VADAPT: VISUAL ANCHORED AND DUAL-PROMPT TUNING FOR GENERALIZABLE FACE ATTACK DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizable face attack detection must handle not only conventional physical presentation attacks but also rapidly evolving digital manipulations. While pre-trained vision-language models (VLMs) offer transferable representations, their direct adaptation to this task remains limited by heterogeneous real/fake distributions and concentrated reliance on a narrow set of visual evidence. To address these limitations, we propose VADAPT, a Visual Anchored and Dual-Prompt Tuning framework for generalizable face attack detection. VADAPT introduces Visual Soft-Anchoring, which uses indexed textual anchors to softly differentiate multiple learnable visual part tokens, encouraging diverse fine-grained evidence exploration without predefined anatomical partitions. In parallel, Dual-Branch Latent Context learns multiple continuous contexts to capture heterogeneous genuine and attack-related representations, while Asymmetric Feature Aggregation adaptively combines their responses using branch-specific aggregation rules. Extensive experiments on UniAttackData and HydraFake demonstrate strong cross-attack generalization. VADAPT achieves state-of-the-art performance on UniAttackData under the official cross-form protocol, attaining 13.45% ACER, compared with 15.13% for the strongest previous baseline.

1 INTRODUCTION

Face recognition systems have been widely deployed in security-critical applications, including mobile authentication, financial payment, access control, and video surveillance (Wang & Deng, 2021). Their reliability, however, is increasingly challenged by diverse facial attacks. Conventional threats mainly include physical presentation attacks such as printed photographs, replay videos, and 3D masks (Li et al., 2020; Zhang et al., 2020; Liu et al., 2022a; Fang et al., 2024a), while recent generative models have enabled digital manipulations including face swapping, reenactment, attribute editing, and synthetic face generation (Rossler et al., 2019; Shiohara & Yamasaki, 2022; Liu et al., 2023a; Yan et al., 2024a). A practical detector must therefore generalize across attack mechanisms that differ substantially in media, generation processes, visual artifacts, and deployment conditions.

Existing face attack detectors often perform well on known attacks but degrade when confronted with unseen spoofing media, manipulation techniques, generators, or acquisition conditions (Hong et al., 2023; Fang et al., 2024b; Zhou et al., 2024). Prior studies have improved generalization through auxiliary supervision, data augmentation, adversarial learning, meta-learning, and domain-invariant representation learning (Liu et al., 2023b; Sun et al., 2023; Yang et al., 2025a; Cai et al., 2025). Nevertheless, these approaches still largely learn from visual patterns observed in the training domains, which can be restrictive in unified face attack detection where future attack distributions cannot be exhaustively anticipated.

Vision-Language Models (VLMs), such as CLIP (Radford et al., 2021), offer an appealing alternative by providing transferable representations in a shared vision-language space. Recent studies have explored VLMs for face anti-spoofing and face forgery detection using hand-crafted descriptions or real/fake prompts (Liu et al., 2025; Huang et al., 2025a). However, directly adapting a global CLIP representation and a small set of fixed prompts leaves two task-specific issues unresolved. We illustrate these observations in Figure 1.

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.

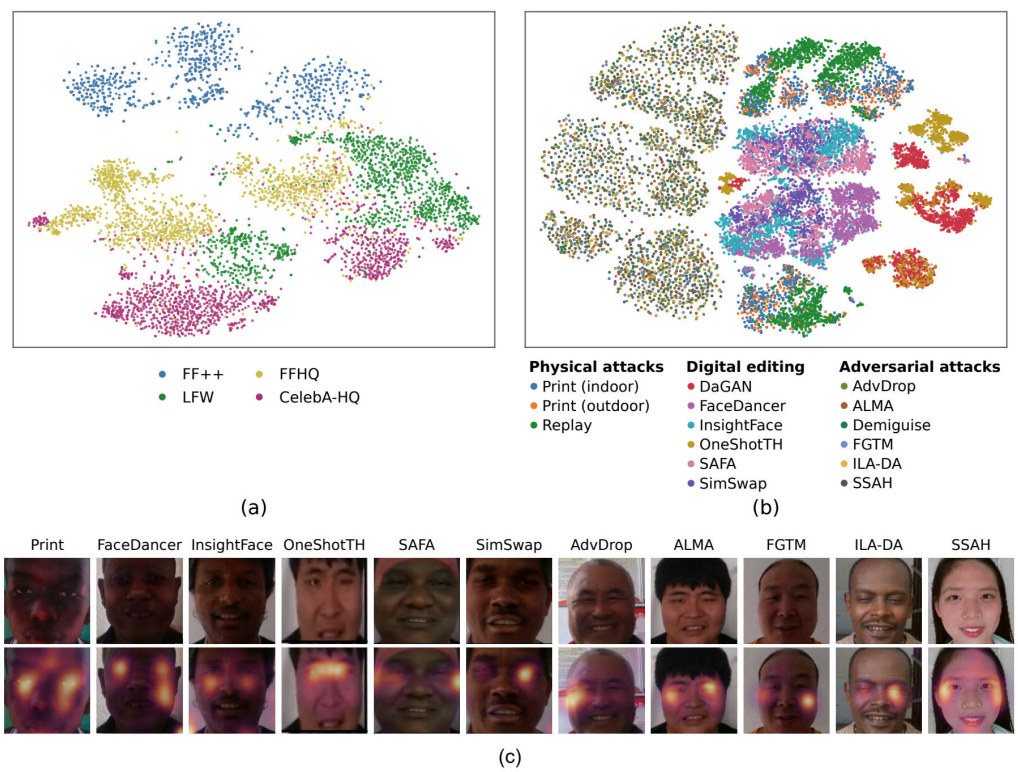


Figure 1: (a) t-SNE visualization of genuine samples from different source datasets, showing source-dependent clustering in the CLIP feature space. (b) t-SNE visualization of attack samples, where physical, digital-editing, and adversarial attacks exhibit coarse group-level structure, while related attack methods show substantial local overlap. (c) Grad-CAM responses of the global `[CLS]` token across heterogeneous attack methods, which frequently concentrate on similar facial regions.

First, as shown in Figure 1(a–b), genuine samples from different sources show clearer separation in the CLIP feature space, while related attack methods exhibit more substantial overlap. This suggests that the real and fake branches need not combine multiple latent contexts in the same way. We therefore learn multiple latent contexts for both branches and adopt branch-specific aggregation strategies, which are evaluated empirically in our ablation studies.

Second, Figure 1(c) shows that the global `[CLS]` token often highlights similar facial regions across substantially different attack methods. As a result, the model may miss useful forensic cues from other parts of the face. Rather than imposing predefined anatomical partitions, we seek multiple learnable evidence-seeking directions that can adaptively explore diverse fine-grained cues.

Motivated by these observations, we propose **VADAPT**, a Visual Anchored and Dual-Prompt Tuning framework for generalizable face attack detection. VADAPT adapts both visual evidence exploration and real/fake context modeling through three complementary components: Visual Soft-Anchoring, Dual-Branch Latent Context, and Asymmetric Feature Aggregation.

The main contributions of this work are summarized as follows:

- We propose **Visual Soft-Anchoring (VSA)**, which uses indexed textual anchors to softly differentiate multiple learnable visual part tokens, encouraging the model to explore diverse fine-grained facial evidence without predefined anatomical regions.
- We introduce **Dual-Branch Latent Context (DLC)**, which learns multiple continuous contexts for genuine and attack samples separately, allowing the model to represent heterogeneous real/fake patterns without relying on manually defined source or attack categories.

- We design **Asymmetric Feature Aggregation (AFA)** to combine the learned real and fake contexts with different aggregation rules and further integrate part-level evidence for final prediction.

  Extensive experiments on UniAttackData and HydraFake demonstrate strong cross-form and cross-distribution generalization, with state-of-the-art performance on UniAttackData.

2 Related Work

2.1 Generalizable Face Attack Detection

Face attack detection has been investigated from two closely related perspectives: face anti-spoofing (FAS), which primarily addresses physical presentation attacks, including print, replay, and mask attacks, and face forgery detection (FFD), which focuses on digital facial manipulations such as face swapping, reenactment, and synthetic face generation (Fang et al., 2024b; Ma et al., 2025). Existing approaches have achieved promising performance against known attack types by exploiting discriminative visual cues derived from texture, depth, reflection, frequency patterns, physiological signals, or manipulation artifacts (He et al., 2024; Jung et al., 2026). Nevertheless, their generalization capability remains limited when encountering unseen spoofing media, emerging digital forgery techniques, or diverse acquisition conditions (Jiang et al., 2024; Zhou et al., 2024). To enhance robustness under such distribution shifts, prior studies have mainly explored domain-invariant visual representation learning through domain alignment, data-centric augmentation, and prompt-based adaptation (Srivatsan et al., 2023; Yang et al., 2025a; Guo et al., 2025a). Although these methods mitigate domain shifts to some extent, they still largely rely on visual feature alignment or assumptions about known attack distributions, thereby limiting their effectiveness in open-ended generalization across both physical and digital attack scenarios.

2.2 Vision-Language Models for Face Attack Detection

Vision-language models (VLMs), such as CLIP, align visual and textual representations within a shared semantic embedding space and exhibit strong transferability across downstream recognition tasks. Recent studies have introduced VLMs into face attack detection by leveraging textual descriptions, real/fake prompts, or facial semantic concepts to guide visual discrimination (Yu et al., 2025; Lin et al., 2025; Sun et al., 2025; Chen et al., 2025b; Jiang et al., 2026). By exploiting pretrained semantic knowledge, these methods improve adaptability in limited-supervision and cross-domain settings. However, most existing VLM-based approaches still depend on global image-text matching or predefined textual prompts, leaving fine-grained local attack evidence insufficiently modeled.

2.3 Prompt Tuning for Vision-Language Adaptation

Prompt tuning has emerged as an effective parameter-efficient strategy for adapting large pretrained models to downstream tasks. It was initially explored in natural language processing, where methods such as P-tuning replace hand-crafted discrete prompts with learnable continuous prompt embeddings to better elicit task-relevant knowledge from pretrained language models (Liu et al., 2024b; 2022b). This paradigm has subsequently been extended to vision-language models. For instance, CoOp optimizes learnable context tokens in the text branch of CLIP, replacing manually designed templates such as ‘a photo of a [CLASS]’ with trainable continuous prompts for visual recognition (Zhou et al., 2022b). Subsequent studies further advance prompt learning by incorporating class-specific prompts, conditional prompts, domain-aware prompts, and multimodal prompts, demonstrating strong adaptation capability in few-shot learning, domain adaptation, and domain generalization scenarios (Zhou et al., 2022a; Khattak et al., 2023; Kan et al., 2023; Wen et al., 2025). Despite these advances, many prompt-based VLM adaptation methods still rely on predefined discrete class names, whereas the coarse labels ‘real’ and ‘fake’ in face attack detection correspond to highly heterogeneous visual distributions (Liu et al., 2024a; Jiang et al., 2026).

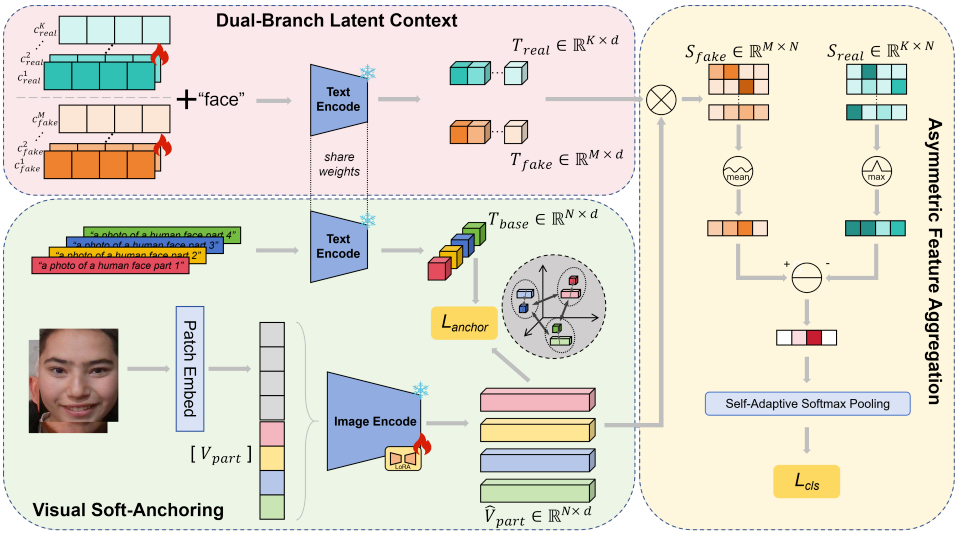


Figure 2: Overview of the VADAPT framework. A CLIP/FaRL backbone is adapted with LoRA for parameter-efficient fine-tuning. (a) Visual Soft-Anchoring introduces learnable visual part tokens $\hat{V}_{part}$ and indexed textual anchors T_{base} to encourage diverse fine-grained evidence representations. (b) Dual-Branch Latent Context learns multiple continuous context prompts for the real and fake branches (C_{real} and C_{fake}), providing multiple latent compatibility representations. (c) Asymmetric Feature Aggregation applies branch-specific pooling over the learned contexts and then adaptively combines the resulting part-level attack evidence for final prediction.

3 Proposed Method

This section presents VADAPT for generalizable face attack detection. As illustrated in Figure 2, VADAPT is built on a CLIP backbone initialized with FaRL (Zheng et al., 2022) and adapted using LoRA (Hu et al., 2022). The framework jointly adapts visual evidence extraction and real/fake context modeling within the shared vision-language space. VADAPT consists of three key components:

- Visual Soft-Anchoring (VSA), which encourages multiple visual part tokens to capture diverse fine-grained facial evidence;
- Dual-Branch Latent Context (DLC), which learns multiple continuous contexts to model heterogeneous real and fake patterns;
- Asymmetric Feature Aggregation (AFA), which combines the two context branches with different aggregation rules and integrates part-level evidence for final prediction.

In practice, LoRA is used to adapt the backbone in a parameter-efficient manner, while VSA, DLC, and AFA constitute the task-specific components of VADAPT.

3.1 Visual Soft-Anchoring

As shown in Figure 1(c), the global `[CLS]` token often focuses on similar facial regions across different attack methods. To encourage broader exploration of fine-grained visual cues, we introduce Visual Soft-Anchoring (VSA). VSA augments the visual encoder with multiple learnable part tokens and differentiates their representations using indexed textual anchors. Importantly, the anchors do not correspond to predefined facial regions or enforce non-overlapping attention; instead, they provide a soft separation signal in the shared vision-language space.

Given an input image x, we introduce N learnable visual part tokens $V_{part} \in \mathbb{R}^{N \times D}$, where D denotes the feature dimension. The part tokens are concatenated with the class token and image patch tokens and processed by the visual encoder $\mathcal{F}(\cdot)$:

$$\hat{V}_{part} = \mathcal{F}(x, V_{part}),$$



Figure 3: Qualitative Grad-CAM comparison of eight visual part tokens for the same input. **Left:** Visual Soft-Anchoring with indexed textual anchors. **Right:** the same multi-token visual design without textual anchoring. The anchored tokens exhibit more diverse attention patterns across facial regions, whereas the unanchored tokens show more similar responses.

where $\hat{V}_{part} \in \mathbb{R}^{N \times D}$ contains the resulting part-level visual representations.

To differentiate these representations, we use CLIP's frozen text encoder to obtain N indexed textual anchors $T_{base} \in \mathbb{R}^{N \times D}$. The anchor texts are intentionally generic and differ only by their indices, serving as latent references rather than semantic descriptions of facial regions. For the i-th visual part representation $\hat{v}_i$ and anchor feature t_i, we define the soft-anchoring loss as

$$\mathcal{L}_{anchor} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\hat{v}_i t_i^\top)}{\sum_{j=1}^{N} \exp(\hat{v}_i t_j^\top)}.$$

This objective encourages different part tokens to align with different anchor references while allowing spatial overlap between their attention patterns. Figure 3 qualitatively shows that textual anchoring leads to more diverse attention responses than the same multi-token design without anchoring. Quantitative ablations in Section 4.3 further evaluate its effect on detection performance.

3.2 Dual-Branch Latent Context

Figure 1(a–b) shows that both genuine and attack samples contain substantial intra-class variation. Genuine samples vary across different data sources, while attack samples cover diverse physical, digital, and adversarial manipulations. A single textual prompt is therefore unlikely to capture the full variation of either class. Rather than manually defining prompts for individual sources or attack types, we introduce Dual-Branch Latent Context (DLC), which learns multiple continuous contexts independently for the real and fake branches.

Specifically, we introduce K learnable contexts for the real branch, $C_{real} \in \mathbb{R}^{K \times L \times D}$, and M learnable contexts for the fake branch, $C_{fake} \in \mathbb{R}^{M \times L \times D}$, where L denotes the prompt length and D the feature dimension. Each learnable context is concatenated with the fixed suffix 'face' and then passed through the pretrained text encoder $\mathcal{G}(\cdot)$:

$$T_{real} = \mathcal{G}(C_{real}), \qquad T_{fake} = \mathcal{G}(C_{fake}),$$

where $T_{real} \in \mathbb{R}^{K \times D}$ and $T_{fake} \in \mathbb{R}^{M \times D}$.

The resulting text features serve as learnable real/fake prototypes in the shared vision-language space. Individual contexts are not assigned to specific datasets or attack methods, and we do not impose an explicit diversity or orthogonality constraint among them. Instead, they are jointly optimized by the classification objective. Ablation results and context visualizations further support the benefit of using multiple learnable contexts over a single real/fake prompt pair.

3.3 Asymmetric Feature Aggregation

Given the visual part representations and the dual-branch textual contexts, we compute the scaled cosine similarities $S_{real} \in \mathbb{R}^{K \times N}$ and $S_{fake} \in \mathbb{R}^{M \times N}$. We first aggregate these similarities along the textual-context dimension, using different rules for the real and fake branches.

As suggested by Figure 1(a–b), genuine and attack samples exhibit different feature structures in the pretrained CLIP space. Genuine samples from different sources are relatively well separated, indicating multiple distinct real modes. For a given visual part, it is therefore sufficient to match one plausible real context rather than requiring agreement with all real contexts. We accordingly use Max pooling over the K real contexts.

In contrast, related attack methods show substantially more overlap in the feature space, suggesting a less clearly separated structure on the fake side. In this case, relying only on the single largest fake-context response may make the decision overly sensitive to one context. We therefore average the responses over the M fake contexts to obtain a more stable aggregated attack response. The feature visualization provides the motivation for this asymmetric design, while the pooling ablation in Section 4.3 evaluates the choice empirically.

Accordingly, the attack logit for the n-th visual part is

$$l_n = \frac{1}{M}\sum_{j=1}^{M} S_{fake}^{(j,n)} - \max_{1\le i\le K} S_{real}^{(i,n)}.$$

The Max/Mean operations above are performed only over textual contexts, not over visual parts. We subsequently aggregate the N part-level logits using Self-Adaptive Softmax Pooling (SASP):

$$\alpha_n = \frac{\exp(l_n)}{\sum_{i=1}^{N}\exp(l_i)}, \qquad L_{final} = \sum_{n=1}^{N}\alpha_n l_n.$$

SASP assigns larger weights to visual parts with stronger attack responses, allowing local evidence to contribute adaptively to the final prediction.

3.4 Overall Optimization Objective

Given the final prediction score L_{final}, we use the standard binary cross-entropy loss:

$$\mathcal{L}_{cls} = -\left[y\log\sigma(L_{final}) + (1-y)\log\left(1-\sigma(L_{final})\right)\right],$$

where $y \in \{0,1\}$ denotes the ground-truth label, with 0 for genuine samples and 1 for attack samples.

The overall objective combines the classification loss with the soft-anchoring loss introduced in Section 3.1:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{anchor},$$

where λ controls the contribution of the anchoring term.

Note that $\mathcal{L}_{anchor}$ regularizes only the visual part tokens with respect to the indexed textual anchors. The dual-branch contexts are optimized through $\mathcal{L}_{cls}$ without an explicit pairwise diversity or orthogonality constraint.

4 Experiments

We evaluate VADAPT on UniAttackData and HydraFake, covering physical presentation attacks and digital face forgeries. We compare against existing methods and conduct ablation and qualitative analyses to examine the effectiveness of each component.

4.1 Implementation Setup

Datasets and Evaluation Metrics. We evaluate VADAPT on UniAttackData (Fang et al., 2024b) and HydraFake (Tan et al., 2025). For UniAttackData, we follow the official cross-form evaluation protocol, which jointly reports performance over the physical-to-digital and digital-to-physical transfer directions. We report ACER, ACC, AUC, and EER according to this official setting. HydraFake evaluates cross-model, cross-forgery, and cross-domain robustness for digital forgeries, for which we report average ACC across the evaluation subsets.

Method	ACER(%)↓	ACC(%)↑	AUC(%)↑	EER(%)↓
ResNet50	34.60±5.31	53.69±6.39	87.89±6.11	19.48±9.10
VIT-B/16	33.69±9.33	52.43±25.88	83.77±2.35	25.94±0.88
Auxiliary	42.98±6.77	37.71±26.45	76.27±12.06	32.66±7.91
CDCN	34.33±0.66	53.10±12.70	77.46±17.56	29.17±14.47
FFD	44.20±1.32	40.43±14.88	80.97±2.86	26.18±2.77
UAD	22.42±10.57	67.35±23.22	91.97±4.55	15.72±3.08
HML-FAS	16.02±5.97	75.24±4.81	92.77±6.10	15.61±6.22
MoAE-CR	15.13±12.10	85.41±6.85	92.09±7.11	13.81±8.71
VADAPT	**13.45±2.51**	**86.02±1.95**	**93.47±1.75**	**13.53±2.52**

Table 1: Performance comparison on UniAttackData under the official cross-form evaluation protocol. Results follow the official aggregate setting over the physical-to-digital and digital-to-physical transfer directions. The best results are highlighted in bold.

Method	ID	CM	CF	CD	Avg.
MLLM-based Forgery Detectors					
M2F2-Det	-	58.7	69.3	61.8	63.2
SIDA-7B	-	87.9	71.7	69.4	76.3
FakeVLM	-	76.9	75.7	79.1	77.3
VERITAS	97.3	98.6	90.3	82.2	90.7
Small Vision Models					
UniFD	82.7	87.5	73.7	71.9	78.0
ProDet	90.5	92.3	76.4	71.4	80.6
Co-SPY	86.6	93.3	85.9	74.7	84.7
D^3	87.3	93.8	75.5	73.1	81.1
Effort	**94.7**	90.7	**86.6**	67.1	82.2
VADAPT (Ours)	93.2	**97.3**	77.3	**80.4**	**85.4**

Table 2: Comparison with existing methods on HydraFake. Results are grouped by model type. Bold indicates the best results in small vision models. ID, CM, CF, and CD denote Intra-Domain, Cross-Model, Cross-Forgery, and Cross-Domain, respectively.

Implementation Details. We use CLIP with a ViT-B/16 visual encoder as the vision-language backbone and initialize it with FaRL. LoRA with rank 8 is used for parameter-efficient adaptation. Input images are resized to 224×224, and the batch size is set to 32. We use $N = 8$ visual part tokens, $K = 4$ real contexts, and $M = 2$ fake contexts, with a prompt length of $L = 4$. The numbers of real and fake contexts are selected on the UniAttackData validation set, where $K = 4$ and $M = 2$ achieve the best validation performance. We use the same setting on HydraFake without further tuning. The indexed textual anchors use the CLIP-style template "a photo of a human face part i", $i \in \{1, \ldots, N\}$. The index distinguishes the anchors without assigning them to predefined anatomical regions. The model is optimized using AdamW with a fixed learning rate of 1×10^{-4}, and the anchoring-loss weight is set to $\lambda = 0.5$. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

4.2 Performance

Results on UniAttackData. We compare VADAPT with representative baselines, including ResNet50 (He et al., 2016), ViT-B/16 (Dosovitskiy et al., 2020), FFD (Dang et al., 2020), CDCN (Yu et al., 2020), Auxiliary(Depth) (Liu et al., 2018), UniAttackDetection (UAD) (Fang et al., 2024b), HML-FAS (Mao & Li, 2025), and MoAE-CR (Chen et al., 2025a).

Table 1 reports results under the official cross-form generalization protocol, covering both physical-to-digital and digital-to-physical transfer. This setting requires models to generalize across substantially different attack formation mechanisms rather than only within a single attack domain. VADAPT achieves the best performance on all four metrics, with the lowest ACER and EER and the highest ACC and AUC among the compared methods. In particular, it reduces ACER from

Method	ACER(↓)	ACC(↑)	AUC(↑)	EER(↓)
CLIP	24.45±2.20	67.85±4.41	90.82±5.95	16.70±6.66
CLIP w/ FaRL	24.21±0.05	75.64±1.71	79.35±2.69	24.81±1.11
w/o FaRL	14.13±1.17	85.54±1.39	93.03±0.95	14.25±1.10
VADAPT	**13.45±2.51**	**86.02±1.95**	**93.47±1.75**	**13.53±2.52**

Table 3: Ablation study on the role of face-specific pretraining in VADAPT. w/ FaRL denotes CLIP initialized with FaRL, and w/o FaRL denotes VADAPT without face-specific pretraining.

Method	ACER(↓)	ACC(↑)	AUC(↑)	EER(↓)
VADAPT	**13.45±2.51**	**86.02±1.95**	**93.47±1.75**	**13.53±2.52**
w/ random T_b	18.62±3.69	81.62±3.89	89.52±3.68	18.62±3.70
w/ same T_b	20.34±0.92	80.18±0.25	87.67±1.31	20.40±0.74
w/o V_p&T_b	33.82±4.61	57.62±13.33	65.78±10.87	37.86±8.61

Table 4: Ablation study on the Visual Soft-Anchoring module. T_b and V_p denote the T_{base} and V_{part}.

15.13% for the strongest previous baseline to 13.45%, demonstrating strong generalization across physical and digital face attacks.

Results on HydraFake. We compare VADAPT with small vision models, including UniFD (Ojha et al., 2023), ProDet (Cheng et al., 2024), Co-SPY (Cheng et al., 2025), D^3 (Yang et al., 2025b), and Effort (Yan et al., 2024b), as well as recent MLLM-based detectors, including M2F2-Det (Guo et al., 2025b), SIDA (Huang et al., 2025b), FakeVLM (Wen et al., 2026), and VERITAS (Tan et al., 2025).

As shown in Table 2, VADAPT achieves the highest average accuracy among the compared small vision models. It also obtains the best Cross-Model and Cross-Domain results in this group, demonstrating strong robustness to digital forgery distribution shifts. Among the MLLM-based methods, VERITAS achieves a higher overall average accuracy, while VADAPT remains competitive with approximately 150M parameters using a compact CLIP-style ViT-B/16 backbone, without requiring large language model inference.

4.3 Ablation Study

We conduct ablation studies to examine the contributions of the main components of VADAPT, including Visual Soft-Anchoring, Dual-Branch Latent Context, and Asymmetric Feature Aggregation. Additional analyses are provided in the appendix.

Role of Face-Specific Pretraining in VADAPT. We compare CLIP, CLIP w/ FaRL, VADAPT w/o FaRL, and the full VADAPT to examine the effect of face-specific initialization.

As shown in Table 3, FaRL initialization alone does not consistently improve the CLIP baseline across all metrics. In contrast, VADAPT without FaRL already achieves strong performance, showing that the main gains come from the proposed adaptation framework rather than face-specific pretraining alone. Incorporating FaRL further improves the full model, indicating that facial-domain priors provide complementary initialization for VADAPT.

Effectiveness of Visual Soft-Anchoring. We evaluate both the visual part tokens and the textual anchoring design. As shown in Table 4, replacing the indexed anchors with randomly selected textual sentences increases ACER from 13.45% to 18.62%, indicating that arbitrary textual references do not provide the same benefit as the proposed indexed anchors. When all part tokens share the same textual anchor, the anchoring regularization becomes mathematically equivalent to removing the anchoring signal, and ACER further increases to 20.34%. Removing both V_{part} and T_{base} leads to the largest performance drop, confirming the contribution of the proposed visual part representation.

Figure 3 provides a qualitative comparison of the learned attention patterns. Without textual anchoring, several visual part tokens produce similar responses around central facial regions, whereas VSA yields more diverse attention patterns. This visualization complements the quantitative ab-

Method	ACER(↓)	ACC(↑)	AUC(↑)	EER(↓)
VADAPT	**13.45±2.51**	**86.02±1.95**	**93.47±1.75**	**13.53±2.52**
w/o multi-ctx	19.06±3.08	78.62±0.73	89.33±3.52	18.50±3.65
w/o learn-ctx	21.44±5.76	79.49±4.66	86.04±5.89	21.64±5.62

Table 5: Ablation study on the Dual-Branch Latent Context module.

Real/Fake Pool	ACER(↓)	ACC(↑)	AUC(↑)	EER(↓)
Avg/Avg	26.24±0.16	72.96±0.23	80.69±0.21	26.39±0.23
Avg/Max	25.12±4.01	72.08±6.71	82.13±4.90	25.46±4.24
Max/Max	14.93±6.10	84.57±5.88	91.41±4.40	15.18±6.09
Max/Avg	**13.45±2.51**	**86.02±1.95**	**93.47±1.75**	**13.53±2.52**

Table 6: Ablation of context-level pooling strategies for the real and fake branches.

lations and supports the role of soft anchoring in reducing redundancy among visual part tokens, without requiring a fixed or non-overlapping facial decomposition.

Analysis of Dual-Branch Latent Context. We evaluate the role of multiple learnable contexts in the text branch. As shown in Table 5, reducing the K real and M fake contexts to a single real/fake context pair (w/o multi-ctx) leads to a clear performance drop, indicating that a single prompt pair is insufficient to capture the variation within the two classes. Replacing the learnable contexts with fixed hand-crafted prompts (w/o learn-ctx) further degrades performance, demonstrating the benefit of adapting the textual representations to the task.

Although DLC does not impose an explicit diversity or orthogonality constraint among the contexts, the context visualizations in the supplementary material show non-uniform and distinguishable response patterns. Together with the multi-context ablation, these results suggest that the learned contexts capture useful and non-identical representations rather than simply collapsing to copies of a single real/fake prompt pair.

Effectiveness of Asymmetric Feature Aggregation. We further evaluate the context-level pooling strategy by comparing all combinations of Max and Average pooling for the real and fake branches. As shown in Table 6, using Max pooling for the real branch is particularly important: both variants with real-branch Average pooling perform substantially worse. With Max pooling fixed for the real branch, averaging the fake-context responses further reduces ACER from 14.93% to 13.45%. These results empirically support the proposed Max/Avg aggregation strategy.

5 Conclusion

We presented VADAPT, a visual-anchored dual-prompt tuning framework for generalizable face attack detection. The method is motivated by two observations of pretrained CLIP representations in this task: genuine and attack samples exhibit different feature structures, and the global `[CLS]` token often concentrates on similar facial regions across diverse attacks. To address these issues, Visual Soft-Anchoring encourages multiple visual part tokens to explore diverse fine-grained evidence, while Dual-Branch Latent Context and Asymmetric Feature Aggregation provide flexible real/fake context modeling with branch-specific aggregation. Experiments on UniAttackData and HydraFake demonstrate strong generalization across physical attacks, digital manipulations, and distribution shifts, including state-of-the-art performance on UniAttackData. These results show the benefit of adapting both visual evidence exploration and text-side context modeling for generalizable face attack detection.

AI use statement

In this work, we used generative AI tools for language editing, primarily to improve grammar, clarity, and wording. We also used generative AI tools for limited assistance in writing a small portion of the ablation-study code and for improving the formatting and visual presentation of figures. We did not use generative AI tools to develop the core research ideas or method, implement the main

experimental pipeline, conduct the primary experiments, generate or modify experimental results, or draw scientific conclusions; the remaining disclosure categories are not applicable to this work.

All AI-assisted content was reviewed and verified by the authors. AI-assisted code was manually inspected, tested, and executed by the authors, and all reported experimental results were obtained from the authors' own experimental pipeline. AI-assisted figure preparation was limited to visualization and presentation, and the resulting figures were checked against the underlying author-generated experimental results. All AI-edited text was reviewed for technical accuracy and consistency with the authors' intended claims. We take responsibility for the final content of this work, including all text, code, figures, claims, and other artifacts produced with the aid of generative AI.

REFERENCES

Rizhao Cai, Cecelia Soh, Zitong Yu, Haoliang Li, Wenhan Yang, and Alex C Kot. Towards data-centric face anti-spoofing: Improving cross-domain generalization via physics-based data synthesis. *International Journal of Computer Vision*, 133(4):1689–1710, 2025.

Shunxin Chen, Ajian Liu, Junze Zheng, Jun Wan, Kailai Peng, Sergio Escalera, and Zhen Lei. Mixture-of-attack-experts with class regularization for unified physical-digital face attack detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2195–2203, 2025a.

Xihong Chen, Yutong Xie, Hui Ma, Ning Li, Yanyan Liang, and Jiabao Guo. Crossgap: Unified face anti-spoofing via cross-modal global-aware prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3208–3215, 2025b.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *Advances in Neural Information Processing Systems*, 37:21979–21998, 2024.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13455–13465, 2025.

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pp. 5781–5790, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z. Li, and Zhen Lei. Surveillance face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 19: 1535–1546, 2024a. doi: 10.1109/TIFS.2023.3337970.

Hao Fang, Ajian Liu, Haocheng Yuan, Junze Zheng, Dingheng Zeng, Yanhong Liu, Jiankang Deng, Sergio Escalera, Xiaoming Liu, Jun Wan, and Zhen Lei. Unified physical-digital face attack detection, 2024b. URL https://arxiv.org/abs/2401.17699.

Jiabao Guo, Ajian Liu, Yunfeng Diao, Jin Zhang, Hui Ma, Bo Zhao, Richang Hong, and Meng Wang. Domain generalization for face anti-spoofing via content-aware composite prompt engineering. *IEEE Transactions on Multimedia*, 2025a.

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, 2025b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Xianhua He, Dashuang Liang, Song Yang, Zhanlong Hao, Hui Ma, Binjie Mao, Xi Li, Yao Wang, Pengfei Yan, and Ajian Liu. Joint physical-digital facial attack detection via simulating spoofing clues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 995–1004, 2024.

Zong-Wei Hong, Yu-Chen Lin, Hsuan-Tung Liu, Yi-Ren Yeh, and Chu-Song Chen. Domain-generalized face anti-spoofing with unknown attacks. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 820–824, 2023. doi: 10.1109/ICIP49359.2023.10223078.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Pei-Kai Huang, Jun-Xiong Chong, Cheng-Hsuan Chiang, Tzu-Hsien Chen, Tyng-Luh Liu, and Chiou-Ting Hsu. Slip: Spoof-aware one-class face anti-spoofing with language image pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3697–3706, 2025a.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025b.

Fangling Jiang, Qi Li, Weining Wang, Min Ren, Wei Shen, Bing Liu, and Zhenan Sun. Open-set single-domain generalization for robust face anti-spoofing. *International Journal of Computer Vision*, 132(11):5151–5172, 2024.

Fangling Jiang, Qi Li, Weining Wang, Wei Shen, Bing Liu, and Zhenan Sun. Learning unknown spoof prompts for generalized face anti-spoofing using only real face images: F. jiang et al. *International Journal of Computer Vision*, 134(5):250, 2026.

Seungjin Jung, Yonghyun Jeong, Minha Kim, Jimin Min, Youngjoon Yoo, and Jongwon Choi. Domain-generalizable face anti-spoofing with patch-based multi-tasking and artifact pattern conversion. *Pattern Recognition*, pp. 113640, 2026.

Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15670–15680, 2023.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 2023.

Lei Li, Zhaoqiang Xia, Xiaoyue Jiang, Yupeng Ma, Fabio Roli, and Xiaoyi Feng. 3d face mask presentation attack detection based on intrinsic image analysis. *IET Biometrics*, 9(3):100–108, 2020. doi: https://doi.org/10.1049/iet-bmt.2019.0155. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-bmt.2019.0155.

Kun-Hsiang Lin, Yu-Wen Tseng, Kang-Yang Huang, Jhih-Ciang Wu, and Wen-Huang Cheng. Instructflip: Exploring unified vision-language model for face anti-spoofing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2987–2996, 2025.

Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, Guodong Guo, Zhen Lei, Stan Z. Li, and Du Zhang. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17:2497–2507, 2022a. doi: 10.1109/TIFS.2022.3188149.

Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang, Zhen Lei, Du Zhang, Stan Z. Li, and Guodong Guo. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 18:4775–4786, 2023a. doi: 10.1109/TIFS.2023.3296330.

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

Ajian Liu, Haocheng Yuan, Xiao Guo, Hui Ma, Wanyi Zhuang, Changtao Miao, Yan Hong, Chuanbiao Song, Jun Lan, Qi Chu, et al. Benchmarking unified face attack detection via hierarchical prompt tuning. *arXiv preprint arXiv:2505.13327*, 2025.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, 2022b.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI open*, 5:208–215, 2024b.

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, 2018.

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

Hui Ma, Ajian Liu, Yongze Li, Chuanbiao Song, Xiao Guo, Changtao Miao, Wanyi Zhuang, Junze Zheng, Shunxin Chen, Yan Hong, Jiabao Guo, Jiankang Deng, Jun Lan, Weiqiang Wang, Tao Gong, Qi Chu, Sergio Escalera, Hugo Jair Escalante, Xiaoming Liu, Zhen Lei, Isabelle Guyon, Yanyan Liang, and Jun Wan. Uniattackdata+: Unified physical-digital attack detection+ challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3202–3211, October 2025.

Shiyun Mao and Huibin Li. Hyperbolic metric learning for generalizable face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2025.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24480–24489, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18720–18729, June 2022.

Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19685–19696, 2023.

Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19576–19586, 2025.

Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, and Wen-Sheng Chu. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24563–24574, June 2023.

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.

Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.10.081. URL https://www.sciencedirect.com/science/article/pii/S0925231220316945.

Changsong Wen, Zelin Peng, Yu Huang, Xiaokang Yang, and Wei Shen. Domain generalization in clip via learning with diverse text prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9559–9569, 2025.

Siwei Wen, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Conghui He, Weijia Li, et al. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *Advances in Neural Information Processing Systems*, 38:58972–59005, 2026.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8984–8994, June 2024a.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024b.

Jingyi Yang, Xun Lin, Zitong Yu, Liepiao Zhang, Xin Liu, Hui Li, Xiaochen Yuan, and Xiaochun Cao. Dadm: Dual alignment of domain and modality for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12045–12056, October 2025a.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. Dˆ 3: scaling up deepfake detection by learning from discrepancy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23850–23859, 2025b.

Jeongmin Yu, Susang Kim, Kisu Lee, Taekyoung Kwon, Won-Yong Shin, and Ha Young Kim. Multi-view slot attention using paraphrased texts for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21117–21128, 2025.

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, 2020.

Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z. Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. doi: 10.1109/TBIOM.2020.2973001.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18697–18709, June 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

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 (CVPR)*, pp. 175–187, June 2024.

N	ACER(↓)	ACC(↑)	AUC(↑)	EER(↓)
1	26.47±4.66	71.54±6.59	81.91±4.44	26.14±4.30
2	19.79±0.35	78.16±2.07	87.07±0.12	20.01±0.38
4	16.26±5.88	82.36±4.63	91.16±5.09	16.51±5.67
8	**13.45±2.51**	**86.02±1.95**	**93.47±1.75**	**13.53±2.52**
12	27.79±5.78	69.15±8.71	77.55±7.13	28.26±6.15

Table 7: Effect of the number of visual part tokens N in Visual Soft-Anchoring.

$K \backslash M$	1	2	4	8
1	19.06±3.08	21.12±2.80	21.12±1.41	26.14±6.40
2	15.13±4.19	22.48±3.38	16.50±1.39	21.14±3.40
4	22.00±8.76	**13.45±2.51**	24.12±5.73	16.27±1.88
8	18.31±3.12	15.97±1.15	19.68±5.70	19.77±5.75

Table 8: ACER for different numbers of real contexts K and fake contexts M in Dual-Branch Latent Context.

A Appendix

A.1 Additional Ablation Studies

A.1.1 Effect of the Number of Visual Part Tokens

We first study the effect of the number of visual part tokens N in Visual Soft-Anchoring. We evaluate $N \in \{1, 2, 4, 8, 12\}$ while keeping the remaining settings unchanged.

As shown in Table 7, performance improves as the number of visual part tokens increases from 1 to 8, with the best result obtained at $N = 8$. Increasing N further to 12 degrades performance, indicating that simply adding more part tokens does not necessarily improve the learned fine-grained visual representations.

A.1.2 Effect of the Number of Latent Contexts

We evaluate the numbers of real and fake contexts, K and M, over $\{1, 2, 4, 8\}$, resulting in 16 configurations.

Table 8 shows that the model is sensitive to the numbers of real and fake contexts, and the performance is not monotonic with respect to either K or M. We select $K = 4$ and $M = 2$ based on the UniAttackData validation set and use the same configuration for HydraFake without additional tuning.

A.1.3 Effect of Part-Level Aggregation

We further compare different strategies for combining the part-level logits produced after context aggregation.

As shown in Table 9, Self-Adaptive Softmax Pooling (SASP) outperforms direct averaging and concatenation followed by an MLP. This result supports adaptively weighting part-level attack evidence rather than treating all visual parts equally.

A.2 Additional Visualization and Analysis

A.2.1 Branch Response and Margin Analysis

Figure 4(a) shows that genuine samples generally produce stronger responses from the real branch, whereas attack samples tend to produce stronger responses from the fake branch. Figure 4(b) further visualizes the resulting response margin $S_{fake} - S_{real}$, illustrating how the relative evidence from the two branches contributes to the final prediction.

Method	ACER(↓)	ACC(↑)	AUC(↑)	EER(↓)
Average	16.21±5.47	82.99±4.88	90.75±4.71	16.20±5.42
Concat+MLP	15.46±1.41	83.57±2.62	92.15±0.93	15.48±1.00
SASP	**13.45±2.51**	**86.02±1.95**	**93.47±1.75**	**13.53±2.52**

Table 9: Ablation of part-level logit aggregation strategies.

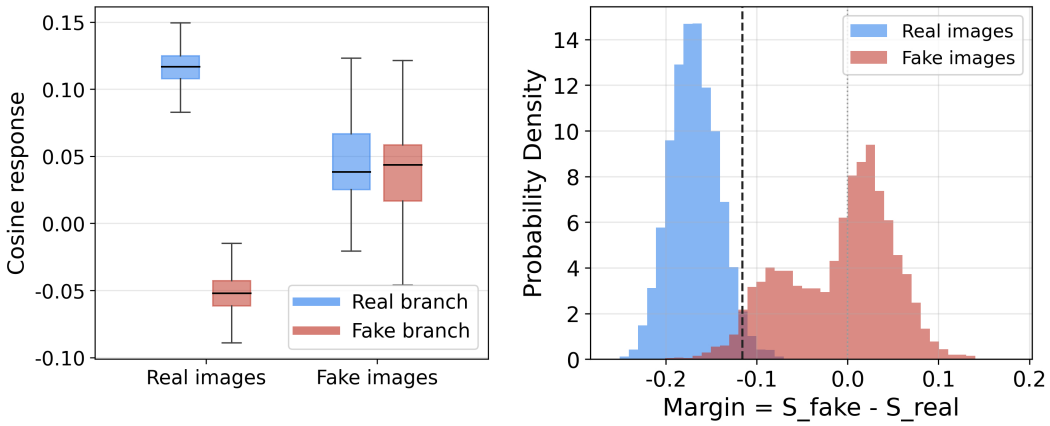


Figure 4: Branch-level response analysis. (a) Aggregate cosine responses of the real and fake branches for genuine and attack samples. (b) Distribution of the response margin $S_{fake} - S_{real}$, where larger values indicate stronger relative attack evidence.

A.2.2 Visualization of Learned Latent Contexts

We visualize the learned context-dependent representations to inspect how the multiple contexts behave in practice.

As shown in Figure 5, different context groups exhibit distinguishable patterns rather than identical behavior. Together with the multi-context ablation in the main paper, this visualization provides qualitative evidence that the learned contexts capture different task-relevant responses in practice, despite the absence of an explicit pairwise diversity or orthogonality constraint.

A.2.3 Context Response Analysis

We further examine the average responses of the learned real and fake contexts to genuine and attack samples.

As shown in Figure 6, real contexts generally respond more strongly to genuine samples, while fake contexts respond more strongly to attack samples. These response patterns are consistent with the use of the two context branches as learnable real/fake representations rather than predefined semantic categories.

A.2.4 Effect of Textual Anchoring on Response Margins

To further examine the role of the indexed textual anchors, we visualize the response-margin distribution after removing T_{base} in Figure 7. The margin is defined as $S_{fake} - S_{real}$, where larger values indicate stronger relative attack evidence. Compared with the full model shown in Figure 4(b), removing textual anchoring leads to greater overlap between genuine and attack samples. This result further supports the role of textual anchoring in improving the alignment between visual part representations and the learned real/fake contexts.

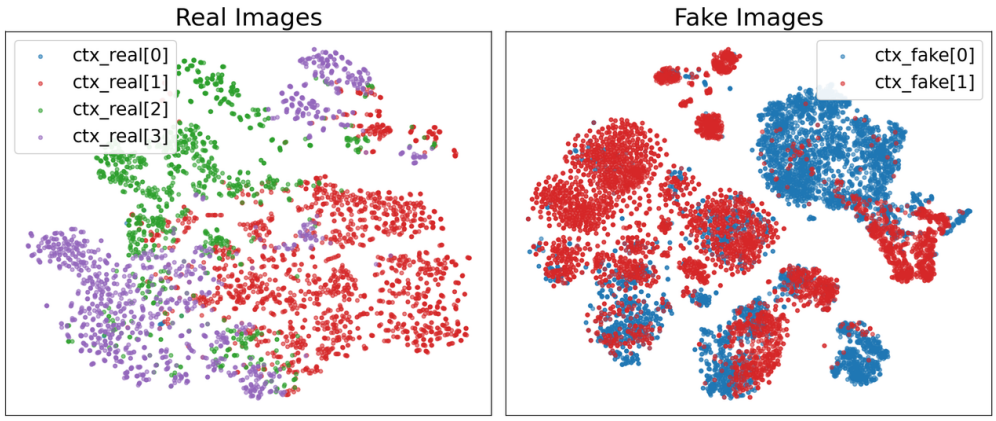


Figure 5: t-SNE visualization of representations associated with the learned real and fake contexts.

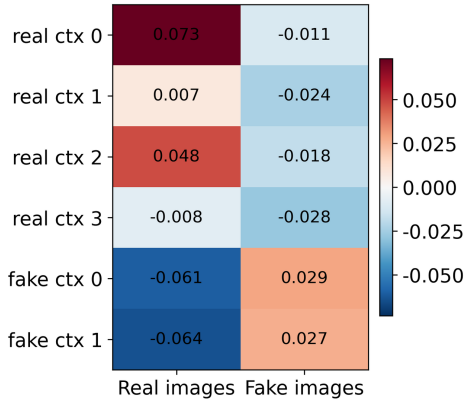


Figure 6: Average normalized responses of the learned real and fake contexts to genuine and attack samples.

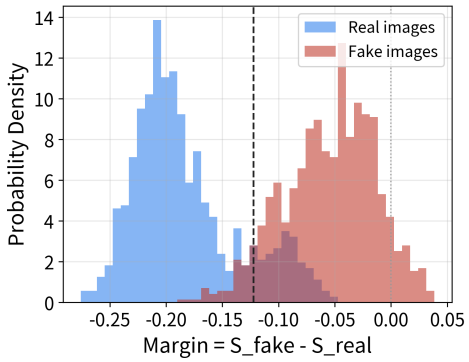


Figure 7: Response-margin distribution without textual anchoring. Removing T_{base} produces greater overlap between genuine and attack samples.