reasoning	Accuracy ↑	**0.69**	**0.84**	0.95	0.74	0.59	0.77	0.59	0.93	0.74	**0.89**	**0.98**	0.68	**0.91**	**0.96**	0.80
	HTER ↓	**0.31**	**0.16**	0.05	0.26	0.41	0.23	0.41	0.07	0.26	**0.11**	**0.02**	0.31	**0.09**	**0.04**	0.20
SA-FAS[15]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	0.46	0.36	0.27	0.37	**0.26**	**0.12**	**0.31**	0.19	0.37	0.26	0.06	0.27	0.22	0.10	0.26
DiVT-M[19]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	0.38	0.35	0.23	**0.23**	0.35	0.17	0.40	0.17	**0.22**	0.12	0.07	0.25	0.17	0.13	0.23

Table 4. Performance comparison across spoof types for the MOC-to-SiWMv2 protocol, where the model is trained on MSU[6], OULU[9], and CASIA[7], and evaluated on SiW-Mv2[32] to assess generalization to unseen spoof types. The table reports binary classification accuracy and HTER across various attack categories. The best result for each column is shown in **bold**, and identical values are indicated with underlines.

Method	Metric	Covering (Partial)				Makeup			3D Attack (Mask)					2D Attack		Average
		Fun.	Eye.	Mou.	Pap.	Ob.	Im.	Cos.	Imp.(Full)	Sil.	Tra.	Pap.	Man.	Rep.	Print	Avg.
Vanilla	Accuracy ↑	0.79	0.86	0.93	0.76	0.59	0.85	0.61	0.94	0.88	**0.93**	0.95	**0.88**	0.85	0.93	0.84
	HTER ↓	0.21	0.14	0.07	0.24	0.41	0.15	0.39	0.06	0.12	**0.07**	0.05	**0.12**	0.15	0.07	0.16
Vanilla+Reasoning	Accuracy ↑	**0.87**	**0.93**	**0.95**	**0.85**	**0.69**	**0.92**	0.69	0.94	**0.93**	0.92	**0.97**	0.82	**0.91**	**0.95**	**0.88**
	HTER ↓	**0.13**	**0.07**	**0.05**	**0.15**	**0.31**	**0.08**	0.31	0.06	**0.07**	0.08	**0.03**	0.18	**0.09**	**0.05**	**0.12**
SA-FAS[15]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	0.32	0.38	0.36	0.27	0.49	0.12	**0.30**	0.25	0.30	0.34	0.18	0.39	0.19	0.13	0.29
DiVT-M[19]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	0.30	0.39	0.27	0.25	0.50	0.20	0.40	0.23	0.31	0.26	0.25	0.40	0.21	0.13	0.29

Table 5. Performance comparison across spoof types for the MIC-to-SiWMv2 protocol, where the model is trained on MSU[6], Idiap[8], and CASIA[7], and evaluated on SiW-Mv2[32] to assess generalization to unseen spoof types. The table reports binary classification accuracy and HTER across various attack categories. The best result for each column is shown in **bold**, and identical values are indicated with underlines.

finding suggests that encouraging the model to generate explanatory cues is a key factor in enhancing its ability to generalize and perform stably across diverse, unseen attack types.

Ablation Studies

Effect of classification loss

Table 7 presents an ablation study on the effect of the classification loss applied to the vision encoder features. As part of our training strategy, we attach a binary classifier to the vision encoder output to directly predict live/spoof labels. This auxiliary classification loss is designed to enhance the model's discriminative capability by encouraging the vision branch to capture spoof-relevant features more explicitly. When the classification loss is removed (*w/o cls loss*), the model achieves an average accuracy of 0.86 and HTER of 0.20 across the four MCIO leave-one-out splits. In contrast, with the classification loss enabled (*w cls loss*), the model records improved results with 0.87 accuracy and a notably lower average HTER of 0.16. The improvement is particularly consistent in MOI-C and MOC-I splits, where the HTER drops significantly. These results demonstrate that incorporating classification supervision at the visual feature level improves spoof detection performance.

Limitations and Future Directions

While our approach trains the model to generate not only spoof predictions but also explanatory text, a key limitation of this work lies in the evaluation of explanation quality. In our experiments, explanations are primarily assessed through their downstream utility for spoof classification, for example by inferring binary decisions from generated text using a GPT-based interpreter. This evaluation captures correctness at a coarse level, but does not directly measure finer-grained properties such as faithfulness to visual evidence, semantic completeness, or usefulness to human operators. We briefly examined standard text-based metrics such as BLEU and ROUGE; however, we found them insufficient for assessing semantic reasoning quality in our setting, particularly under the constrained and structured instruction prompts used for explanation generation. As a result, these metrics were not adopted in our evaluation.

Consequently, the quality of generated explanations is not directly evaluated in terms of their alignment with visual cues or logical coherence, but only through a minimal functional criterion—whether the explanation implies the correct live/spoof judgment. While this design choice is sufficient for analyzing the effect of explanation-based supervision on classification and generalization, it leaves open the question of how faithful or informative the generated rationales are from a human perspective.

Method	Metric	Covering (Partial)				Makeup			3D Attack (Mask)					2D Attack		Average
		Fun.	Eye.	Mou.	Pap.	Ob.	Im.	Cos.	Imp.(Full)	Sil.	Tra.	Pap.	Man.	Rep.	Print	Avg.
Vanilla	Accuracy ↑	0.72	0.78	**0.95**	0.76	0.51	0.81	0.57	**0.95**	0.78	0.83	0.92	**0.78**	0.92	**0.97**	0.80
	HTER ↓	0.28	0.22	**0.05**	0.24	0.49	0.19	0.43	**0.05**	0.22	0.17	0.08	**0.22**	0.08	**0.03**	0.20
Vanilla+Reasoning	Accuracy ↑	**0.73**	**0.85**	0.94	0.76	0.55	0.84	0.58	0.93	**0.80**	0.83	0.96	0.61	**0.93**	0.96	0.80
	HTER ↓	**0.27**	**0.15**	0.06	0.24	0.45	0.16	0.42	0.07	**0.20**	0.17	0.04	0.39	**0.07**	0.04	0.20
SA-FAS[15]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	0.41	0.48	0.39	0.27	**0.44**	**0.13**	**0.30**	0.15	0.29	0.27	**0.03**	0.23	0.20	0.15	0.27
DiVT-M[19]	Accuracy ↑	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	HTER ↓	0.30	0.32	0.20	0.22	0.31	0.14	0.34	0.14	0.23	**0.14**	0.08	0.23	0.18	0.09	0.21

Table 6. Performance comparison across spoof types for the MICO-to-SiWMv2 protocol, where the model is trained on MSU[6], Idiap[8], CASIA[7], and OULU[9], and evaluated on SiW-Mv2[32] to assess generalization to unseen spoof types. The table reports binary classification accuracy and HTER across various attack categories. The best result for each column is shown in **bold**, and identical values are indicated with underlines.

	Metric	OCI-M	MOI-C	MOC-I	MIC-O	avg.
w/o cls loss	Accuracy ↑	**0.93**	0.87	0.87	**0.75**	0.86
	HTER ↓	0.11	0.25	0.26	0.18	0.20
w cls loss	Accuracy ↑	0.92	**0.94**	**0.90**	0.71	**0.87**
	HTER ↓	0.11	**0.12**	**0.22**	0.18	**0.16**

Table 7. Ablation study on the effect of classification loss across MCIO leave-one-out protocols. The model is trained with or without an auxiliary classification head applied to the vision encoder features. The best result for each column is shown in **bold**, and identical values are indicated with underlines.

Another related limitation concerns the reliance on large language models for both caption generation and explanation interpretation. Prompt bias and inherent limitations of models such as GPT-4o may influence the quality of automatically generated explanations. In this work, we do not attempt to eliminate such biases entirely, but instead aim to control their impact through fixed and structured instruction templates that constrain free-form generation and promote consistent reasoning patterns.

Nevertheless, large language models may emphasize spurious or frequently occurring visual attributes during caption generation, which could inadvertently bias the supervision signal and affect model fairness or generalization across demographic groups and acquisition conditions. A systematic analysis of how such language-model-induced biases propagate into visual representation learning remains an open challenge.

In addition, using GPT-4o-mini to infer binary predictions from generated explanations introduces a degree of circular dependency on large language models during evaluation. Although this interpreter is employed only as a lightweight consistency checker rather than as a quality evaluator, it does not constitute a fully independent assessment mechanism.

Developing rigorous metrics for evaluating explanation quality, analyzing robustness across language models, and incorporating human-centered evaluation protocols are important directions for future work.

Conclusion

In this work, we investigated explainable face anti-spoofing from a supervision-centric perspective, treating explanations not as post-hoc outputs but as explicit training signals that shape model behavior. By constructing an explanation-augmented benchmark and controlling for backbone architecture and data scale, we systematically isolated the effect of reasoning-structured supervision on spoof detection and domain generalization. Quantitatively, the vanilla+reasoning model achieved consistent improvements across MCIO leave-one-out protocols, reducing the average HTER from 0.19 to 0.16 while increasing accuracy from 0.84 to 0.87. In cross-domain evaluation on SiW-Mv2, the reasoning-based supervision demonstrated strong robustness, achieving the lowest HTER (0.12) in the MIC-to-SiW-Mv2 setting and competitive or superior performance compared to state-of-the-art domain generalization methods such as SA-FAS and DiVT-M. These results indicate that structured reasoning supervision can enhance cross-dataset generalization without increasing model capacity or modifying backbone architecture. Importantly, our analysis also reveals that reasoning supervision introduces an inductive bias toward specific spoof cues emphasized during training. While this bias often improves robustness, it may degrade performance when unseen attack types exhibit substantially different discriminative evidence. This finding highlights that explanation-based supervision is not universally beneficial, but rather a controllable mechanism that influences representation learning and generalization behavior. Despite these contributions, several limitations remain. The quality and faithfulness of generated explanations are evaluated

only through their downstream classification implications, rather than through human-centered or fine-grained semantic metrics. In addition, the reliance on large language models for caption generation and evaluation may introduce bias. Future work should focus on developing rigorous explanation-quality metrics, investigating bias propagation from language supervision to visual representation learning, and conducting human-in-the-loop evaluation to better assess interpretability in real-world deployment scenarios. Overall, this study reframes explainability in face anti-spoofing as a quantitative learning problem, demonstrating that explanation structure can measurably influence domain generalization performance.

References

1. Guo, J. *et al.* Learning meta face recognition in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6163–6172 (2020).
2. Ramachandra, R. & Busch, C. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Comput. Surv.* **50**, 1–37 (2017).
3. Liu, A. *et al.* Cross-ethnicity face anti-spoofing recognition challenge: A review. *IET Biom.* **10**, 24–43 (2021).
4. Jia, S., Guo, G. & Xu, Z. A survey on 3d mask presentation attack detection and countermeasures. *Pattern recognition* **98**, 107032 (2020).
5. Yu, Z. *et al.* Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis Mach. Intell.* **45**, 5609–5631 (2022).
6. Wen, D., Han, H. & Jain, A. K. Face spoof detection with image distortion analysis. *IEEE Transactions on Inf. Forensics Secur.* **10**, 746–761 (2015).
7. Zhang, Z. *et al.* A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, 26–31 (IEEE, 2012).
8. Chingovska, I., Anjos, A. & Marcel, S. 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)*, 1–7 (IEEE, 2012).
9. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X. & Hadid, A. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE international conference on automatic face and gesture recognition (FG 2017)*, 612–618 (IEEE, 2017).
10. Shao, R., Lan, X. & Yuen, P. C. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 11974–11981 (2020).
11. Jia, Y., Zhang, J., Shan, S. & Chen, X. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8484–8493 (2020).
12. Liu, S. *et al.* Feature generation and hypothesis verification for reliable face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 1782–1791 (2022).
13. Wang, Z. *et al.* Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEECVF Conference on Computer Vision and Pattern Recognition*, 4123–4133 (2022).
14. Wang, C.-Y., Lu, Y.-D., Yang, S.-T. & Lai, S.-H. 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*, 20281–20290 (2022).
15. Sun, Y., Liu, Y., Liu, X., Li, Y. & Chu, W.-S. Rethinking Domain Generalization for Face Anti-spoofing: Separability and Alignment. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24563–24574, DOI: 10.1109/CVPR52729.2023.02353 (IEEE, Vancouver, BC, Canada, 2023).
16. Le, B. M. & Woo, S. S. Gradient alignment for cross-domain face anti-spoofing. In *Proceedings of the IEEECVF Conference on Computer Vision and Pattern Recognition*, 188–199 (2024).
17. Huang, H.-P. *et al.* Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *Proceedings of the European Conference on Computer Vision*, 37–54 (Springer, 2022).
18. Cai, R. *et al.* S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens. *IEEE Transactions on Inf. Forensics Secur.* **19**, 8385–8397 (2024).
19. Liao, C.-H. *et al.* Domain Invariant Vision Transformer Learning for Face Anti-spoofing. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6087–6096, DOI: 10.1109/WACV56688.2023.00604 (IEEE, Waikoloa, HI, USA, 2023).

20. Li, D., Chen, G., Wu, X., Yu, Z. & Tan, M. Face anti-spoofing with cross-stage relation enhancement and spoof material perception. *Neural Networks* **175**, 106275 (2024).

21. Kong, C., Zheng, K., Wang, S., Rocha, A. & Li, H. Beyond the pixel world: A novel acoustic-based face anti-spoofing system for smartphones. *IEEE Transactions on Inf. Forensics Secur.* **17**, 3238–3253 (2022).

22. Kong, C. *et al.* Fas: An accurate and robust multimodal mobile face anti-spoofing system. *IEEE Transactions on Dependable Secur. Comput.* **21**, 5650–5666 (2024).

23. Srivatsan, K., Naseer, M. & Nandakumar, K. FLIP: Cross-domain Face Anti-spoofing with Language Guidance. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 19628–19639, DOI: 10.1109/ICCV51070.2023.01803 (IEEE, Paris, France, 2023).

24. Liu, A. *et al.* Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 222–232 (2024).

25. Zhang, H. *et al.* Concept discovery in deep neural networks for explainable face anti-spoofing. *arXiv preprint arXiv:2412.17541* (2024).

26. Singh, R. P., Dash, R. & Mohapatra, R. K. Unveiling explainability in face anti-spoofing: Hybrid feature extraction with xai-guided feature aggregation. *Pattern Recognit.* **169**, 111905 (2026).

27. Zhang, G. *et al.* Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models. *arXiv preprint arXiv:2501.01720* (2025).

28. Wang, H. *et al.* Faceshield: Explainable face anti-spoofing with multimodal large language models. *arXiv preprint arXiv:2505.09415* (2025).

29. Tian, J. *et al.* Admm-based adversarial false data injection attacks against multi-label locational detection. *IEEE Transactions on Dependable Secur. Comput.* (2025).

30. Tian, J. *et al.* Evade: targeted adversarial false data injection attacks for state estimation in smart grid. *IEEE Transactions on Sustain. Comput.* (2024).

31. Kong, C., Wang, S. & Li, H. Digital and physical face attacks: Reviewing and one step further. *APSIPA Transactions on Signal Inf. Process.* **12**, 1–51 (2023).

32. Guo, X., Liu, Y., Jain, A. & Liu, X. Multi-domain learning for updating face anti-spoofing models. In *European Conference on Computer Vision*, 230–249 (Springer, 2022).

Funding

This research was supported by the research fund of Hanbat National University in 2024. This research was supported by the Regional Innovation System & Education(RISE) program through the Daejeon RISE Center, funded by the Ministry of Education(MOE) and the Daejeon Metropolitan City, Republic of Korea (2025-RISE-06-002).

Author contributions statement

Conceptualization and methodology, J.M., K.L. and H.J.; software, data curation and visualization, J.M. and M.K.; validation, formal analysis and investigation, J.M., S.O. and D.K.; writing—original draft preparation, J.M.; writing—review and editing, E.K. and H.J.; supervision, project administration and funding acquisition, E.K. and H.J. All authors have read and agreed to the published version of the manuscript.

Data availability

The four benchmark datasets used in this study (MSU-MFSD, CASIA-FASD, Replay-Attack, and OULU-NPU) are publicly available benchmarks for research purposes. The official datasets download links:

- MSU-MFSD: https://drive.google.com/drive/folders/1nJCPdJ7R67xOiklF1omkfz4yHeJwhQsz
- CASIA-FASD: http://www.cbsr.ia.ac.cn/english/FaceAntiSpoofDatabases.asp
- Replay-Attack: https://www.idiap.ch/en/scientific-research/data/replayattack
- OULU-NPU: https://sites.google.com/site/oulunpudatabase
- SiW-Mv2: https://cvlab.cse.msu.edu/siw-mv2-dataset.html

To support reproducibility, we publicly release only the additional metadata(captions) generated for training our model, excluding all images, at the following Hugging Face repository: https://huggingface.co/datasets/DescriptiveFAS/MCIO_public.

Ethics Declarations

This study was conducted using publicly available datasets and did not involve any new human participants or the collection of private human data; therefore, ethics committee or institutional review board approval was not required.

Competing interests

The authors declare no competing interests.