EVITRACE: STRUCTURED EVIDENCE-GUIDED PROGRESSIVE REASONING FOR INTERPRETABLE FACE ANTI-SPOOFING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing face anti-spoofing methods typically apply a uniform one-pass pipeline to all samples, despite the substantial variation in the saliency and diagnostic difficulty of spoofing cues. While some attacks can be identified from obvious artifacts with minimal reasoning, others exhibit only subtle and hard-to-perceive clues that require more careful evidence aggregation. To address this limitation, we propose **EviTrace**, a structured Evidence-guided progressive reasoning framework for interpretable face anti-spoofing. Specifically, we first introduce **Structured Evidence Slots (SES)**, which leverage structured priors during training to learn explicit responses to diverse spoof-related cues beyond sample-level binary live/spoof supervision. Building on these structured evidence responses, we further develop **Difficulty-Aware Progressive Reasoning (DPR)**, which pools high-confidence slot evidence for coarse screening and selectively routes low-confidence samples for refinement by revisiting the most ambiguous slots and their most relevant local patch regions. Rather than replacing the coarse prediction, DPR progressively corrects it through residual evidence aggregation. Based on this evidence-to-decision process, EviTrace naturally yields an **Interpretable Reasoning Trace (IRT)** that explicitly reveals evidence activation, uncertainty estimation, selective refinement, and final correction as a transparent decision chain. Experiments under a generalization setting show that EviTrace achieves competitive performance while providing a clearer and more transparent reasoning process. Code is available at `https://anonymous.4open.science/r/EviTrace`

1 INTRODUCTION

Face anti-spoofing (FAS) is a fundamental component of reliable face recognition systems, aiming to distinguish bona fide face from presentation attacks. Despite significant progress Liu et al. (2018b); Yu et al. (2020b); Liu et al. (2021b); George & Marcel (2021); Huang et al. (2024) in improving in-domain performance, achieving robust generalization of FAS models to unseen domains remains challenging, due to variations in acquisition devices, environments, presentation media, and attack patterns. This challenge is particularly critical in real-world deployments, where a model must not only remain robust under domain shift but also provide transparent justification for security-sensitive decisions. However, most existing methods Cai et al. (2022); Guo et al. (2022); Keresh & Shamoi (2024); Zhou et al. (2024); Lin et al. (2024); He et al. (2024); Jiang et al. (2025); Liu et al. (2026) still formulate FAS as a flat binary classification problem and process all test samples with a uniform one-pass inference pipeline.

This uniform inference paradigm overlooks the intrinsic heterogeneity of spoofing evidence. In practice, some attacks can be identified from salient artifacts, such as color distortions in printed attacks Mostaani et al. (2020); Liu et al. (2021a), screen artifacts (e.g., moiré patterns) in replay attacks Costa-Pazo et al. (2016); Guo et al. (2022); Li et al. (2018), and abnormal facial geometry in mask-based attacks Yu et al. (2020a); Mostaani et al. (2020), and therefore require little additional reasoning. Others, however, contain only weak, localized, or ambiguous cues, such as planar facial appearance Boulkenafet et al. (2017); Liu et al. (2018b), globally visible paper-like textures Zhang et al. (2020); Wang et al. (2023), reduced contrast Wen et al. (2015); Mostaani et al. (2020), abnormal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

color tone, missing fine skin microstructures Liu et al. (2022; 2018a), whose discriminative value emerges only after more careful evidence aggregation. As a result, the model may lack confidence in exploiting salient cues for easy samples, instead over-analyzing subtle evidence; for hard cases, it may commit prematurely, fail to progressively resolve uncertainty, and is unlikely to refine or correct an initially incorrect decision through iterative reasoning.

A fundamental limitation of conventional binary supervision is that it compresses heterogeneous spoof cues into a global live/fake decision rather than organizing them into explicit evidence components, limiting interpretability and potentially increasing reliance on domain-specific patterns. A desirable FAS framework should therefore learn structured intermediate evidence representations while supporting difficulty-aware progressive reasoning. Applying exhaustive refinement to all samples, however, may introduce redundant computation and disturb reliable predictions. While easy samples with salient spoof cues can often be classified directly, hard samples with weak or ambiguous evidence require more targeted scrutiny. This motivates a progressive reasoning strategy that adapts the depth of analysis to sample difficulty, preserving efficiency on easy cases while enabling focused evidence refinement for hard cases.

To this end, we propose **EviTrace**, a structured evidence-guided progressive reasoning framework for interpretable face anti-spoofing. Our method first introduces **Structured Evidence Slots (SES)**, which leverage structured priors during training to learn explicit responses to different spoof-related cues beyond flat binary supervision. Built on these structured evidence responses, we further develop **Difficulty-Aware Progressive Reasoning (DPR)**, which performs coarse screening by pooling high-confidence evidence and selectively routes low-confidence samples to a refinement stage that revisits the most ambiguous cues and their most relevant local regions, thereby further correcting the coarse decision through residual evidence aggregation. This evidence-guided reasoning process yields an **Interpretable Reasoning Trace (IRT)**, which reveals evidence activation, uncertainty estimation, selective refinement, and final decision correction as a transparent reasoning chain.

2 METHODS

2.1 OVERVIEW

Our goal is to learn structured spoof-related evidence for adaptive coarse-to-fine reasoning while exposing an interpretable reasoning trace. Given a labeled source-domain dataset $\mathcal{D}_s = \{(x_n, y_n, \mathbf{c}_n)\}_{n=1}^{N}$, where x_n is a face sample, $y_n \in \{0, 1\}$ is the live/spoof label, and $\mathbf{c}_n \in \{0, 1\}^M$ denotes the structured annotation over M evidence slots, the model learns to generalize to unlabeled target-domain samples. As shown in Fig. 1, **EviTrace** comprises **Structured Evidence Slots (SES)**, **Difficulty-Aware Progressive Reasoning (DPR)**, and the resulting **Interpretable Reasoning Trace (IRT)**. SES introduces learnable evidence slot tokens into a CLIP-based ViT encoder to extract global, slot-wise, and local patch features, guided by slot supervision and slot-text alignment. DPR consists of two reasoning stages connected by a confidence-based routing component. Stage I aggregates reliable slot evidence for coarse screening, after which the routing component identifies samples requiring further refinement. For routed hard samples, Stage II revisits ambiguous slots and their relevant local regions to refine the coarse prediction through residual score correction. This process naturally yields IRT, exposing evidence activation, routing decisions, selective refinement, and final score updates.

2.2 STRUCTURED EVIDENCE SLOTS (SES)

Existing face anti-spoofing models are typically trained with only sample-level binary live/spoof supervision, which does not explicitly distinguish different spoof-related cues. To address this limitation, we introduce **Structured Evidence Slots (SES)**, which augment the visual encoder with a set of learnable evidence slot tokens and train them to produce explicit responses to predefined spoof-related evidence.

Slot-augmented visual encoding. Given an input face sample x, we use a CLIP-based *ViT Encoder* and insert M learnable *Evidence Slot Tokens* between the CLS token and patch tokens. The resulting token sequence is processed as:

$$\big[\mathbf{f}_{\text{cls}},\, \mathbf{S},\, \mathbf{R}\big] = f_v\big([\mathbf{z}_{\text{cls}};\, \mathbf{Q};\, \mathbf{P}(x)]\big), \tag{1}$$

where $\mathbf{Q} \in \mathbb{R}^{M\times D}$ denotes the learnable evidence slot tokens, $\mathbf{P}(x) \in \mathbb{R}^{P\times D}$ denotes the patch tokens, $\mathbf{f}_{\text{cls}} \in \mathbb{R}^D$ is the global CLS feature, $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_M]^\top \in \mathbb{R}^{M\times D}$ are the slot-wise evidence

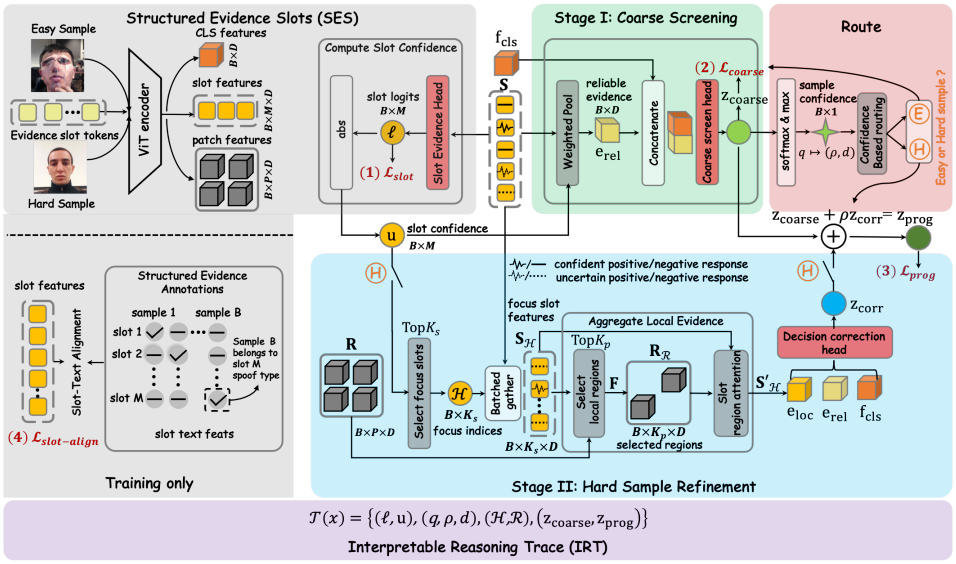


Figure 1: Overview of **EviTrace**. SES learns structured slot-wise evidence from a CLIP-based visual encoder with learnable evidence slot tokens. Built on these slot responses, DPR performs confidence-guided coarse-to-fine reasoning through coarse screening, routing, and selective refinement. The resulting evidence-to-decision process naturally yields an interpretable reasoning trace.

features, and $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_P]^\top \in \mathbb{R}^{P \times D}$ are the local patch features. In this way, the encoder outputs exactly the three feature groups: *CLS features*, *slot features*, and *patch features*.

Slot-wise evidence supervision. Let $\mathbf{c} \in \{0, 1\}^M$ denote the structured evidence annotation vector for sample x, where each element indicates whether the corresponding spoof-related cue is present. We apply a shared *Slot Evidence Head* to all slot features and obtain:

$$\boldsymbol{\ell} = g_{\text{slot}}(\mathbf{S}) \in \mathbb{R}^M, \qquad \mathcal{L}_{\text{slot}} = \text{BCEWithLogits}(\boldsymbol{\ell}, \mathbf{c}), \tag{2}$$

where $\boldsymbol{\ell} = [\ell_1, \ldots, \ell_M]$ are the slot logits. This supervision, marked as (1) $\mathcal{L}_{\text{slot}}$ in Fig. 1, encourages different slots to respond explicitly to different structured spoof-related cues rather than relying only on flat binary supervision.

In addition, the same slot logits are passed through the *Compute Slot Confidence* module, where the absolute-value operation transforms them into the slot confidence used later by DPR for confidence estimation and ambiguous-slot selection.

Slot-text semantic alignment. To further ground slot features in structured semantics, we construct a slot-specific text description for each slot according to its annotation, using either "`name`$_m$`: Detected.`" or "`name`$_m$`: Not Detected.`". A CLIP text encoder is then used to obtain the corresponding text features $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_M]^\top \in \mathbb{R}^{M \times D}$. After ℓ_2 normalization, we compute the slot-text similarity matrix:

$$\mathbf{A} = \alpha \hat{\mathbf{S}} \hat{\mathbf{T}}^\top, \qquad \mathcal{L}_{\text{slot-align}} = \frac{1}{2}\Big[\text{CE}(\mathbf{A}, \mathbf{m}) + \text{CE}(\mathbf{A}^\top, \mathbf{m})\Big], \tag{3}$$

where α is the learned logit scale, $\mathbf{m} = [1, \ldots, M]$ denotes the diagonal matching targets, and $\text{CE}(\cdot, \mathbf{m})$ is the row-wise cross-entropy loss. This objective encourages each visual slot to align with its own semantic description while remaining discriminative from other slot semantics.

SES objective and output. The SES-related training objective is defined as:

$$\mathcal{L}_{\text{SES}} = \lambda_{\text{slot}} \mathcal{L}_{\text{slot}} + \lambda_{\text{align}} \mathcal{L}_{\text{slot-align}}. \tag{4}$$

After training, SES no longer requires structured evidence annotations or the *Slot-Text Alignment* branch during inference. Instead, it outputs the global feature $\mathbf{f}_{\text{cls}}$, the slot-wise evidence features $\mathbf{S}$, and the local patch features $\mathbf{R}$, which are forwarded to the DPR module.

2.3 DIFFICULTY-AWARE PROGRESSIVE REASONING (DPR)

Based on the structured slot-wise evidence learned by SES, we further develop Difficulty-Aware Progressive Reasoning (DPR) for adaptive coarse-to-fine inference. As shown in Fig. 1, DPR consists of two reasoning stages connected by a confidence-based routing component. Stage I forms a coarse prediction from reliable slot-wise evidence. The routing component estimates sample difficulty and determines whether the sample requires further refinement. Stage II is activated only for routed hard samples, where the model revisits ambiguous slots and their relevant local patch regions to predict a residual score correction. The final prediction is obtained by adding the gated correction to the coarse prediction.

Stage I: Coarse evidence screening. Given the global CLS feature $\mathbf{f}_{\text{cls}}$ and the slot-wise evidence features $\mathbf{S} \in \mathbb{R}^{M\times D}$ from SES, we first reuse the *Slot Evidence Head* to compute slot logits

$$\boldsymbol{\ell} = g_{\text{slot}}(\mathbf{S}) \in \mathbb{R}^{M}, \qquad \mathbf{u} = |\boldsymbol{\ell}| \in \mathbb{R}^{M}, \tag{5}$$

where $\mathbf{u}$ denotes the slot confidence, and larger values indicate more decisive slot responses.

We then aggregate reliable evidence from all slots through a confidence-weighted pooling operation:

$$\mathbf{a} = \text{softmax}(\mathbf{u}), \qquad \mathbf{e}_{\text{rel}} = \mathbf{a}^{\top}\mathbf{S}, \tag{6}$$

where $\mathbf{a} \in \mathbb{R}^{M}$ is the normalized slot-weight vector, and $\mathbf{e}_{\text{rel}} \in \mathbb{R}^{D}$ is the resulting reliable evidence feature. This operation corresponds to the *Weighted Pool* module in Fig. 1.

The reliable evidence feature is then concatenated with the global CLS feature and fed into the *Coarse Screen Head* to obtain the coarse prediction:

$$\mathbf{z}_{\text{coarse}} = g_{\text{coarse}}\big([\mathbf{f}_{\text{cls}}; \mathbf{e}_{\text{rel}}]\big), \qquad q = 2\max\big(\text{softmax}(\mathbf{z}_{\text{coarse}})\big) - 1. \tag{7}$$

This stage produces the coarse decision corresponding to the supervision term (2) $\mathcal{L}_{\text{coarse}}$ in Fig. 1. Under the binary live/spoof setting, $q \in [0, 1]$ measures the confidence margin of the coarse prediction, where lower values indicate higher decision ambiguity.

Routing component. The routing component connects Stage I and Stage II by determining whether a sample can be handled directly by coarse screening or requires further refinement. Given the coarse prediction confidence q, it generates a soft correction gate ρ and a hard routing decision d:

$$\rho = \sigma\big(\beta(\tau - q)\big), \qquad d = \mathbb{I}(q < \tau), \tag{8}$$

where τ is the routing threshold and $\beta > 0$ is a learnable scaling factor. The gate ρ controls the strength of residual score correction, while $d \in \{0, 1\}$ determines whether Stage II is activated. Consequently, high-confidence samples are handled directly by Stage I, whereas low-confidence samples are routed to Stage II for further refinement.

Stage II: Hard-sample refinement. For routed hard samples ($d = 1$), DPR revisits the most ambiguous slots and their most relevant local regions. We first select the K_s least-confident slots:

$$\mathcal{H} = \text{TopK}_{\text{smallest}}(\mathbf{u}, K_s), \qquad \mathbf{S}_{\mathcal{H}} = \text{Gather}(\mathbf{S}, \mathcal{H}), \qquad \mathbf{u}_{\mathcal{H}} = \text{Gather}(\mathbf{u}, \mathcal{H}), \tag{9}$$

where $\text{Gather}(\cdot, \cdot)$ denotes sample-wise row selection according to the given indices. Thus, $\mathbf{S}_{\mathcal{H}} \in \mathbb{R}^{K_s\times D}$ and $\mathbf{u}_{\mathcal{H}} \in \mathbb{R}^{K_s}$ are the selected focus-slot features and their confidences, respectively. This step corresponds to the *TopK_s*, *Select Focus Slots*, and *Batched Gather* modules in Fig. 1.

We then retrieve the K_p most relevant local regions according to slot-patch affinity:

$$\mathbf{F} = |\mathbf{S}_{\mathcal{H}}\mathbf{R}^{\top}|, \qquad \mathbf{b} = \max_{\text{row}}(\mathbf{F}), \qquad \mathcal{R} = \text{TopK}_{\text{largest}}(\mathbf{b}, K_p), \qquad \mathbf{R}_{\mathcal{R}} = \text{Gather}(\mathbf{R}, \mathcal{R}), \tag{10}$$

where $\mathbf{b} \in \mathbb{R}^{P}$ collects the maximum affinity of each patch over the selected focus slots, and $\mathbf{R}_{\mathcal{R}} \in \mathbb{R}^{K_p\times D}$ denotes the selected local patch features. This step corresponds to the *Aggregate Local Evidence*, *TopK_p*, and *Select Local Regions* modules in Fig. 1.

The selected patch features are then used to refine the focus slots through the *Slot Region Attention* module:

$$\mathbf{S}'_{\mathcal{H}} = \text{LN}\Big(\text{MHA}(\mathbf{S}_{\mathcal{H}}, \mathbf{R}_{\mathcal{R}}, \mathbf{R}_{\mathcal{R}}) + \mathbf{S}_{\mathcal{H}}\Big), \qquad \tilde{\mathbf{a}} = \text{softmax}(-\mathbf{u}_{\mathcal{H}}), \qquad \mathbf{e}_{\text{loc}} = \tilde{\mathbf{a}}^{\top}\mathbf{S}'_{\mathcal{H}}, \tag{11}$$

where $\mathrm{MHA}(\cdot)$ denotes multi-head attention. The refined slots are pooled into the local evidence feature $\mathbf{e}_{\mathrm{loc}}$, so that refinement emphasizes the least-confident slots while aggregating additional local evidence around them.

Residual score refinement and decision correction. DPR predicts a residual correction conditioned on the global feature, reliable evidence, and refined local evidence. As shown by the *Decision Correction Head*, the correction branch takes $[\mathbf{f}_{\mathrm{cls}}; \mathbf{e}_{\mathrm{rel}}; \mathbf{e}_{\mathrm{loc}}]$ as input and predicts the correction logits:

$$\mathbf{z}_{\mathrm{corr}} = g_{\mathrm{corr}}\big([\mathbf{f}_{\mathrm{cls}}; \mathbf{e}_{\mathrm{rel}}; \mathbf{e}_{\mathrm{loc}}]\big), \qquad \mathbf{z}_{\mathrm{prog}} = \mathbf{z}_{\mathrm{coarse}} + \rho\, \mathbf{z}_{\mathrm{corr}}. \tag{12}$$

Rather than only changing the final binary prediction, the residual term continuously refines the prediction logits of routed samples. It can therefore improve the live/spoof score margin and class separability even when the final discrete label remains unchanged, while sufficiently large corrections may also reverse an incorrect coarse decision. The final progressive prediction $\mathbf{z}_{\mathrm{prog}}$ is obtained by adding the gated residual correction to the coarse prediction, as indicated by the residual path in Fig. 1. When $d = 0$, the refinement branch is skipped and we set $\mathbf{z}_{\mathrm{corr}} = \mathbf{0}$, so easy samples are handled directly by the coarse stage, whereas hard samples receive additional score refinement and, when necessary, decision correction.

DPR objective and trace. We supervise the coarse prediction for all samples, while applying progressive supervision only to routed hard samples:

$$\mathcal{L}_{\mathrm{DPR}} = \mathrm{CE}(\mathbf{z}_{\mathrm{coarse}}, y) + \lambda_{\mathrm{prog}}\, d\, \mathrm{CE}(\mathbf{z}_{\mathrm{prog}}, y), \tag{13}$$

where $d = \mathbb{I}(q < \tau)$ is the routing decision and λ_{prog} controls the supervision strength of the progressive correction branch. In Fig. 1, these two supervision terms are marked as $(2)\,\mathcal{L}_{\mathrm{coarse}}$ and $(3)\,\mathcal{L}_{\mathrm{prog}}$, respectively. In this way, easy samples are optimized by the coarse prediction alone, while routed hard samples additionally receive supervision on the progressively corrected prediction. At inference time, DPR exposes the coarse confidence, routing decision, selected ambiguous slots, selected local regions, and final progressive prediction, thereby naturally revealing the evidence-to-decision process as an interpretable reasoning trace.

2.4 Interpretable Reasoning Trace (IRT)

The proposed **Interpretable Reasoning Trace (IRT)** is not an auxiliary explainer, but the explicit inference trace naturally induced by SES and DPR. For an input sample x, we define

$$\mathcal{T}(x) = \{(\boldsymbol{\ell}, \mathbf{u}),\ (q, \rho, d),\ (\mathcal{H}, \mathcal{R}),\ (\mathbf{z}_{\mathrm{coarse}}, \mathbf{z}_{\mathrm{prog}})\}, \tag{14}$$

where $(\boldsymbol{\ell}, \mathbf{u})$ denote slot-wise evidence activation and confidence, (q, ρ, d) denote confidence-based routing, $(\mathcal{H}, \mathcal{R})$ denote the selected ambiguous slots and local regions, and $(\mathbf{z}_{\mathrm{coarse}}, \mathbf{z}_{\mathrm{prog}})$ denote the prediction before and after refinement. Therefore, IRT provides a faithful evidence-to-decision path by construction, rather than a separate post-hoc explanation module.

3 Experiments

3.1 Experimental Setup

Datasets, Protocols and Metrics. Following recent cross-domain FAS studies Zhang et al. (2025a;b), we use CelebA-Spoof Zhang et al. (2020) as the source domain and evaluate the same trained model on 11 unseen datasets, namely CASIA-FASD Zhang et al. (2012), 3DMask Yu et al. (2020a), HKBU-MARs V1+ Liu et al. (2018a), CASIA-HiFiMask Liu et al. (2022), MSU-MFSD Wen et al. (2015), OULU-NPU Boulkenafet et al. (2017), ReplayAttack Chingovska et al. (2012), Rose-Youtu Li et al. (2018), SiW Liu et al. (2018b), SiW-Mv2 Guo et al. (2022), and IDIAP-WMCA George et al. (2019). CelebA-Spoof provides 10 official spoof-type annotations, namely *Photo*, *Poster*, *A4*, *Face Mask*, *Upper Body Mask*, *Region Mask*, *PC*, *Pad*, *Phone*, and *3D Mask*. These annotations are used only during source-domain training to provide structured supervision for SES, while no spoof-type annotations are required for any unseen target dataset or during inference.

We use the held-out development split for model selection and threshold determination, and report the final results on the corresponding test split. For each target dataset, the decision threshold is determined from its development split and then fixed when evaluating the test split. We report

Table 1: Comparison under the single-source One-to-Eleven protocol, where all models are trained solely on CelebA-Spoof Zhang et al. (2020) and evaluated on 11 unseen target datasets. Lower HTER and higher AUC are better. For EviTrace, per-dataset results are from Seed 0, while Avg. is reported as mean ± std over three runs (Seeds 0, 1, and 2). Best results within each group are marked in bold.

Methods	CASIA-FASD		3DMask		HKBU-MARs-V1+		CASIA-HiFiMask	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MLLM-based generative methods								
I-FAS (AAAI'25)	1.11	99.88	6.18	98.40	18.64	88.77	28.23	77.17
FaceCoT (CVPR'26)	**0.00**	**100.00**	6.21	98.73	6.96	99.41	28.68	79.74
TAR-FAS (CVPR'26)	**0.00**	**100.00**	**2.09**	**99.65**	**3.48**	**99.71**	**17.97**	**90.23**
Discriminative and vision-language alignment methods								
ViTAF (ECCV'22)	3.11	99.48	32.44	75.20	49.29	57.28	37.30	67.10
ViT-B (ICLR'21)	**0.70**	99.86	24.89	84.26	45.08	62.28	37.33	67.35
ViT-L (ICLR'21)	0.93	**99.95**	23.54	84.22	33.33	73.88	32.81	72.58
FLIP (ICCV'23)	4.88	98.48	**8.83**	**96.93**	17.25	88.31	**28.32**	**76.50**
EviTrace (Ours)	1.33	99.87	18.67	88.06	**8.26**	**98.04**	33.80	72.01

Methods	MSU-MFSD		OULU-NPU		Replay-Attack		Rose-Youtu	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MLLM-based generative methods								
I-FAS (AAAI'25)	**5.63**	**98.73**	14.86	91.68	9.15	95.12	5.52	98.48
FaceCoT (CVPR'26)	8.33	98.29	**9.47**	**96.01**	9.87	**96.84**	**2.45**	**99.66**
TAR-FAS (CVPR'26)	5.71	98.29	14.45	92.63	**5.75**	96.69	3.61	99.16
Discriminative and vision-language alignment methods								
ViTAF (ECCV'22)	12.86	93.14	26.73	81.28	12.38	95.73	69.34	74.22
ViT-B (ICLR'21)	16.67	89.89	28.53	78.59	24.80	84.47	82.69	63.22
ViT-L (ICLR'21)	20.87	85.65	29.42	78.07	16.58	92.00	80.47	71.69
FLIP (ICCV'23)	19.37	89.95	20.57	87.30	25.67	81.37	80.73	73.60
EviTrace (Ours)	**5.71**	**98.49**	**5.28**	**98.78**	**2.14**	**99.74**	**5.33**	**98.77**

Methods	SiW		SiW-M-V2		WMCA		Avg.	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MLLM-based generative methods								
I-FAS (AAAI'25)	4.02	98.34	10.89	95.02	20.07	89.17	11.30	93.71
FaceCoT (CVPR'26)	**0.79**	**99.98**	**9.85**	**96.29**	11.51	95.12	8.56	96.37
TAR-FAS (CVPR'26)	8.35	96.93	11.72	94.75	**9.78**	**95.29**	**7.54**	**96.67**
Discriminative and vision-language alignment methods								
ViTAF (ECCV'22)	14.74	92.51	26.72	80.70	29.88	77.14	23.85	82.82
ViT-B (ICLR'21)	9.13	96.24	22.60	84.59	34.72	73.10	23.48	82.98
ViT-L (ICLR'21)	9.03	96.56	17.26	90.37	34.39	75.13	21.08	85.61
FLIP (ICCV'23)	11.01	95.40	25.95	80.78	19.36	88.73	18.73	87.90
EviTrace (Ours)	**1.69**	**99.88**	**17.00**	**91.29**	**10.55**	**96.10**	9.98 ± 0.07	94.58 ± 0.06

HTER and AUC. HTER is computed as the average of the False Rejection Rate (FRR) and False Acceptance Rate (FAR), while AUC measures threshold-independent score ranking; lower HTER and higher AUC are better.

3.2 Implementation Details

Unless otherwise specified, EviTrace is instantiated on the CLIP visual backbone **ViT-B/16**. We insert $M = 10$ learnable evidence slot tokens, corresponding to the 10 spoof-related evidence slots defined in CelebA-Spoof Zhang et al. (2020). The CLIP text encoder is used only during training for the slot-text alignment objective $\mathcal{L}_{\text{slot-align}}$, while optimization is applied to the visual encoder and the proposed reasoning modules. The input resolution is 224×224. During training, we adopt spoof-type-aware balanced sampling to form fake-live paired mini-batches of size 60, with a balanced live/spoof ratio and near-uniform coverage of the 10 spoof types. We optimize the model using SGD with an initial learning rate of 5×10^{-6} for all 15 epochs under a cosine schedule. Newly introduced modules use a $12\times$ larger learning rate than the visual backbone. Unless otherwise specified, we set $\lambda_{\text{slot}} = 0.15$, $\lambda_{\text{align}} = 0.01$, $\lambda_{\text{prog}} = 0.65$, $K_s = 4$, $K_p = 49$, and $\tau = 0.45$. The slot-region interaction uses 8 attention heads, and the routing scale factor β is learnable and initialized to 8.0.

3.3 Comparison with State-of-the-Art Methods

In Tab. 1, MLLM-based generative methods demonstrate strong cross-domain generalization. I-FAS Zhang et al. (2025a), FaceCoT Zhang et al. (2025b), and TAR-FAS Zhang et al. (2026) achieve average HTERs of 11.30%, 8.56%, and 7.54%, respectively, with corresponding AUCs of 93.71%, 96.37%, and 96.67%. Although EviTrace does not surpass the latter two methods, it achieves an

Table 2: Ablation of EviTrace components and spoof-type supervision. The plain ViT-B control evaluates the benefit of 10-way supervision without CLIP pre-training, while EviTrace ablations use a CLIP-pretrained visual backbone. SES combines evidence slot tokens, slot supervision, and slot-text alignment, while DPR performs difficulty-aware progressive refinement.

Variant	Evidence Slot Tokens	$\mathcal{L}_{\text{slot}}$	$\mathcal{L}_{\text{slot-align}}$	SES	DPR	Avg. HTER (%)	Avg. AUC (%)
Plain ViT-B (binary only)						23.48	82.98
+ 10-way type supervision						17.82	87.89
EviTrace binary baseline						21.01	85.68
+ Evidence Slot Tokens	✓					21.31	85.43
+ Slot supervision	✓	✓				15.13	89.43
+ Slot-text alignment (SES)	✓	✓	✓	✓		13.56	92.15
+ DPR (EviTrace)	✓	✓	✓	✓	✓	**9.98**	**94.64**

average HTER of $9.98 \pm 0.07\%$ and an average AUC of $94.58 \pm 0.06\%$ over three random seeds, outperforming I-FAS by 1.32 and 0.87 percentage points, respectively.

The remaining performance gap may stem from the different reasoning paradigms. Generative MLLMs can leverage broad vision-language knowledge and flexible semantic reasoning to analyze unfamiliar spoofing patterns, whereas EviTrace relies on structured evidence learned from the source domain. Nevertheless, EviTrace achieves competitive generalization within a discriminative framework, without relying on large-scale generative reasoning, while providing an explicit evidence-to-decision process.

At the per-dataset level, EviTrace achieves the best HTER and AUC on 8 of the 11 target datasets among discriminative and vision-language alignment methods, covering diverse print, replay, and mask-based attacks. The dataset-level slot-response profiles in Fig. 3 reveal different attack-type-related activation patterns across unseen target datasets. Furthermore, the counterfactual slot-suppression experiment in Appendix B.1 shows that suppressing selected high-response slots degrades detection performance on representative datasets. These observations support the functional relevance of the learned structured evidence under heterogeneous domain shifts, although they do not fully explain the performance differences across all target datasets. Nevertheless, EviTrace remains less effective on several mask-dominant targets, particularly 3DMask, CASIA-HiFiMask, and SiW-M-V2. The substantial material and realism gaps between source-domain masks and unseen attacks may limit the transferability of learned evidence, highlighting a remaining challenge for structured evidence-based generalization.

3.4 Ablation Studies

Ablation on SES and DPR. Tab. 2 evaluates the contributions of spoof-type supervision and the proposed components through two separate ablation tracks. On a plain ViT-B backbone, introducing 10-way spoof-type supervision improves Avg. HTER/AUC from 23.48/82.98 to 17.82/87.89, demonstrating the substantial benefit of richer annotations. We then incrementally evaluate the proposed components using the CLIP-pretrained visual backbone of EviTrace. Introducing *Evidence Slot Tokens* without structured supervision slightly degrades Avg. HTER from 21.01% to 21.31%, suggesting that additional tokens alone are insufficient. Adding slot-wise supervision $\mathcal{L}_{\text{slot}}$ substantially improves HTER/AUC from 21.31/85.43 to 15.13/89.43, highlighting the benefit of richer supervision and structured evidence learning. Further introducing slot-text alignment improves performance to 13.56/92.15, demonstrating additional gains from semantic grounding. Finally, DPR achieves 9.98% HTER and 94.64% AUC, improving upon SES alone by 3.58% and 2.49%, respectively. These results demonstrate the complementary contributions of SES and DPR through semantic alignment and progressive evidence refinement beyond richer supervision alone.

Ablation on DPR design. Tab. 3 evaluates the key design choices in DPR. Compared with SES only, always refining all samples improves performance, indicating that the refinement branch itself is beneficial. However, random routing remains inferior to confidence-based routing despite routing substantially more samples, showing that the gain does not simply arise from applying refinement more frequently. Replacing relevant patches with random patches also degrades performance, while

Table 3: Ablation on the DPR design. *Routed* denotes the percentage of samples entering the refinement branch. W→C and C→W are computed among routed samples only and measure discrete prediction changes at the operating threshold. These statistics characterize hard-label transitions rather than continuous score refinement.

Variant	Routed (%)	W→C (%)	C→W (%)	Avg. HTER (%)	Avg. AUC (%)
SES only	0.00	–	–	13.56	92.15
Always refine	100.00	1.07	**0.09**	12.98	92.81
Random routing	55.48	0.71	0.16	11.46	93.30
Confidence routing + random patches	17.76	3.35	0.38	10.85	93.95
Confidence routing + most-confident slots	17.63	1.71	0.40	11.05	93.65
EviTrace	16.75	**5.14**	0.12	**9.98**	**94.64**

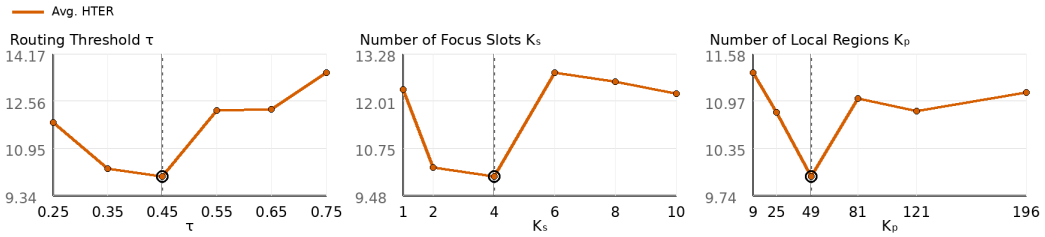


Figure 2: Sensitivity of EviTrace to the main DPR hyper-parameters, measured by average HTER. We vary the routing threshold τ, the number of focus slots K_s, and the number of selected local regions K_p. The best results are achieved at $\tau = 0.45$, $K_s = 4$, and $K_p = 49$, and the overall trends remain stable within a reasonable range.

selecting the most-confident slots is inferior to focusing on ambiguous ones. These results confirm the importance of confidence-based routing and ambiguity-focused local refinement.

The W→C and C→W statistics characterize the discrete prediction changes induced by DPR at a fixed operating point. EviTrace routes only 16.75% of samples and achieves a 5.14% W→C rate with only a 0.12% C→W rate among routed samples, indicating that refinement rarely corrupts reliable coarse predictions. However, these hard-label transitions capture only part of the effect of DPR, since its residual branch directly modifies continuous prediction logits.

Table 4: DPR mechanism analysis.

Variant	HTER (%)	AUC (%)
SES baseline	13.56	92.15
+ Decision-only	12.62	92.15
+ Full DPR	**9.98**	**94.64**

Mechanism analysis of DPR. To disentangle discrete decision correction from continuous-score refinement, Tab. 4 compares three inference settings using the matched Seed-0 SES and full-DPR models. The SES baseline achieves 13.56% Avg. HTER and 92.15% Avg. AUC. Decision-only correction retains the SES scores and threshold while replacing the binary decisions of DPR-routed samples, reducing HTER to 12.62% (a 0.94-point improvement), with AUC unchanged by construction. Full DPR instead uses refined continuous scores and an independently selected threshold, further reducing HTER to 9.98% and increasing AUC to 94.64%. The additional 2.64-point HTER reduction and improved AUC indicate that most of the observed gain is associated with continuous-score refinement rather than discrete label changes alone. These results support the role of DPR in improving score separability through evidence-guided refinement.

Sensitivity analysis. Fig. 2 shows the sensitivity of EviTrace to the main DPR hyper-parameters in terms of average HTER. For the routing threshold τ, performance first improves and then degrades, with the best result achieved at $\tau = 0.45$, suggesting that both insufficient and excessive refinement are suboptimal. A similar trend is observed for the number of focus slots K_s, where using a moderate number of ambiguous slots is more effective than using too few or too many. For the number of selected local regions K_p, the best result is obtained at $K_p = 49$, indicating that moderate local evidence selection provides the best balance between missing useful cues and introducing redundant details. Overall, the curves change smoothly around the optimum, showing that DPR is reasonably

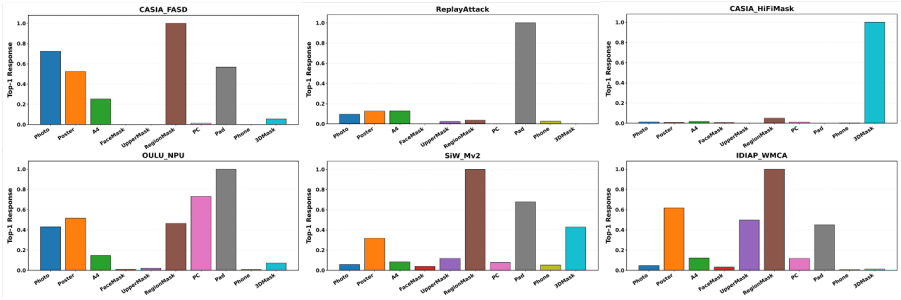


Figure 3: Dataset-level top-1 slot response profiles on spoof samples from six unseen datasets. ReplayAttack and OULU-NPU are dominated by *Pad*, CASIA-HiFiMask by *3DMask*, while the remaining datasets exhibit more distributed responses, suggesting transferable attack related evidence.

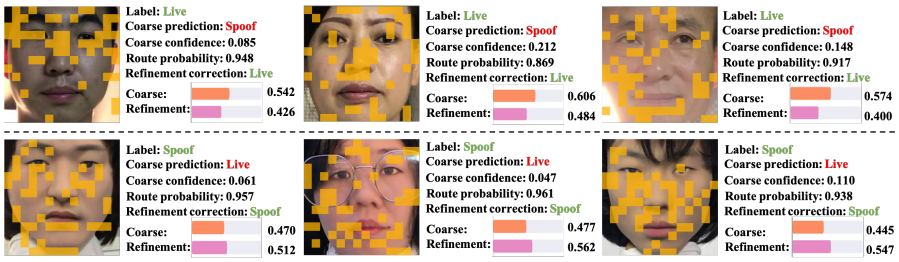


Figure 4: DPR prediction correction examples. Top: three live samples corrected from spoof to live; bottom: three spoof samples corrected from live to spoof. Orange patches mark selected refinement regions, while orange and pink bars show spoof probabilities before and after DPR, respectively.

robust to hyper-parameter choices. A complementary sensitivity analysis of the loss weights is provided in Appendix Tab. 8.

3.5 Reasoning Trace Visualization

Dataset-level slot response analysis. Fig. 3 presents the dataset-level normalized top-1 slot response profiles on spoof samples from six unseen target datasets, revealing distinct dataset-dependent patterns. ReplayAttack and OULU-NPU are dominated by the *Pad* slot, while CASIA-HiFiMask is dominated by the *3DMask* slot. In contrast, CASIA-FASD, SiW-Mv2, and IDIAP-WMCA exhibit more distributed slot responses. These patterns suggest that SES learns attack-type-related evidence representations that remain interpretable across unseen domains, providing a structured basis for DPR. The corresponding analysis of bona fide samples in the Appendix (Fig. 5) shows generally weaker and less concentrated slot activations.

Case Studies of Prediction Correction Fig. 4 shows representative hard samples corrected by DPR. The top row contains live faces that are initially biased toward spoof, while the bottom row contains spoof faces that are initially biased toward live. After routing to refinement, the model revisits discriminative local facial regions, and the spoof probability is consistently shifted toward the ground-truth class. This qualitatively verifies that DPR improves difficult cases by correcting unreliable coarse decisions with localized evidence. Additional failure cases, where refinement changes a correct coarse prediction to an incorrect one, are provided in the Appendix (see Fig. 6).

4 Conclusion

We propose **EviTrace**, a structured evidence-guided progressive reasoning framework for interpretable face anti-spoofing. Through **SES** and **DPR**, it learns explicit spoof-related evidence, selectively refines hard samples, and naturally yields an **IRT**. Experiments on 11 unseen target datasets show strong cross-domain generalization and improved interpretability.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, language editing, LaTeX formatting, and drafting and revising portions of the paper. They also assisted in presenting and interpreting experimental results. All AI-assisted content was reviewed and verified by the authors against the experimental results, source code, and cited literature. Generative AI tools were not used to generate or modify experimental data, execute experiments, or produce mathematical proofs. The research methodology, experimental design, scientific interpretation, and final conclusions were determined and validated by the authors, who take full responsibility for the manuscript.

ETHICS STATEMENT

This work investigates interpretable face anti-spoofing using existing research datasets and does not involve new human-subject experiments or the collection of additional personal data. Face anti-spoofing can improve the security of biometric authentication systems, but incorrect predictions may lead to false rejection of legitimate users or acceptance of presentation attacks. Moreover, performance may vary across acquisition conditions, attack types, and unseen domains. The proposed evidence-to-decision trace is intended to support model analysis and transparency, rather than guarantee the correctness of individual predictions. We therefore recommend appropriate validation and human oversight before deploying such systems in security-sensitive applications.

REPRODUCIBILITY STATEMENT

We provide detailed methodological and experimental information in the main paper and appendix. The method section specifies the Structured Evidence Slots (SES), Difficulty-Aware Progressive Reasoning (DPR), training objectives, confidence-based routing, and test-time inference procedure. The experimental section documents the CelebA-Spoof-to-Eleven evaluation protocol, dataset splits, evaluation metrics, implementation details, and hyperparameter settings. We additionally report three-seed performance statistics, component and mechanism ablations, hyperparameter sensitivity, counterfactual slot-suppression experiments, and deployment latency and throughput measurements. Results adopted from prior studies are distinguished from our own experiments. The anonymized source code is available at `https://anonymous.4open.science/r/EviTrace`.

REFERENCES

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

Rizhao Cai, Zhi Li, Renjie Wan, Haoliang Li, Yongjian Hu, and Alex C Kot. Learning meta pattern for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17:1201–1213, 2022.

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

Artur Costa-Pazo, Sushil Bhattacharjee, Esteban Vazquez-Fernandez, and Sebastien Marcel. The replay-mobile face presentation-attack database. In *2016 international conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–7. IEEE, 2016.

Anjith George and Sébastien Marcel. Cross modal focal loss for rgbd face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7882–7891, 2021.

Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE transactions on information forensics and security*, 15:42–55, 2019.

Xiao Guo, Yaojie Liu, Anil Jain, and Xiaoming Liu. Multi-domain learning for updating face anti-spoofing models. In *European conference on computer vision*, pp. 230–249. Springer, 2022.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 105–116, 2025.

Yan He, Fei Peng, Rizhao Cai, Zitong Yu, Min Long, and Kwok-Yan Lam. Category-conditional gradient alignment for domain adaptive face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 19:10071–10085, 2024.

Pei-Kai Huang, Cheng-Hsuan Chiang, Tzu-Hsien Chen, Jun-Xiong Chong, Tyng-Luh Liu, and Chiou-Ting Hsu. One-class face anti-spoofing via spoof cue map-guided feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 277–286, 2024.

Fangling Jiang, Qi Li, Bing Liu, Weining Wang, Caifeng Shan, Zhenan Sun, and Ming-Hsuan Yang. Learning knowledge-based prompts for robust 3d mask presentation attack detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Arman Keresh and Pakizar Shamoi. Liveness detection in computer vision: Transformer-based self-supervised learning for face anti-spoofing. *IEEE Access*, 2024.

Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13 (7):1794–1809, 2018.

Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Wenzhong Tang, Zitong Yu, and Alex Kot. Suppress and rebalance: Towards generalized multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 211–221, 2024.

Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1179–1187, 2021a.

Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z Li. Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16:2759–2772, 2021b.

Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17: 2497–2507, 2022.

Ajian Liu, Xun Lin, Ruicong Zhi, Yanyan Liang, Xinshan Zhu, Zhanchuan Cai, Jun Wan, Sergio Escalera, and Zhen Lei. Dgpdl: Domain-guided prompt distribution learning for generalizable face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Si-Qi Liu, Xiangyuan Lan, and Pong C Yuen. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 558–573, 2018a.

Yao Liu, Ying Tai, Jilin Li, Shouhong Ding, Chengjie Wang, Feiyue Huang, Dongyang Li, Wenshuai Qi, and Rongrong Ji. Aurora guard: Real-time face anti-spoofing via light reflection. *arXiv preprint arXiv:1902.10311*, 2019.

Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 389–398, 2018b.

Zohreh Mostaani, Anjith George, Guillaume Heusch, David Geissbuhler, and Sebastien Marcel. The high-quality wide multi-channel attack (hq-wmca) database. *arXiv preprint arXiv:2009.09703*, 2020.

Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10023–10031, 2019.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. *arXiv preprint arXiv:2508.21048*, 2025.

Dong Wang, Jia Guo, Qiqi Shao, Haochi He, Zhian Chen, Chuanbao Xiao, Ajian Liu, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, et al. Wild face anti-spoofing challenge 2023: Benchmark and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6380–6391, 2023.

Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5042–5051, 2020.

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

Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3005–3023, 2020a.

Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5295–5305, 2020b.

Guosheng Zhang, Keyao Wang, Haixiao Yue, Ajian Liu, Gang Zhang, Kun Yao, Errui Ding, and Jingdong Wang. Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9896–9904, 2025a.

Haoyuan Zhang, Keyao Wang, Guosheng Zhang, Haixiao Yue, Zhiwen Tan, Siran Peng, Tianshuo Zhang, Xiao Tan, Kunbin Chen, Wei He, et al. From intuition to investigation: A tool-augmented reasoning mllm framework for generalizable face anti-spoofing. *arXiv preprint arXiv:2603.01038*, 2026.

Honglu Zhang, Zhiqin Fang, Ningning Zhao, Saihui Hou, Long Ma, Renwang Pei, and Zhaofeng He. Facecot: A benchmark dataset for face anti-spoofing with chain-of-thought reasoning. *arXiv preprint arXiv:2506.01783*, 2025b.

Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *European conference on computer vision*, pp. 70–85. Springer, 2020.

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

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

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025.

A Related Work

A.1 Binary Face Anti-Spoofing

Most existing FAS methods formulate spoof detection as a binary recognition task. Given an input face image, the model extracts visual features and maps them to a single prediction score indicating whether the sample is bona fide or spoof. The optimization objective is therefore usually defined by a binary classification loss, sometimes combined with auxiliary supervision such as depth maps Liu et al. (2018b), reflection cues Liu et al. (2019), temporal consistency Wang et al. (2020), or domain alignment constraints Shao et al. (2019). During inference, the final live/fake score is compared with a threshold to make the decision. Following this formulation, Deep FAS methods use CNN- or Transformer-based representations, enabling stronger feature learning from RGB, depth, infrared, or multi-modal inputs Liu et al. (2018b); Yu et al. (2020b); Liu et al. (2021b); George & Marcel (2021); Huang et al. (2024). More recent studies further improve cross-domain robustness through domain generalization, meta-learning, adversarial learning, test-time adaptation, and prompt-based representation learning Cai et al. (2022); Guo et al. (2022); Zhou et al. (2024); Lin et al. (2024); He et al. (2024); Jiang et al. (2025); Liu et al. (2026). Although these methods differ in feature design, network architecture, or domain adaptation strategy, their decision mechanism remains largely unchanged: diverse spoofing patterns are ultimately compressed into a single live/fake decision space and all samples are processed by the same one-pass inference pipeline.

A.2 Text-Interpretable Face Forgery Detection

Recent Multimodal Large Language Model (MLLM) based methods have introduced textual reasoning into face forgery analysis. Representative works use tool-augmented MLLM reasoning for generalizable FAS Zhang et al. (2026), pattern-aware reasoning for deepfake detection Tan et al. (2025), chain-of-thought annotations for FAS Zhang et al. (2025b), multimodal interpretable forged-face detection Guo et al. (2025), and MLLM-based explanations for AI-generated image detection Zhou et al. (2025). These methods improve interpretability by generating natural-language rationales or reasoning chains, showing the potential of language-guided evidence analysis. However, their reliance on large generative vision-language models usually leads to high computational cost, slow inference, and substantial memory requirements. Moreover, the generated explanations are not always tightly coupled with compact visual evidence representations, which limits their suitability for efficient and stable FAS deployment.

B Dataset-level slot response analysis on live samples.

Fig. 5 presents the aggregated top-1 slot response statistics on bona fide samples from six representative unseen target datasets. In contrast to the spoof-sample analysis in the main paper, the responses on live samples are generally much weaker and do not exhibit the same sharp slot dominance patterns. Most datasets show only low-amplitude and relatively scattered activations, suggesting that the learned slots are not indiscriminately triggered on bona fide faces.

More specifically, ReplayAttack, SiW-Mv2, and IDIAP-WMCA exhibit very weak overall responses across nearly all slots, indicating good suppression of spoof-oriented evidence on live samples. CASIA-FASD and OULU-NPU show slightly higher responses on a few slots such as *Poster* and *Pad*, but the activations remain substantially weaker and less concentrated than those observed on spoof samples. CASIA-HiFiMask yields the strongest live-sample responses among the six datasets, with moderate activation spread over several slots rather than being dominated by a single spoof-specific cue.

Overall, this comparison complements the spoof-sample analysis by showing that SES is not only capable of producing semantically meaningful slot activations for spoof samples, but also tends to remain relatively inactive or weakly activated on bona fide samples. This behavior supports the evidence specificity of SES and improves the interpretability of the subsequent reasoning process in DPR.

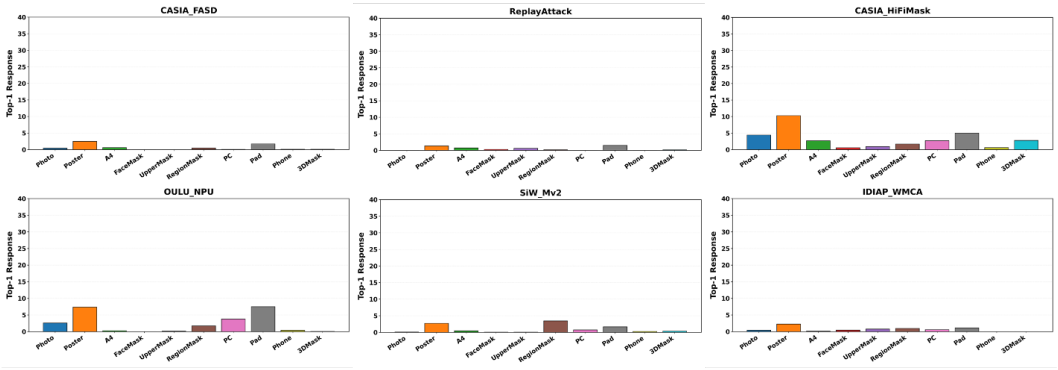


Figure 5: Aggregated top-1 slot response statistics on bona fide samples from six representative unseen target datasets. Compared with spoof samples, live samples generally produce much weaker and less concentrated slot responses. ReplayAttack, SiW-Mv2, and IDIAP-WMCA show very low overall activation, while CASIA-FASD, OULU-NPU, and CASIA-HiFiMask exhibit only moderate responses distributed across multiple slots rather than a single dominant spoof-specific slot. This suggests that SES does not indiscriminately activate spoof-related evidence on live samples, supporting the specificity and interpretability of the learned evidence slots.

Table 5: Counterfactual slot-suppression analysis on three representative unseen target datasets. We suppress the dominant slots identified in Fig. 3 and report the resulting HTER degradation. ΔHTER denotes the increase in HTER after suppression.

Target dataset	**Suppressed slots**	**Original HTER (%)**	**Suppressed HTER (%)**	**ΔHTER (pp)**
CASIA-FASD	Photo + Poster + A4	1.33	3.92	+2.59
Replay-Attack	Pad	2.14	7.06	+4.92
CASIA-HiFiMask	3DMask	33.80	39.47	+5.67

B.1 Counterfactual Analysis of Slot Faithfulness

To quantitatively examine whether the visualized slot responses are functionally relevant to the model's predictions, we conduct a counterfactual slot-suppression experiment on three representative unseen target datasets. Specifically, we suppress the dominant attack-type slots identified in Fig. 3 during inference and evaluate the resulting performance degradation. This experiment aims to establish a direct connection between the observed slot responses and the model's actual prediction behavior, rather than relying solely on qualitative visualizations.

As shown in Tab. 5, suppressing the dominant slots consistently degrades detection performance across all three target datasets. Specifically, removing the print-related slots on CASIA-FASD increases HTER by 2.59 percentage points, while suppressing the Pad slot on Replay-Attack and the 3DMask slot on CASIA-HiFiMask leads to increases of 4.92 and 5.67 points, respectively. These results provide quantitative evidence that the dominant slot responses visualized in Fig. 3 are functionally relevant to the model's predictions rather than merely serving as visually plausible explanations.

Importantly, IRT exposes intermediate variables directly involved in the inference process, including slot activation and confidence, difficulty estimation and routing, selected slots and local regions, and predictions before and after refinement. The slot-suppression results provide initial quantitative support for the functional relevance of these intermediate representations. Nevertheless, this experiment does not fully establish the semantic accuracy of individual slots or the faithfulness of DPR-selected local regions. We therefore regard IRT as an explicit and auditable evidence-to-decision trace, with the present analysis providing complementary evidence beyond qualitative visualization.

Table 6: Model complexity of EviTrace instantiated on CLIP ViT-B/16. FLOPs are computed for a single 224×224 RGB image. Average inference FLOPs are estimated using a routing ratio of 16.75%.

Component / Metric	**Params (M)**	**FLOPs (G)**
Visual backbone (trainable part)	86.19	–
SES added modules	0.27	–
Coarse branch (exclusive)	0.53	0.006
Coarse branch (incl. shared slot head)	0.79	–
Refinement branch	1.84	0.058
Visual encoder with slot tokens	–	35.580
Total model	152.26	–
Trainable parameters	88.84	–
Frozen parameters	63.43	–
Coarse-only inference	–	35.586
Hard-sample full inference	–	35.644
Average inference (route ratio = 16.75%)	–	35.596

Table 7: Measured inference efficiency on an NVIDIA RTX 4090. We compare coarse-only, all-routed, and mixed-batch inference at the default 16.75% routing ratio. Latency is reported in milliseconds per batch, and throughput in images per second. The p95 and p99 values characterize mixed-batch tail latency.

Batch	**Coarse-only**		**All-routed**		**Mixed-batch (16.75%)**			
	Mean (ms)	Throughput (img/s)	Mean (ms)	Throughput (img/s)	Mean (ms)	p95 (ms)	p99 (ms)	Throughput (img/s)
1	8.564	116.76	9.506	105.19	8.686	9.563	9.899	115.13
8	8.697	919.83	9.633	830.48	9.424	10.125	10.898	848.87
16	9.447	1693.58	10.450	1531.17	10.451	10.749	11.751	1530.91
32	19.055	1679.34	20.030	1597.61	20.085	20.331	20.500	1593.23

Model complexity. Tab. 6 reports the parameter and computational complexity of EviTrace instantiated on CLIP ViT-B/16. The full model contains 152.26M parameters, of which 88.84M are trainable and 63.43M are frozen. Importantly, the additional SES and DPR modules are lightweight relative to the visual backbone: SES introduces only 0.27M parameters, while the refinement branch contains 1.84M parameters.

In terms of computation, the visual encoder with slot tokens dominates the overall cost, requiring 35.58 GFLOPs for a single 224×224 RGB image. The coarse branch adds only 0.006 GFLOPs, and the refinement branch introduces an additional conditional cost of 0.058 GFLOPs for each routed hard sample. As a result, coarse-only inference costs 35.586 GFLOPs, while the worst-case full inference for a routed hard sample costs 35.644 GFLOPs. Using the observed routing ratio of 16.75%, the estimated average inference cost is 35.596 GFLOPs, indicating that DPR introduces only a marginal overhead on top of the visual backbone.

B.2 Deployment Efficiency

To evaluate the practical inference cost of dynamic routing, we measure the latency and throughput of coarse-only, all-routed, and mixed-batch inference on an idle NVIDIA RTX 4090 using 224×224 inputs. We perform 100 warm-up iterations followed by 1,200 timed iterations. The measurements cover the GPU-resident model forward pass and exclude image decoding, preprocessing, and host-to-device transfer. Mixed-batch inference uses the default routing ratio of 16.75%, measured over the complete inference stream for batch size 1.

Tab. 7 shows that mixed-batch inference introduces mean-latency increases of 1.4%, 8.4%, 10.6%, and 5.4% at batch sizes of 1, 8, 16, and 32, respectively, compared with coarse-only inference. The corresponding throughput reductions are 1.4%, 7.7%, 9.6%, and 5.1%. Although selective routing

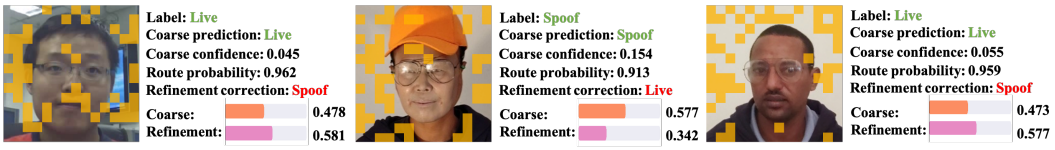


Figure 6: DPR failure cases where refinement reverses correct coarse predictions. Selected patches focus mainly on irrelevant background regions, leading to incorrect decisions. Orange and pink bars show spoof probabilities before and after refinement, respectively.

Table 8: Ablation study of loss weights in EviTrace, reported with HTER (%). Lower is better.

Hyperparameter	Value_1	HTER	Value_2	HTER	Value_3	HTER	Value_4	HTER
λ_{align}	0.000	10.04	0.005	10.64	0.010	**9.98**	0.030	10.17
λ_{slot}	0.000	12.05	0.075	13.83	0.150	**9.98**	0.300	10.65
λ_{prog}	0.000	13.56	0.350	12.12	0.650	**9.98**	1.000	10.43

reduces the number of refined samples, mixed-batch latency approaches all-routed latency at larger batch sizes, suggesting that dynamic control flow and routed-subbatch handling introduce additional execution overhead. These results demonstrate the importance of measuring actual latency and throughput, rather than relying solely on FLOPs, when assessing the deployment cost of DPR.

B.3 Ablation on Loss Weights

Tab. 8 presents the ablation results of the three loss weights in EviTrace. We observe that the default setting, i.e., $\lambda_{\text{slot}} = 0.15$, $\lambda_{\text{align}} = 0.01$, and $\lambda_{\text{prog}} = 0.65$, achieves the best overall performance with an average HTER of 9.98. For λ_{align}, either decreasing the alignment weight or increasing it to 0.03 leads to degraded performance, indicating that a mild slot-text alignment constraint is most effective. For λ_{slot}, the performance drops substantially when the supervision weight is reduced, especially at 0.075, showing that slot-level supervision is important for learning stable discriminative cues. For λ_{prog}, both removing the progressive reasoning term and assigning it a larger weight result in worse average HTER, suggesting that a moderate reasoning coefficient provides the best trade-off. Overall, these results validate the effectiveness of the adopted default weighting strategy.

C Failure Case Analysis

Fig. 6 shows representative failure cases where DPR degrades a correct coarse decision. Although the coarse branch already predicts the correct class, the refinement branch attends primarily to background regions instead of informative facial cues, and consequently shifts the spoof probability toward the wrong class. This suggests that DPR may occasionally overreact to contextual noise, especially when the selected local regions are not aligned with the true discriminative evidence.