DF-PRISM: BEYOND AGGREGATE METRICS VIA VISUAL-ATTRIBUTE PROFILING FOR DEEPFAKE DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deepfake detection benchmarks typically summarize each detector with aggregate metrics such as AUC over an entire test set, but provide little guidance on which detector to select for a given image. This gap stems from a lack of systematic analysis of how detector performance varies across different image conditions. We operationalize such image conditions as visual attributes, such as illumination, background, occlusion, image stylization, and facial appearance. We present **DF-PRISM** (**D**eep**F**ake **PR**ofiling v**I**a **S**ystematic **M**ulti-attribute Analysis), a pipeline that uses a vision-language model to annotate images with 60 visual attributes and evaluates detectors on attribute-conditioned subsets, yielding a *Detector Profile* for each detector. The resulting profiles reveal systematic, detector-specific vulnerabilities: for instance, detectors trained primarily on blending artifacts fail to detect cartoon-style fake images. As a proof of concept, we use these profiles to select a subset of detectors for each image based on its annotated attributes; this adaptive selection improves macro-average AUC over an equal-weight ensemble of all 10 detectors across 13 evaluation settings. Overall, DF-PRISM complements aggregate metrics with attribute-conditioned profiling, enabling more informative benchmarking and analysis.

1 INTRODUCTION

Deepfake detection has advanced rapidly in recent years, with detectors spanning diverse architectures, training datasets, and learning paradigms, from frequency-domain and spatial artifact cues (Qian et al., 2020; Chollet, 2017) to contrastive learning (Sun et al., 2022) and generalization-oriented approaches (Yan et al., 2025). In parallel, benchmark resources have expanded steadily, from earlier datasets such as FaceForensics++ (Rossler et al., 2019) and Celeb-DF (Li et al., 2020b) to more recent evaluation platforms including DeepfakeBench (Yan et al., 2023b), DF40 (Yan et al., 2024b), ScaleDF (Wang et al., 2025), and HydraFake (Tan et al., 2025). Taken together, these developments suggest steady progress in deepfake detection when detectors are compared by aggregate benchmark metrics such as dataset-level AUC.

However, aggregate metrics provide limited guidance for the practical setting of detecting a single image. Benchmarks typically summarize each detector with one score over an entire dataset, but such dataset-level summaries do not tell us which detector is most suitable for a given image. Figure 1 illustrates this gap with a simple three-detector comparison on ScaleDF. Although the detectors have clearly different global AUCs, their relative performance on individual images, after score calibration, is much more balanced. Even the detector with the lowest AUC achieves the best calibrated prediction on 30.4% of images. In other words, higher aggregate AUC does not imply consistently better performance on individual images. This mismatch suggests that aggregate rankings alone are insufficient for deployment-oriented detector selection.

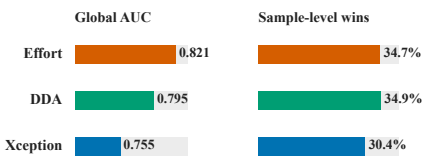

Figure 1: Higher aggregate AUC does not guarantee better per-image performance. In a three-detector comparison on ScaleDF, the lowest-AUC detector still achieves the best calibrated prediction on 30.4% of images.

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.

This gap stems from a lack of systematic analysis of how detector performance varies across different image conditions. In this work, we operationalize such conditions as *visual attributes*: directly observable properties of an image, such as illumination, background, occlusion, image stylization, and facial appearance. These attributes are closely tied to the challenges detectors face in practice, yet standard benchmark evaluations rarely characterize detector behavior along these dimensions. As a result, current benchmarks say little about when a detector performs well, what kinds of images expose its weaknesses, or how to choose among detectors for a particular test image.

To address this limitation, we present **DF-PRISM** (**D**eep**F**ake **PR**ofiling v**I**a **S**ystematic **M**ulti-attribute Analysis), a visual-attribute profiling pipeline for deepfake detectors. DF-PRISM uses a vision-language model to annotate each image with 60 fine-grained binary visual attributes, and then evaluates detectors on attribute-conditioned subsets to produce a *Detector Profile* for each detector. These profiles reveal both shared weaknesses and detector-specific vulnerabilities. For example, some attributes, such as black-and-white imagery, degrade many detectors simultaneously, whereas others expose more selective failures, such as detectors that miss cartoon-style fake images.

Our main contributions are threefold. **First**, we introduce a scalable pipeline that uses a vision-language model to annotate 60 visual attributes and evaluate each detector's attribute-conditioned performance. **Second**, we construct Detector Profiles for 10 heterogeneous pretrained detectors, providing a systematic view of detector behavior across visual attributes. **Third**, as a proof of concept, we show that these profiles enable adaptive detector selection: given an image's annotated attributes, selecting a small subset of detectors improves macro-average AUC over an equal-weight ensemble of all 10 detectors across 13 evaluation settings. **Overall**, DF-PRISM complements aggregate metrics with attribute-conditioned profiling, enabling more informative benchmarking and deployment-aware detector selection.

2 Related Work

2.1 Deepfake Detection

The deepfake detection landscape spans frequency-domain models such as F3-Net (Qian et al., 2020) and SRM (Luo et al., 2021), spatial artifact detectors such as Xception (Chollet, 2017) and EfficientNet (Tan & Le, 2019), warping-artifact detectors (Li & Lyu, 2019) and blending-based detectors (Li et al., 2020a), identity-consistency approaches (Dong et al., 2022; Huang et al., 2023), disentanglement-based methods (Yang et al., 2021; Yan et al., 2023a), reconstruction-based approaches (Cao et al., 2022) and attention-guided forgery mining (Wang & Deng, 2021), contrastive methods such as DCL (Sun et al., 2022) and SeeABLE (Larue et al., 2023), and generalization-oriented detectors including LSDA (Yan et al., 2024a), UCF (Yan et al., 2023a), SBI (Shiohara & Yamasaki, 2022), CDFA (Lin et al., 2024), DDA (Chen et al., 2025), D^3 (Yang et al., 2025), Effort (Yan et al., 2025), and ForensicsAdapter (Cui et al., 2025). This diversity producing a heterogeneous set of pretrained detectors with distinct reliability patterns. Yet most prior work summarizes each detector through aggregate benchmark scores, leaving detector behavior across directly observable visual attributes under-characterized.

2.2 Benchmarks, Attribute-Level Evaluation, and VLM Annotation

Forgery detection benchmarks and codebases, including FaceForensics++ (Rossler et al., 2019), Celeb-DF (Li et al., 2020b), DFDC (Dolhansky et al., 2020), DF40 (Yan et al., 2024b), HydraFake (Tan et al., 2025), ScaleDF (Wang et al., 2025), DeepfakeBench (Yan et al., 2023b), and ForensicHub (Du et al., 2025), provide shared testbeds for comparing detectors. However, they usually report aggregate scores at the dataset or forgery-method level, such as face swapping, reenactment, or diffusion-based synthesis. This is useful for benchmarking, but it is not available where the forgery method is unknown and only the image is directly observable. A related line of work studies demographic fairness in deepfake detection, showing that detectors can vary across race, gender, and skin tone (Nadimpalli & Rattani, 2022; Ju et al., 2024; Lin et al., 2025), but these analyses focus on a narrow attribute set and are typically framed around bias mitigation. Multi-detector fusion methods, including vanilla ensemble, weighted ensembles, and learned fusion networks (Bonettini et al., 2021), also treat all detectors uniformly rather than conditioning on the input image. Recent VLM-based forensic systems use multimodal models for detection, explanation, or structured reasoning (Jia et al., 2024; Shi et al.,

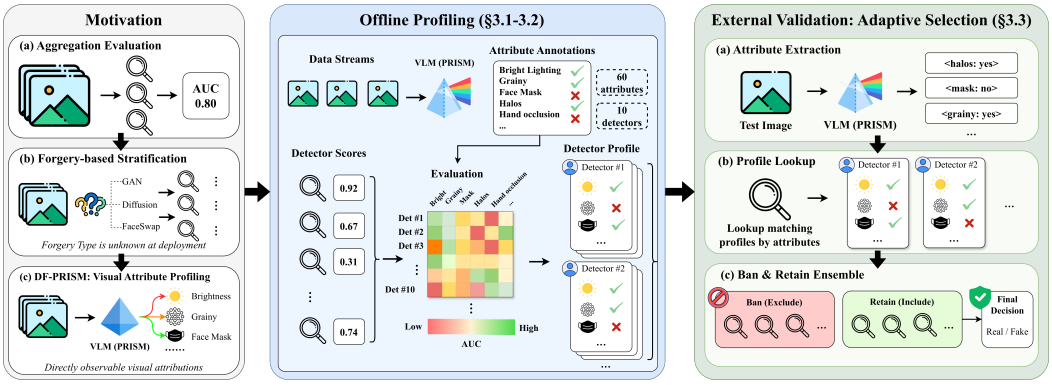

Figure 2: Overview of DF-PRISM. **Left: Motivation.** Aggregate evaluation collapses each detector to a single score and hides when or why it fails, while forgery-based stratification depends on forgery types that are unknown at deployment time. DF-PRISM instead profiles detectors through directly observable visual attributes. **Center: Offline profiling (Sec. 3.1–3.2).** A VLM annotates fine-grained visual attributes across the image corpus; detectors are run on the same images to produce detector scores; attribute labels and detector scores are then combined to construct Detector Profiles, with matched control used internally to define the attribute-absent reference and the within-detector contrast. **Right: External validation protocol (Sec. 3.3).** For a test image, the VLM extracts active attributes; detectors with low AUC on those attributes are excluded based on profile-conditioned rankings; the remaining detectors are ensembled to produce the final prediction.

2024). In contrast, DF-PRISM uses a VLM as an attribute annotator rather than as a detector or explainer, enabling attribute-conditioned profiling for constructing Detector Profiles.

3 DF-PRISM Pipeline

In this paper, we propose DF-PRISM, a pipeline for profiling individual detectors with respect to directly observable visual attributes. DF-PRISM proceeds in three stages. First, we define a vocabulary of visual attributes and annotate an image corpus (Sec. 3.1). Second, we combine the attribute annotations with detector prediction scores to construct Detector Profiles through attribute-conditioned evaluation across attribute subsets (Sec. 3.2). Third, we use these profiles in a simple external validation protocol to test whether they capture actionable performance signals for multi-detector prediction (Sec. 3.3). The first two stages form the core profiling pipeline, while the third is included only as a exploratory downstream evaluation. An overview is provided in Figure 2.

3.1 Visual Attribute Design and Annotation

Attribute Design. To profile detector behavior under observable visual conditions, we design a vocabulary of fine-grained binary visual attributes according to four principles, ensuring that the resulting attribute space suitable for profiling. First, each attribute describes a *directly observable* visual property (for instance, "Is the person wearing a face mask?" or "Is the makeup extremely heavy?"), instead of any forgery-specific artifact. This design allows the same vocabulary to be applied uniformly to both real and fake images without requiring forensic expertise. Second, all attributes are annotated using *binary yes/no* judgments, where **yes** indicates the presence of the attribute. This produces a clean $N \times K$ binary matrix that supports efficient conditional slicing. Third, Each prompt asks about one directly observable visual condition without conjunctions, reducing ambiguity within an individual annotation. Fourth, *extreme qualifiers* ("extremely," "heavily," "intensely") are leveraged to raise the activation threshold, avoiding trivially universal activation and ensuring that a positive label corresponds to a genuinely challenging visual condition. Together, these constraints define an attribute space that remains directly observable at deployment time and captures conditions under which detector performance may systematically vary.

Attribute Vocabulary and Annotation. We construct a canonical vocabulary of $K{=}60$ profiling attributes spanning six high-level categories: *Lighting* (e.g., extreme backlighting, harsh shadows), *Texture* (e.g., extreme graininess, waxy skin), *Geometry* (e.g., camera distortion, asymmetric features), *Semantics* (e.g., heavy makeup, occluding objects), *Post-Processing* (e.g., heavy compression, color banding), and *Context* (e.g., multiple faces, text overlays). We exclude attributes with trivially universal activation (activation rate $>50\%$) or insufficient supports (fewer than 50 samples per class), ensuring each attribute provides stable conditional statistics. The full attribute set is listed in Appendix A. For scalable annotation, we leverage Qwen3.5 9B served via vLLM (Kwon et al., 2023) as an attribute annotator. Each image is independently queried with each attribute prompt, and responses are parsed into a binary attribute matrix $\mathbf{A} \in \{0,1\}^{N\times K}$. Prompt templates and annotation details are provided in Appendix B. Representative visual examples of the 60 profiling attributes are provided in Appendix K.

The result of this stage is a binary attribute matrix that defines the observable condition space over which detector behavior is profiled.

3.2 Detector Profiling via Attribute-Conditioned Evaluation

Profiling Inputs. Detector profiling takes two aligned inputs. The first is the attribute matrix $\mathbf{A} \in \{0,1\}^{N\times K}$ introduced in Sec. 3.1. The second is a detector score matrix $\mathbf{S} \in \mathbb{R}^{N\times D}$, obtained by running each of the D detectors on the same image corpus to produce image-level real/fake scores and stack these outputs to form the detector score matrix. DF-PRISM constructs each Detector Profile by jointly analyzing the attribute matrix and the detector score matrix.

Three Evaluation Scores. For each detector-attribute (d, k) pair, DF-PRISM defines three complementary evaluation scores, which together form the profile body, the sensitivity layer and the aggregate reference.

AUC_{yes}: conditional AUC (primary profile entry). $\text{AUC}^d_{\text{yes}}(k)$ measures the performance of detector d on the subset of images where attribute k is active, denoted by $\mathcal{P}_k = \{x \mid \mathbf{A}_{x,k} = 1\}$. This serves as the primary entry of the profile, describing the detector's performance when the target visual condition is present. It directly supports cross-detector comparison under the same attribute, and provides the primary signal for downstream profile-guided selection. Note that $\text{AUC}_{\text{yes}}(k)$ reflects the conditional performance under the observed attribute distribution.

AUC_{no}: matched control (attribute-absent reference). $\text{AUC}^d_{\text{no}}(k)$ measures the performance of detector d on a matched control set $\mathcal{N}_k$, where attribute k is absent but the remaining annotated attributes are controlled as much as possible. Specifically, each image in $\mathcal{P}_k$ is paired with its nearest neighbor from the set $\{x \mid \mathbf{A}_{x,k} = 0\}$ using Hamming distance over the remaining $K{-}1$ attributes, with matching restricted to the same forgery label (Real↔Real, Fake↔Fake). This construction aligns $\mathcal{P}_k$ and $\mathcal{N}_k$ in label distribution and observable attribute profile, differing as much as possible only in the target attribute.

AUC_{base}: global aggregate reference. $\text{AUC}^d_{\text{base}}$ denotes the performance of detector d over the full image corpus, corresponding to the aggregate view used in conventional evaluation. It is not attribute-specific, and provides a reference for contextualizing profile entries relative to standard benchmarking.

Two Complementary Profile Comparisons. The three evaluation scores support two complementary types of comparisons.

The first is *cross-detector conditional comparison*. For a fixed attribute k, comparing detectors by $\text{AUC}^d_{\text{yes}}(k)$ reveals which detector achieves the highest AUC when k is active. This comparison drives profile-guided selection by identifying which detector performs better under each observed attribute.

The second is *within-detector attribute sensitivity*. For a fixed detector d, comparing $\text{AUC}^d_{\text{yes}}(k)$ with $\text{AUC}^d_{\text{no}}(k)$ reveals whether attribute k is associated with an increase or a decrease in performance relative to a matched baseline. We define the matched contrast as:

$$\Delta^d(k) = \text{AUC}^d_{\text{yes}}(k) - \text{AUC}^d_{\text{no}}(k). \tag{1}$$

A negative $\Delta^d(k)$ indicates lower detection AUC when attribute k is present for detector d. A positive value indicates higher AUC under the same condition.

$\text{AUC}^d_{\text{base}}$ contextualizes both comparisons by showing what conventional aggregate evaluation would report without attribute conditioning. It places the profile entries in relation to standard benchmarking practice.

Detector Profile Definition and Sensitivity Interpretation. A *Detector Profile* for detector d summarizes its behavior across visual attributes. Its primary component is the attribute-conditioned performance vector $\{\text{AUC}^d_{\text{yes}}(k)\}_{k=1}^K$, where each entry measures detector d on the subset of images where attribute k is present. This vector supports cross-detector comparison under the same visual condition and serves as the main signal for profile-guided selection. The second component, referred to as the **sensitivity layer**, consists of the matched contrasts $\{\Delta^d(k)\}_{k=1}^K$. These contrasts capture the attribute-related performance change relative to a matched reference. The third component is the **aggregate reference** $\text{AUC}^d_{\text{base}}$, corresponding to the global view used in conventional evaluation.

Based on the sensitivity contrasts, we categorize each detector-attribute pair into one of three levels. We label a pair as *strong negative sensitivity* when $\Delta^d(k) < -0.05$, indicating the attribute is associated with a substantial AUC drop. We label a pair as *strong positive sensitivity* when $\Delta^d(k) > 0.05$, indicating a notable AUC gain. All remaining pairs are labeled as *weak sensitivity*. Together, these components provide a structured view of detector behavior across visual conditions, supporting both cross-detector comparison and profile-guided selection.

3.3 External Validation Protocol

We evaluate the practical utility of Detector Profiles through a simple training-free selection protocol. This stage provides an external validation of whether the profiles constructed in Sec. 3.2 can support effective detector selection at test time.

Adaptive-Ensemble. We profile-guided detector selection through *Adaptive-Ensemble*, a minimal attribute-conditioned routing protocol. Detector ranking is based on $\text{AUC}^d_{\text{yes}}(k)$, since AUC_{yes} directly reflects detector performance under the observed attribute condition.

Step 1: Build an exclusion list per attribute (offline). For each attribute k, we rank all detectors by $\text{AUC}^d_{\text{yes}}(k)$ and exclude the bottom-B detectors (B=3 by default):

$$\text{EXCLUDE}(k) = \text{bottom-B}\big\{\text{AUC}^d_{\text{yes}}(k)\big\}_{d=1}^{D}. \tag{2}$$

This yields a simple profile-relative routing rule. When an attribute is active, detectors that perform worst under that condition are excluded from the final judgment procedure.

Step 2: Exclude and fuse (online). At test time, the same Vision-Language Model (VLM) is used to annotate the active attributes of the input image. For each active attribute, the corresponding precomputed exclusion list is retrieved. **Any** detector appearing in at least one active exclusion list is excluded by taking the union over all active attributes. The remaining detectors are then fused by averaging their normalized scores with equal weight to produce the final real/fake prediction.

4 Experiments

4.1 Experimental Setup

Datasets. We construct Detector Profiles from *ScaleDF* (Wang et al., 2025). From this source corpus, we draw a class balanced profiling set of 5.85M images with equal numbers of real and fake samples, and apply the profiling methodology (Sec. 3.2) on this balanced distribution. We use ScaleDF as the profiling source because its scale and visual diversity provide broader coverage of attribute-conditioned visual conditions than narrower alternatives.

For evaluation, we consider a comprehensive set of 13 settings: (1) *7 established benchmarks*: Celeb-DF v2 (Li et al., 2020b), DeepFakeDetection (Dufour & Gully, 2019), DeepFakeFace (Song et al., 2023), ForgeryNet (He et al., 2021), ScaleDF (test split), WildDeepfake (Zi et al., 2020), and

DF40 (Yan et al., 2024b); and (2) *6 HydraFake (Tan et al., 2025) cross-domain splits*, deepfacelab (Perov et al., 2020), dreamina (Dreamina AI team, 2025), FFIW (Zhou et al., 2021), gpt4o (Hurst et al., 2024), hailuo (Hailuo AI team, 2025), and InfiniteYou (Jiang et al., 2025).

Detectors. We profile 10 pretrained detector checkpoints: Xception (Chollet, 2017), ConvNeXt-S (Liu et al., 2022), Swin-S (Liu et al., 2021), SegFormer-B3 (Xie et al., 2021), SegFormer-B3-AvgPooling (Xie et al., 2021), ForensicsAdapter (Cui et al., 2025), DDA (Chen et al., 2025), CDFA (Lin et al., 2024), D^3 (Yang et al., 2025), and Effort-FF (Yan et al., 2025). These checkpoints are obtained from official releases or public benchmark codebases, including DeepfakeBench (Yan et al., 2023b) and ForensicHub (Du et al., 2025). They differ in backbone, training data, objectives, and training recipes. We therefore treat each checkpoint as a fixed detector instance and evaluate all detectors under the same profiling protocol. Checkpoint-source details are deferred to Appendix E.

Implementation Details. Visual attribute annotation uses Qwen3.5 9B (Qwen Team, 2026) served via vLLM (Kwon et al., 2023). We apply sigmoid to detector logits where applicable, then map each detector's scores to empirical percentiles using its score distribution on the ScaleDF training set. This mapping is fixed for all evaluation sets and yields a shared $[0, 1]$ score scale across detectors. All profiling analyses and ensemble predictions use the mapped scores.

4.2 Cross-VLM Annotation Consistency

VLM-based annotation scales attribute labeling by leveraging broad visual priors, but it depends on whether the resulting labels are stable. We therefore use Gemma-4-26B-A4B (Google DeepMind, 2026) as an independent annotator and compare its responses with Qwen3.5 9B. We randomly sample 500 real and 500 fake images from ScaleDF and measure the per-image Hamming distance between their 60-attribute vectors from each model.

The resulting attribute labels by different VLMs present consistent distributions. Under the natural image distribution, the two models disagree on a median of only **2 out of 60 attributes** per image. **49 attributes** (82%) achieve agreement $\geq$0.90. High-agreement attributes are visually unambiguous binary concepts, such as *baby*, *black-and-white*, and *solid background*. Lower-agreement attributes concentrate on threshold-sensitive visual judgments, such as *rotation borders* and *shadow severity*, where each model applies its own implicit threshold. Full per-attribute agreement details are provided in Appendix I.

Table 1: Cross-VLM annotation consistency between Qwen3.5 9B and Gemma-4-26B on randomly sampled ScaleDF images. Hamming distance is computed over 60 binary attributes per image. Lower is better.

Group	N	Mean	Median
Random-Real	500	2.17	2
Random-Fake	500	2.86	2

Human annotation audit. Two annotators independently labeled all 60 attributes on 200 class-balanced ScaleDF images (100 real and 100 fake), without seeing Qwen annotations or authenticity labels. Qwen agreed with the two human annotators on 94.40% and 94.45% of the original 60-attribute judgments, respectively. For the 50 attributes retained in the Adaptive-Ensemble evaluation, the corresponding agreement rates are 95.60% and 95.91%. Agreement varies across attributes, and these overall rates should be interpreted in light of the rarity of positive labels. The original 60-attribute list appears in Appendix A, with the 10 filtered attributes shown in gray. The human labeling procedure is illustrated in Appendix I.

4.3 Detector Sensitivity Analysis

We next examine what VLM-derived attributes reveal about detector performance. If aggregate AUC fully described detector behavior, performance would be similar across visual attributes.

Instead, we observe large variance in detector performance across attributes, indicating that visual conditions reveal aspects of detector behavior that are hidden by aggregate evaluation. We quantify this effect using the matched contrast $\Delta\mathrm{AUC} = \mathrm{AUC}_{\mathrm{yes}} - \mathrm{AUC}_{\mathrm{no}}$: negative values indicate that the active attribute is associated with lower detector AUC, whereas positive values indicate higher detector AUC. Figure 3 visualizes this sensitivity landscape. We summarize the main observations

in three parts: shared performance drops across many detectors, detector-specific sensitivities, and attribute-specific strengths or stability.

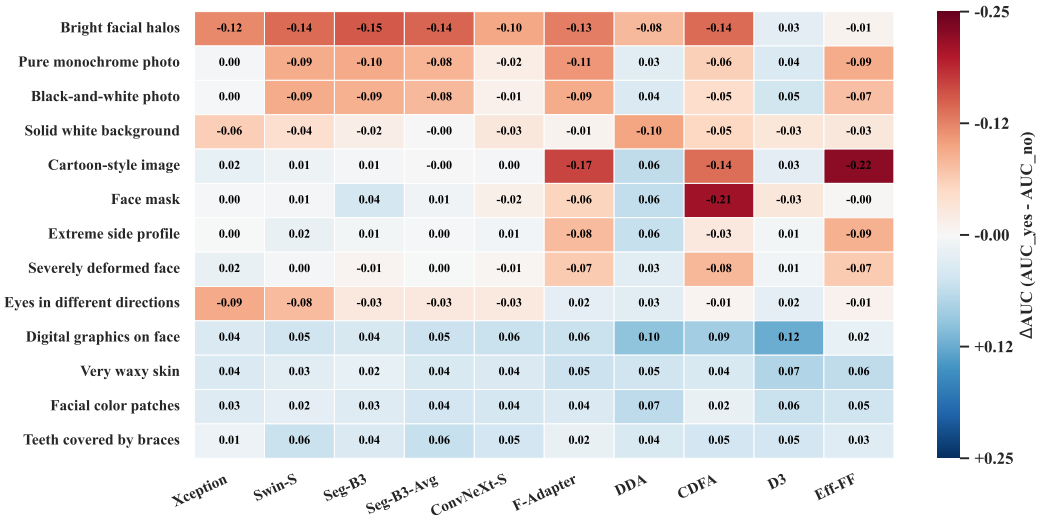

Figure 3: Attribute sensitivity landscape on ScaleDF. Each cell reports $\Delta\mathrm{AUC}$ for one detector-attribute pair, computed between attribute-present images and nearest-neighbor matched attribute-absent controls. Red indicates degradation and blue indicates amplification. Attributes are grouped into shared difficulties, detector-specific sensitivities, and positive cues; the full 60-attribute landscape is provided in Appendix D.

Shared Performance Drops. We observe broad degradation when color information is suppressed: monochrome and black-and-white images reduce the performance of many detectors. A plausible explanation is that these attributes weaken color-related cues that detectors trained on RGB face crops may exploit, such as skin-tone inconsistencies and color-blending artifacts around manipulated regions. This does not imply that color is the only evidence used by these detectors; rather, it suggests that some detectors partially depend on such cues, making them less effective when color information is removed.

Detector-Specific Sensitivities. Most degradation patterns are detector-specific rather than shared: under the $\Delta\mathrm{AUC} < -0.05$ threshold, only 6 of 60 attributes strongly degrade three or more detectors. Cartoon-style images provide a concrete example of this detector-specific behavior: Effort-FF, CDFA, and Forensics Adapter show large performance drops under this attribute, while DDA improves on the same subset, creating a gap of over 0.38 in $\mathrm{AUC}_{\mathrm{yes}}$ between the weakest and strongest detectors.

Occlusion reveals a different form of detector-specific weakness. Face masks most strongly degrade CDFA and also reduce Forensics Adapter, while leaving most other detectors largely unaffected. Unlike cartoon stylization, this failure is tied to missing visual evidence rather than a shift in image style. Face masks occlude the lower face, where standard face-swap artifacts often appear, including blending boundaries, mouth-region inconsistencies, and skin-tone transitions. A detector that relies on these localized cues can benefit from them in unobstructed face-swap detection, but the same regional bias becomes brittle when the region is covered. Thus, a useful prior under standard conditions can become a failure mode under occlusion.

Together, these examples show that visual attributes should not be interpreted as universal difficulty labels. Instead, they act as interaction variables between image conditions and detector-specific learned cues: the same attribute can suppress one detector's evidence while leaving another detector stable or even improving it.

This complementarity extends beyond individual examples. Across all 60 attributes, DDA achieves higher $\mathrm{AUC}_{\mathrm{yes}}$ than Effort-FF on 36 attributes, while Effort-FF leads on the remaining 24. The split is structured rather than random: DDA is stronger on non-photorealistic or strongly perturbed facial appearances, whereas Effort-FF is stronger on more conventional photographic variations.

This suggests that Detector Profiles capture qualitatively different detector priors, rather than simply re-ranking detectors by AUC.

This interaction is exactly what a profile-guided selector can exploit. When an attribute is active, the profile indicates which detectors are likely to perform poorly under that condition, instead of assuming that all detectors perform equally well under every attribute.

Attribute-Specific Strengths and Stability. Detector Profiles also reveal attribute-specific strengths. Some attributes increase the performance of particular detectors rather than degrading them. For example, digital face graphics increase D^3 performance by $\Delta\mathrm{AUC} = +0.124$, suggesting that unusual facial overlays may provide discrepancy cues that this detector can exploit. Attributes such as waxy skin and color patches similarly produce positive sensitivity for some detectors. These cases suggest that future routing strategies could go beyond excluding weak detectors; they could also actively favor detectors whose profiles indicate strengths under the active attributes.

D^3 further illustrates how profiles capture stability, not just amplification. Across all 60 attributes, it has no strong degradation under the $\Delta\mathrm{AUC} < -0.05$ threshold, and its worst negative sensitivity is only -0.034. This pattern is consistent with its discrepancy-based design, which aims to suppress generator-specific shortcuts and rely on more transferable discrepancy cues. As a result, D^3 is less easily disrupted by individual visual attributes than detectors whose cues are more tied to specific appearance conditions.

4.4 Evaluation via Profile-Guided Selection

The sensitivity analysis above shows that Detector Profiles contain structured performance differences across visual attributes. We next provide useful signals for a simple rule: can the profiles guide detector selection across the 13 evaluation settings? To test this, we use the ScaleDF-derived $\mathrm{AUC}_{\mathrm{yes}}$ rankings in the simple Adaptive-Ensemble protocol, which excludes detectors ranked as weak under the active attributes of each test image.

Table 2 reports AUC across 13 evaluation settings: 7 established benchmarks and 6 HydraFake cross-domain splits. We compare the 10 individual detectors, a simple but strong ensemble baseline (**Vanilla-Ensemble**, the equal-weight average of all detector scores), and **Adaptive-Ensemble**, which uses ScaleDF-derived Detector Profiles to exclude detectors with low profile scores under the active attributes.

Table 2: AUC across 13 evaluation settings. Upper rows report individual detectors; lower rows report ensemble methods. Adaptive-Ensemble uses 50 routing attributes. Underlining marks the best result in each column, ranked by unrounded AUC.

	Established Benchmarks							HydraFake Cross-Domain						
Detector	CDFv2	DFD	DFF	ForgNet	ScaleDF	WildDF	DF40	deepfcl	dreamina	FFIW	gpt4o	hailuo	InfYou	**Avg.**
CDFA (Lin et al., 2024)	0.936	0.732	0.561	0.755	0.723	<u>0.865</u>	0.741	0.797	0.836	0.743	0.341	0.635	0.747	0.724
ConvNeXt-S (Liu et al., 2022)	0.839	0.721	0.682	0.674	0.777	0.786	0.871	0.754	0.884	0.743	0.824	0.858	0.991	0.800
D^3 (Yang et al., 2025)	0.660	0.697	0.906	0.663	0.819	0.736	0.814	0.678	0.872	0.648	0.725	0.542	0.929	0.745
DDA (Chen et al., 2025)	0.550	0.585	0.873	0.596	0.838	0.624	0.803	0.675	<u>0.972</u>	0.590	<u>0.893</u>	<u>0.976</u>	0.985	0.766
Effort-FF (Yan et al., 2025)	0.900	0.768	0.863	<u>0.808</u>	0.830	0.855	0.908	<u>0.933</u>	0.495	0.827	0.412	0.629	0.642	0.759
ForensicsAdapter (Cui et al., 2025)	0.921	0.743	0.787	0.765	0.762	0.824	0.838	0.903	0.502	0.794	0.496	0.549	0.601	0.730
SegFormer-B3 (Xie et al., 2021)	0.823	0.643	0.587	0.641	0.693	0.745	0.905	0.742	0.722	0.737	0.726	0.701	0.979	0.742
SegFormer-B3-Avg (Xie et al., 2021)	0.873	0.680	0.618	0.668	0.724	0.743	0.864	0.768	0.750	0.744	0.722	0.689	0.962	0.754
Swin-S (Liu et al., 2021)	0.827	0.645	0.659	0.662	0.722	0.793	0.853	0.733	0.868	0.752	0.773	0.808	0.992	0.776
Xception (Chollet, 2017)	0.808	0.620	0.590	0.625	0.721	0.768	0.840	0.681	0.791	0.649	0.704	0.706	0.876	0.721
Vanilla-Ensemble	<u>0.940</u>	0.785	0.877	0.755	0.847	0.856	0.938	0.890	0.925	0.835	0.791	0.865	0.993	0.869
Adaptive-Ensemble	0.931	<u>0.792</u>	<u>0.919</u>	0.750	<u>0.868</u>	0.856	<u>0.939</u>	0.903	0.938	<u>0.841</u>	0.827	0.887	<u>0.995</u>	<u>0.880</u>

Analysis. The results show that aggregate detector rankings are highly unstable across evaluation settings. No individual detector is uniformly strong: the best single detector by average AUC, ConvNeXt-S, reaches 0.800 on average but still drops to 0.674 on ForgeryNet, while detectors that are strong on some domains can fail sharply on others. Vanilla-Ensemble is therefore a strong baseline, improving the average AUC to 0.869 by pooling all 10 detectors.

Adaptive-Ensemble improves macro-average AUC from 0.869 to 0.880 (+1.1%). The largest gains over Vanilla-Ensemble occur on DFF (+4.2 %), gpt4o (+3.6 %), and hailuo (+2.2 %). These results suggest that ScaleDF-derived profiles can provide useful selection signals. Adaptive-Ensemble trails Vanilla-Ensemble by 1.0 % and 0.5 % on CDFv2 and ForgeryNet, respectively, indicating that ScaleDF-derived rankings do not transfer uniformly. A post-hoc diagnostic shows that CDFA is frequently excluded despite its strong target-domain performance; retaining it improves AUC on both datasets (Appendix F). We therefore treat Adaptive-Ensemble as a downstream probe of profile utility rather than a universally reliable router.

5 Conclusion

We presented DF-PRISM, a profiling pipeline that characterizes deepfake detector performance over directly observable visual attributes. The resulting Detector Profiles reveal behavior hidden by aggregate AUC: a small set of attributes exposes shared blind spots across many detectors, while many others reveal detector-specific sensitivities and attribute-specific strengths. These profiles therefore provide a more detailed view of detector behavior than dataset-level scores alone. As a proof of concept, a simple profile-guided ensemble improves macro-average AUC over equal-weight fusion across the 13 evaluation settings.

Limitations and Future Work. DF-PRISM uses VLM-derived attribute labels as operational variables for profiling, supported by a human annotation audit rather than treated as error-free ground truth. The audit informed the choice of attributes used in the routing experiment, but annotation errors may remain. As stronger VLMs become available, the labels and resulting Detector Profiles can be updated and reassessed with human annotations. Our human-designed attribute vocabulary is a diagnostic starting point rather than an exhaustive description of visual conditions.

Appendix Organization. Appendix A lists the complete 60-attribute vocabulary. Appendix B provides the VLM prompt templates and inference configuration. Appendix C reports attribute activation statistics on ScaleDF. Appendix D provides the full 10×60 $\Delta\mathrm{AUC}$ sensitivity heatmap. Appendix E documents checkpoint origins and training configurations for all 10 detectors. Appendix F analyzes the CDFv2 and ForgeryNet routing losses. Appendix G ablates the exclusion threshold B across all 13 benchmarks. Appendix H lists the top-5 negative and positive sensitivities for each detector. Appendix I reports cross-VLM annotation agreement. Appendix J reports a per-stage GPU-hour breakdown for the full pipeline. Appendix K shows one representative visual example per attribute.

Ethics Statement

This work aims to identify gaps in deepfake detector evaluation and support defensive auditing, stronger benchmarks, and more reliable detection. The two human annotators were aware that the images included both real and manipulated faces, their task was to judge general visual attributes, not image authenticity. Reporting detector-specific weaknesses could also help malicious actors evade detection. We present these profiles to support detector improvement and deployment risk assessment while acknowledging this potential misuse.

AI Use Statement

We used Qwen3.5-9B to annotate visual attributes and Gemma-4-26B-A4B-IT to assess annotation consistency, as described in the paper. Generative AI tools also assisted with writing and editing code and polishing the manuscript for clarity.

Reproducibility Statement

The main text describes the datasets, detector profiling procedure, and attribute-guided selection rule. Appendices A and B provide the visual attribute definitions and annotation prompts; Appendix E documents detector checkpoints and preprocessing; Appendices G and J report additional results and computational resources. We plan to release the attribute annotations, Detector Profiles, and evaluation code once paper accepted

REFERENCES

Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5012–5019. IEEE, 2021.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4122, 2022.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025. NeurIPS 2025 Spotlight.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting CLIP for generalizable face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 19207–19217, June 2025.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9468–9478, 2022.

Dreamina AI team. Dreamina. `https://jimeng.jianying.com`, 2025. Accessed: 2025-07-25.

Bo Du, Xuekang Zhu, Xiaochen Ma, Chenfan Qu, Kaiwen Feng, Zhe Yang, Chi-Man Pun, Jian Liu, and Ji-Zhe Zhou. Forensichub: A unified benchmark & codebase for all-domain fake image detection and localization. *arXiv preprint arXiv:2505.11003*, 2025. NeurIPS 2025 Datasets and Benchmarks Track.

Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. Google Research Blog, 2019. URL `https://research.google/blog/contributing-data-to-deepfake-detection-research/`.

Google DeepMind. Gemma 4 model card. Google AI for Developers, 2026. URL `https://ai.google.dev/gemma/docs/core/model_card_4`.

Hailuo AI team. Hailuo video. `https://hailuoai.com/video/create`, 2025. Accessed: 2025-07-25.

Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4324–4333, 2024.

Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infiniteyou: Flexible photo recrafting while preserving your identity. *arXiv preprint arXiv:2503.16418*, 2025.

Yan Ju, Shu Hu, Shan Jia, George H. Chen, and Siwei Lyu. Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4655–4665, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21011–21021, 2023.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020a.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020b.

Li Lin, Santosh Santosh, Mingyang Wu, Xin Wang, and Shu Hu. Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3503–3515, 2025.

Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *European Conference on Computer Vision*, pp. 104–122. Springer, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9992–10002, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16317–16326, 2021.

Aakash Varma Nadimpalli and Ajita Rattani. Gbdf: Gender balanced deepfake dataset towards fair deepfake detection. In *International Joint Conference on Biometrics*, 2022.

Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer, 2020.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://huggingface.co/Qwen/Qwen3.5-9B`.

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 Conference on International Conference on Computer Vision*, 2019.

Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *arXiv preprint arXiv:2402.04178*, 2024.

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

Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. Robustness and generalizability of deepfake detection: A study with diffusion models. *arXiv preprint arXiv:2309.02218*, 2023.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2316–2324, 2022.

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.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14923–14932, 2021.

Wenhao Wang, Longqi Cai, Taihong Xiao, Yuxiao Wang, and Ming-Hsuan Yang. Scaling laws for deepfake detection. *arXiv preprint arXiv:2510.16320*, 2025.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22412–22423, 2023a.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. *Advances in Neural Information Processing Systems*, 36:4534–4565, 2023b.

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*, pp. 8984–8994, 2024a.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: toward next-generation deepfake detection. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 29387–29434, 2024b.

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. In *International Conference on Machine Learning*, 2025.

Tianyun Yang, Juan Cao, Qiang Sheng, Lei Li, Jiaqi Ji, Xirong Li, and Sheng Tang. Learning to disentangle gan fingerprint for fake image attribution. *arXiv preprint arXiv:2106.08749*, 2021.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D^3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23850–23859, 2025.

Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5778–5788, 2021.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2382–2390, 2020.

Appendix Overview

The appendix is organized into four thematic groups:

- **Pipeline Details** - materials describing the visual attribute vocabulary and the VLM annotation procedure.
  - Appendix A: Profiling Attribute List - the complete set of 60 binary visual attributes.
  - Appendix B: Prompt Templates - VLM annotation prompt format, example, and inference configuration.
  - Appendix C: Attribute Annotation Statistics - activation frequency distribution on ScaleDF.
  - Appendix K: Visual Examples of Profiling Attributes - one representative image per attribute.
- **Profiling Results** - extended detector profile outputs.
  - Appendix D: Complete Detector Profile Table - full 10×60 $\Delta\mathrm{AUC}$ sensitivity heatmap.
  - Appendix H: Per-Detector Sensitivity Profiles - top-5 negative and positive sensitivities for each detector.
- **Extended Analyses** - additional experimental results and failure diagnostics.
  - Appendix I: Annotation Consistency Analysis - cross-VLM Hamming distance and per-attribute agreement.
  - Appendix F: Transfer Diagnosis - CDFA exclusion and post-hoc retention analysis.
  - Appendix G: Ablation Study - Adaptive-Ensemble (B) ablation with full per-benchmark results across all 13 settings.
  - Appendix E: Detector Checkpoint Sources - checkpoint origin and training configuration for all 10 detectors.
- **Reproducibility** - compute resource summary.
  - Appendix J: Compute Resources - per-stage GPU-hour breakdown for the full DF-PRISM pipeline.

A PROFILING ATTRIBUTE LIST

We define a canonical set of $K=60$ binary visual attributes spanning directly observable visual factors at inference time. Each attribute is formulated as a single yes/no question. Some questions use qualifiers such as *extremely* or *heavily* to identify pronounced visual conditions. Attributes are selected to avoid trivially universal activation and statistically sparse conditions. Table 3 lists all 60 profiling attributes; 50 are selected by human judgment for Adaptive-Ensemble routing.

Table 3: Canonical profiling attribute list ($K=60$). Each entry is a binary yes/no visual question.

ID	VLM Query (Yes/No)	ID	VLM Query (Yes/No)
01	Is the facial skin extremely smooth?	31	Is the person completely bald?
02	Is the makeup extremely heavy?	32	Is the person wearing a hat?
03	Does the face have many different color patches?	33	Is the face blocked by an object?
04	Does the image have black rotation borders?	34	Is the face heavily covered in shadows?
05	Are there bright halos around the facial outlines?	35	Is the facial structure completely deformed?
06	Is the mouth opened extremely wide?	36	Are there heavy shadows on the face?
07	Is the tongue heavily exposed?	37	Is the person wearing a face mask?
08	Are the facial muscles intensely squeezed?	38	Are there digital graphics on the face?
09	Is the chin cut off by the image border?	39	Are the lips tightly puckered?
10	Are the ears cut off by the image border?	40	Is the forehead completely covered by thick hair?
11	Does the facial skin look extremely waxy?	41	Is the forehead completely covered by a hat?
12	Does the background have a strong bokeh effect?	42	Is this a purely monochromatic photo?
13	Are the background colors extremely vivid?	43	Is this image cartoon-style?
14	Is the background completely solid black?	44	Is the face seen from an extreme side profile?
15	Is the background completely solid white?	45	Are there hands or fingers covering parts of the face?
16	Does the background contain many different colors?	46	Are there multiple faces clearly visible in the image?
17	Is this a formal passport-style photo?	47	Does the image have an extreme color tint?
18	Is there scrambled text in the image?	48	Is the person making an intensely exaggerated facial expression?
19	Is there a prominent logo in the image?	49	Are there intense light reflections or glare on the eyes?
20	Is this a black-and-white photo?	50	Are the eyes looking in completely different directions?
21	Are the teeth completely covered by heavy braces?	51	Are the person's eyes completely closed?
22	Are there huge earrings covering the facial outline?	52	Does the image have visible physical damage?
23	Are large headphones completely covering the ears?	53	Is the person wearing glasses?
24	Is the image in extremely bright light?	54	Does the face have distinct freckles?
25	Is the image in extreme darkness?	55	Is the head heavily tilted to one side?
26	Is the face covered in deep wrinkles?	56	Is the person holding something in their mouth?
27	Does the person look extremely old?	57	Is the inside of the mouth pitch black?
28	Is the person an extremely young baby?	58	Are the lips painted with bright lipstick?
29	Is eyewear blocking the eyes?	59	Is the face illuminated by colored lights?
30	Is there extremely strong backlighting?	60	Is the person looking downwards?

Upon acceptance, we will release the code, including the canonical attribute definitions and annotation prompts used in this paper.

B PROMPT TEMPLATES

Overview. Each attribute annotation query is an independent, single-turn visual question posed to Qwen3.5 9B. No system prompt is used; the entire instruction is contained in the user message shown below. Each image is queried once per attribute; attributes are processed independently to avoid inter-attribute interference.

Prompt Format. The user message follows a fixed template, populated with the per-attribute question from Appendix A:

```
/no_think
Look at the image carefully.
You must answer with valid JSON only.
Return exactly one of these two objects and nothing else:
{"answer":"YES"}
{"answer":"NO"}
Do not output any explanation, markdown, or extra text.
Question: <attribute question>
```

The `/no_think` prefix is a model-level directive that disables Qwen3.5 9B's chain-of-thought reasoning mode. This is intentional: for simple binary visual judgements, extended reasoning tends to introduce over-analysis and higher label flip rates relative to direct visual assessment.

Example (attr_05).

```
Look at the image carefully.
You must answer with valid JSON only.
Return exactly one of these two objects and nothing else:
{"answer":"YES"}
{"answer":"NO"}
Do not output any explanation, markdown, or extra text.
Question:  Are there bright halos around the facial
outlines?
```

Model output: `{"answer":"NO"}`

JSON Output Schema. The response format is constrained to a JSON object via the inference engine's `response_format={"type":"json_object"}` parameter. Valid outputs are `{"answer":"YES"}` or `{"answer":"NO"}`; any response that cannot be parsed as one of these two objects is treated as an annotation failure and flagged for re-annotation. In practice, the JSON failure rate on ScaleDF was below 0.1%.

Inference Configuration. Key inference parameters are listed in Table 4.

Table 4: VLM inference configuration for attribute annotation.

Parameter	Value
Model	Qwen3.5 9B
Precision	bfloat16
Serving backend	vLLM (swift_vllm)
Temperature	0.0 (deterministic)
Max new tokens	32
Response format	`json_object`
Max image tokens	1,024
Max pixel resolution	1,003,520 px
Batch size	64 images per batch

C Attribute Annotation Statistics

Figure 4 shows the visual attribute activation frequency on the ScaleDF training split for all 60 profiling attributes. Attributes are sorted by matched sample count. The long-tail distribution confirms that most attributes are rare-to-moderate in frequency, preventing trivially universal activation.

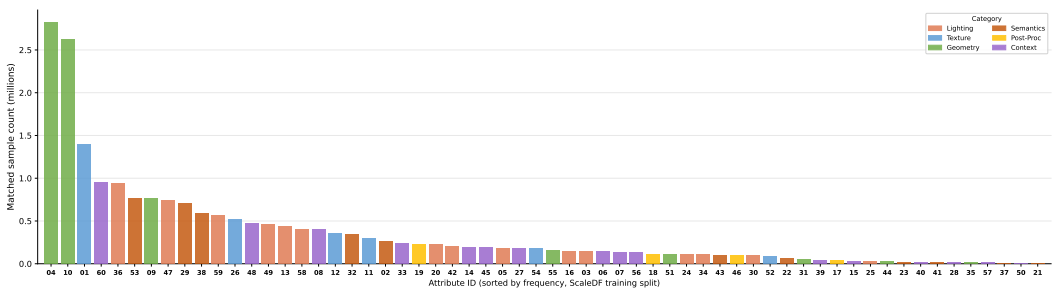

Figure 4: Visual attribute activation frequency on ScaleDF (60 profiling attributes, sorted by matched sample count). Colors indicate macro-category.

D Complete Detector Profile Table

Figure 5 presents the complete 10×60 attribute sensitivity landscape on ScaleDF, extending Figure 3 in the main paper from 13 representative attributes to all 60 profiling attributes. Each cell shows the within-detector sensitivity contrast $\Delta\mathrm{AUC} = \mathrm{AUC}_{\mathrm{yes}} - \mathrm{AUC}_{\mathrm{no}}$; separately, $\mathrm{AUC}_{\mathrm{yes}}$ values support cross-detector comparison under each condition. Attributes are sorted left-to-right by mean $|\Delta\mathrm{AUC}|$ across all detectors, placing the most diagnostically informative attributes first. Detectors are sorted top-to-bottom by baseline AUC. Cells with $|\Delta\mathrm{AUC}| \geq 0.05$ are annotated. This matched-control construction controls for the annotated visual attributes, but unannotated visual factors may still remain.

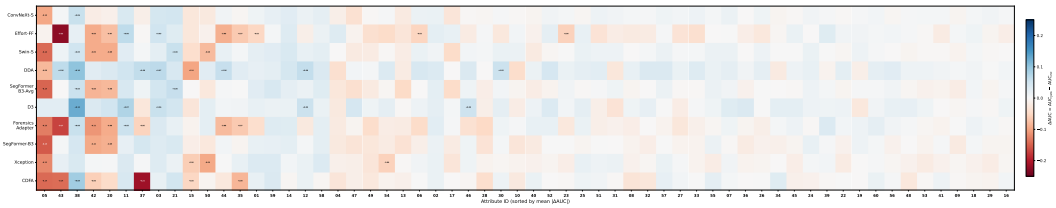

Figure 5: Full attribute sensitivity landscape on ScaleDF (10 detectors $\times$ 60 attributes). $\Delta\mathrm{AUC} = \mathrm{AUC}_{\mathrm{yes}} - \mathrm{AUC}_{\mathrm{no}}$ is the within-detector sensitivity contrast under nearest-neighbour covariate control. Red indicates degradation sensitivity; blue indicates amplifying sensitivity. Columns sorted by mean $|\Delta\mathrm{AUC}|$ (left: highest-impact attributes).

E Detector Checkpoint Sources

Table 5 lists the checkpoint source, for each of the 10 detectors used in our evaluation. All detectors are evaluated *without* fine-tuning on ScaleDF or HydraFake; their weights are loaded directly from the sources listed below.

Table 5: Detector checkpoint sources.

Detector	**Checkpoint Source**
Xception	DeepfakeBench (Yan et al., 2023b)
ConvNeXt-S	ForensicHub
Swin-S	ForensicHub
SegFormer-B3	ForensicHub
SegFormer-B3-Avg	ForensicHub
ForensicsAdapter	Official repository
DDA	Official repository
CDFA	Official repository
D3	Official repository (CVPR 2025)
Effort-FF	Official repository (ICML 2025)

Preprocessing. All face crops, alignment, and score extraction follow the **DeepfakeBench** preprocessing pipeline (Yan et al., 2023b) to ensure consistent input resolution and face-region alignment across all detectors. Detector logits are converted to scores via sigmoid where applicable. The resulting scores are mapped to empirical percentiles using a per-detector score distribution computed on the ScaleDF training set; the mappings are held fixed for all evaluation sets.

F Profile-Guided Selection: Transfer Diagnosis

Adaptive-Ensemble trails Vanilla-Ensemble on CDFv2 by 1.0 percentage point and on ForgeryNet by 0.5 percentage points. No image on either benchmark has an empty retained detector set under the default B=3 rule. We inspect the excluded detectors to understand these losses.

CDFA is the best individual detector on CDFv2 (AUC 0.936), yet the selector excludes it on 57.3% of images. On ForgeryNet, CDFA ranks third (AUC 0.755) and is excluded on 89.1% of images. In the ScaleDF-derived profiles, CDFA ranks last by $\mathrm{AUC}_{\mathrm{yes}}$ on 36 of the 50 routing attributes and falls in the bottom three on all 50. Thus, the union of per-attribute exclusion lists frequently removes a detector that performs well on these target domains.

Post-hoc CDFA retention diagnostic. To assess the effect of these exclusions, we force CDFA into the retained set on the two benchmarks (Table 6). This intervention uses knowledge of target-domain performance and is only a diagnostic. It raises Adaptive-Ensemble AUC from 0.931 to 0.944 on CDFv2 and from 0.750 to 0.765 on ForgeryNet. The recovery supports CDFA over-exclusion as a substantial contributor to the losses, while other transfer effects may also be present.

Table 6: Post-hoc CDFA retention diagnostic (AUC). Retaining CDFA uses target-domain knowledge and is not part of Adaptive-Ensemble.

Benchmark	Vanilla-Ensemble	Adaptive-Ensemble	Retain CDFA
CDFv2	0.940	0.931	**0.944**
ForgeryNet	0.755	0.750	**0.765**

This example shows how Detector Profiles can help audit routing failures: the source-domain rankings identify which exclusions conflict with a detector's target-domain strengths. The current selector does not account for shifts in baseline detector performance between profiling and evaluation datasets.

G Ablation Study

We report the full per-benchmark breakdown for the Adaptive-Ensemble ablation. All routing experiments use 50 attributes from the full 60-attribute ScaleDF Detector Profile and are evaluated across 7 benchmarks (CDFv2, DFD, DFF, ForgeryNet, ScaleDF, WildDF, DF40) and 6 cross-domain generators (DeepFaceLab, Dreamina, FFIW, GPT-4o, Hailuo, InfiniteYou). **Bold** and underlined mark the best and second-best results in each column, ranked by unrounded AUC.

(a) Number of Excluded Detectors B. B controls how many lowest-ranked detectors are excluded per attribute group. B=1 provides minimal routing; larger B increases the aggressiveness of exclusion.

Table 7: Ablation on excluded detector count B (AUC, higher is better).

B	CDFv2	DFD	DFF	FN	ScDF	WildDF	DF40	DFL	Drm	FFIW	GPT	Hail	InfY	**Avg**
Vanilla-Ensemble	**0.940**	0.785	0.877	**0.755**	0.847	0.856	0.938	0.890	0.925	0.835	0.791	0.865	0.993	0.869
1	0.929	0.783	0.893	0.745	0.850	0.845	0.940	0.889	0.916	0.832	0.808	0.871	0.994	0.869
2	0.931	0.788	0.907	0.755	0.857	0.852	**0.941**	0.894	0.924	0.839	0.818	0.871	0.995	0.875
3	0.931	**0.792**	**0.919**	0.750	**0.868**	**0.856**	0.939	**0.903**	0.938	**0.841**	0.827	0.887	0.995	**0.880**
4	0.924	0.783	0.915	0.734	0.866	0.847	0.936	0.888	0.954	0.827	**0.840**	**0.900**	**0.996**	0.878
5	0.914	0.777	0.896	0.726	0.859	0.842	0.936	0.871	**0.955**	0.811	0.828	0.868	0.994	0.867

Summary. Adaptive-Ensemble achieves a macro-average AUC of 0.880 at B=3, compared with 0.869 for Vanilla-Ensemble. Increasing exclusion beyond B=3 reduces the macro-average AUC, showing the sensitivity of hard exclusion to this choice. Under the default B=3 rule, no image in the 13 evaluation settings produces an empty retained set.

H Per-Detector Sensitivity Profiles

For each of the 10 detectors, Tables 8–17 list the top-5 strongest negative and positive sensitivities, ranked by $|\Delta\mathrm{AUC}| = |\mathrm{AUC}_{\mathrm{yes}} - \mathrm{AUC}_{\mathrm{no}}|$. $\mathrm{AUC}_{\mathrm{no}}$ is the matched attribute-absent AUC (the covariate-controlled baseline for that specific attribute); $\mathrm{AUC}_{\mathrm{yes}}$ is the AUC when the attribute is present; $\Delta\mathrm{AUC} = \mathrm{AUC}_{\mathrm{yes}} - \mathrm{AUC}_{\mathrm{no}}$ is the sensitivity. The detector's global AUC is reported in each caption for reference.

Interpretation. CDFA exhibits large failures on face masks (-0.21) and cartoon-style images (-0.14), consistent with incomplete coverage of strong occlusion or stylized domains in its training augmentation. Its strongest amplifying sensitivities are visually salient synthetic-texture cues such as digital face graphics, heavy braces, waxy skin, and colored lighting.

Table 8: CDFA (base AUC = 0.726). Top-5 negative and positive sensitivities, ranked by $|\Delta\mathrm{AUC}|$.

Attribute	$\mathbf{AUC_{no}}$	$\mathbf{AUC_{yes}}$	**ΔAUC**
Negative sensitivities			
Is the person wearing a face mask?	0.768	0.556	−0.212
Are there bright halos around facial outlines?	0.744	0.602	−0.142
Is this image cartoon-style?	0.756	0.616	−0.140
Is the facial structure completely deformed?	0.794	0.711	−0.083
Is this a purely monochromatic photo?	0.722	0.663	−0.059
Amplifying sensitivities			
Are there digital graphics on the face?	0.643	0.730	+0.087
Are the teeth completely covered by heavy braces?	0.748	0.794	+0.046
Does the facial skin look extremely waxy?	0.629	0.669	+0.040
Is the face illuminated by colored lights?	0.738	0.775	+0.037
Is the face heavily covered in shadows?	0.725	0.754	+0.030

Interpretation. ConvNeXt-S has the highest baseline among all detectors and a comparatively flat degradation sensitivity profile. Its primary vulnerability is luminance halos (-0.10), a blending-boundary artifact. Its amplifying sensitivities are high-contrast synthetic cues and controlled backgrounds (graphics, braces, waxy skin, solid black background), which may amplify high-contrast local cues.

Table 9: ConvNeXt-S (base AUC = 0.848). Top-5 negative and positive sensitivities, ranked by $|\Delta\mathrm{AUC}|$.

Attribute	$\mathbf{AUC_{no}}$	$\mathbf{AUC_{yes}}$	**ΔAUC**
Negative sensitivities			
Are there bright halos around facial outlines?	0.862	0.766	−0.096
Does the image have an extreme color tint?	0.869	0.831	−0.038
Is this a formal passport-style photo?	0.804	0.774	−0.030
Is the background completely solid white?	0.866	0.837	−0.029
Are the eyes looking in completely different directions?	0.896	0.871	−0.025
Amplifying sensitivities			
Are there digital graphics on the face?	0.834	0.890	+0.055
Are the teeth completely covered by heavy braces?	0.867	0.913	+0.046
Does the face have many different color patches?	0.831	0.874	+0.044
Does the facial skin look extremely waxy?	0.792	0.830	+0.038
Is the background completely solid black?	0.845	0.874	+0.029

Interpretation. D^3 exhibits the flattest degradation sensitivity profile across all 60 attributes (worst $\Delta\mathrm{AUC} = -0.034$), consistent with its discrepancy-based design for suppressing generator-specific shortcuts. Its failures concentrate on compositional and framing cues (logos, bald heads, rotation borders) rather than classic forgery artifacts. Its strongest amplifying sensitivity in the entire set remains digital face graphics ($+0.124$), consistent with discrepancy signals being more informative when unusual visual elements are present.

Interpretation. DDA is most sensitive to uniform backgrounds and framing/occlusion cues (solid white backgrounds, ear crops, rotation borders, bald heads), rather than to the texture degradations emphasized by some other detectors. Counterintuitively, DDA strongly improves on cartoon-style images ($+0.064$), face masks ($+0.061$), and strong bokeh ($+0.060$), making it a complementary detector exactly where several others fail.

Table 10: D^3 (base AUC = 0.794). Top-5 negative and positive sensitivities, ranked by $|\Delta\mathrm{AUC}|$.

Attribute	$\mathbf{AUC_{no}}$	$\mathbf{AUC_{yes}}$	$\mathbf{\Delta AUC}$
Negative sensitivities			
Is the person completely bald?	0.799	0.765	−0.034
Is the background completely solid white?	0.801	0.771	−0.029
Is the person wearing a face mask?	0.887	0.859	−0.028
Is there a prominent logo in the image?	0.784	0.758	−0.026
Does the image have black rotation borders?	0.804	0.780	−0.024
Amplifying sensitivities			
Are there digital graphics on the face?	0.775	0.899	+0.124
Does the facial skin look extremely waxy?	0.833	0.908	+0.075
Does the face have many different color patches?	0.868	0.923	+0.055
Does the background have a strong bokeh effect?	0.817	0.871	+0.054
Are there multiple faces clearly visible in the image?	0.833	0.884	+0.051

Table 11: DDA (base AUC = 0.795). Top-5 negative and positive sensitivities, ranked by $|\Delta\mathrm{AUC}|$.

Attribute	$\mathbf{AUC_{no}}$	$\mathbf{AUC_{yes}}$	$\mathbf{\Delta AUC}$
Negative sensitivities			
Is the background completely solid white?	0.820	0.721	−0.100
Are there bright halos around facial outlines?	0.835	0.753	−0.082
Are the ears cut off by the image border?	0.820	0.776	−0.044
Does the image have black rotation borders?	0.812	0.773	−0.040
Is the person completely bald?	0.829	0.801	−0.028
Amplifying sensitivities			
Are there digital graphics on the face?	0.755	0.855	+0.100
Does the face have many different color patches?	0.774	0.840	+0.066
Is this image cartoon-style?	0.884	0.948	+0.064
Is the person wearing a face mask?	0.869	0.930	+0.061
Does the background have a strong bokeh effect?	0.819	0.879	+0.060

Interpretation. Effort-FF exhibits the single largest failure in our dataset: cartoon-style images cause a -0.225 drop, consistent with a strong dependence on photorealistic facial priors. Monochrome and side-profile conditions further reveal a dependency on full-color, near-frontal face structure.

Interpretation. ForensicsAdapter shares failure modes with Effort-FF (cartoon, monochrome, side profile), consistent with a shared reliance on photorealistic visual representations. Its additional sensitivity to luminance halos may reflect sensitivity to blending-boundary signals that are disrupted by halo artifacts.

Interpretation. SegFormer-B3 is strongly degraded by luminance halos and desaturation conditions, consistent with sensitivity to boundary and color signals. Its positive response to face masks, braces, graphics, and bright lipstick suggests that it benefits from sharp local structure even when part of the face is occluded.

Interpretation. SegFormer-B3-Avg closely mirrors SegFormer-B3's profile, as expected given that it differs only in the classification head (average-pooling vs. standard). The high similarity between the two profiles is consistent with backbone-driven behavior dominating head design differences.

Interpretation. Swin-S shares the halo and monochrome failures common across transformer-based detectors. Its additional sensitivity to divergent eye gaze (-0.078) may reflect sensitivity to symmetric bilateral face structure, while its gains on braces, digital graphics, solid-black backgrounds, and colored lighting indicate that it still benefits from strong high-contrast local cues.

Interpretation. Xception has the lowest baseline AUC and is sensitive to halos, divergent gaze, and freckles-the last suggesting that natural high-frequency skin texture is misread as a GAN artifact. Its

Table 12: Effort-FF (base AUC = 0.821). Top-5 negative and positive sensitivities, ranked by $|\Delta\text{AUC}|$.

Attribute	AUC_{no}	AUC_{yes}	ΔAUC
Negative sensitivities			
Is this image cartoon-style?	0.789	0.564	−0.225
Is this a purely monochromatic photo?	0.813	0.720	−0.093
Is the face seen from an extreme side profile?	0.767	0.678	−0.089
Is this a black-and-white photo?	0.819	0.744	−0.074
Is the facial structure completely deformed?	0.864	0.798	−0.066
Amplifying sensitivities			
Does the facial skin look extremely waxy?	0.768	0.832	+0.064
Does the face have many different color patches?	0.759	0.809	+0.050
Does the image have black rotation borders?	0.805	0.838	+0.033
Is the background completely solid black?	0.830	0.861	+0.032
Are the teeth completely covered by heavy braces?	0.833	0.862	+0.030

Table 13: ForensicsAdapter (base AUC = 0.784). Top-5 negative and positive sensitivities, ranked by $|\Delta\text{AUC}|$.

Attribute	AUC_{no}	AUC_{yes}	ΔAUC
Negative sensitivities			
Is this image cartoon-style?	0.789	0.621	−0.168
Are there bright halos around facial outlines?	0.812	0.684	−0.128
Is this a purely monochromatic photo?	0.773	0.663	−0.110
Is this a black-and-white photo?	0.779	0.686	−0.093
Is the face seen from an extreme side profile?	0.761	0.683	−0.078
Amplifying sensitivities			
Are there digital graphics on the face?	0.723	0.778	+0.055
Does the facial skin look extremely waxy?	0.695	0.748	+0.053
Does the face have many different color patches?	0.716	0.752	+0.035
Are the lips tightly puckered?	0.852	0.886	+0.034
Is the face illuminated by colored lights?	0.784	0.816	+0.032

positive response to smooth skin and waxy skin is consistent with smooth texture being treated as a forgery indicator, compatible with early GAN-based deepfake characteristics.

I Annotation Consistency Analysis

We conduct an inter-model consistency audit using **Gemma-4-26B-A4B-IT** (Google DeepMind, 2026) as a second independently trained annotator. We sample 2,000 images from ScaleDF train in four balanced groups of 500-Random-Real, Random-Fake (natural distribution) and Stratified-Real, Stratified-Fake (oversampled from Qwen-positive images to ensure $\geq$8 positives per attribute)—and compare each image's 60-dimensional binary attribute vector from Qwen against that from Gemma.

Per-Image Hamming Distance. Table 18 reports the per-image disagreement. On average, the two models disagree on **3.48 out of 60 attributes** (5.8%) per image. Under the natural distribution (random groups), the median Hamming distance is 2; the stratified groups show higher disagreement (median 3–4) by design, since they oversample rare, ambiguous positive attributes.

Per-attribute agreement. Of the 60 attributes: **39** (65%) achieve agreement $\geq$0.95, **49** (82%) achieve $\geq$0.90, and **11** (18%) fall below 0.90. Table 19 lists the five highest- and lowest-agreement attributes. High-agreement attributes are semantically unambiguous binary concepts; low-agreement attributes concentrate on framing and threshold-sensitive visual judgments (border crops, bokeh, rotation borders, shadow severity) where each model applies its own implicit calibration threshold.

Table 14: SegFormer-B3 (base AUC = 0.782). Top-5 negative and positive sensitivities, ranked by $|\Delta\mathrm{AUC}|$.

Attribute	$\mathbf{AUC_{no}}$	$\mathbf{AUC_{yes}}$	$\mathbf{\Delta AUC}$
Negative sensitivities			
Are there bright halos around facial outlines?	0.778	0.624	−0.153
Is this a purely monochromatic photo?	0.758	0.663	−0.095
Is this a black-and-white photo?	0.769	0.680	−0.089
Are the background colors extremely vivid?	0.809	0.773	−0.036
Is the person an extremely young baby?	0.778	0.744	−0.035
Amplifying sensitivities			
Is the person wearing a face mask?	0.791	0.835	+0.044
Are there digital graphics on the face?	0.760	0.802	+0.042
Are the teeth completely covered by heavy braces?	0.728	0.765	+0.038
Does the face have many different color patches?	0.723	0.755	+0.032
Are the lips painted with bright lipstick?	0.809	0.833	+0.024

Table 15: SegFormer-B3-Avg (base AUC = 0.794). Top-5 negative and positive sensitivities, ranked by $|\Delta\mathrm{AUC}|$.

Attribute	$\mathbf{AUC_{no}}$	$\mathbf{AUC_{yes}}$	$\mathbf{\Delta AUC}$
Negative sensitivities			
Are there bright halos around facial outlines?	0.813	0.668	−0.145
Is this a purely monochromatic photo?	0.782	0.697	−0.085
Is this a black-and-white photo?	0.792	0.711	−0.080
Are the background colors extremely vivid?	0.823	0.776	−0.047
Is this a formal passport-style photo?	0.746	0.700	−0.046
Amplifying sensitivities			
Are the teeth completely covered by heavy braces?	0.791	0.849	+0.058
Are there digital graphics on the face?	0.782	0.836	+0.054
Does the face have many different color patches?	0.757	0.802	+0.045
Is the face illuminated by colored lights?	0.805	0.845	+0.040
Does the facial skin look extremely waxy?	0.707	0.747	+0.040

Table 16: Swin-S (base AUC = 0.808). Top-5 negative and positive sensitivities, ranked by $|\Delta\mathrm{AUC}|$.

Attribute	$\mathbf{AUC_{no}}$	$\mathbf{AUC_{yes}}$	$\mathbf{\Delta AUC}$
Negative sensitivities			
Are there bright halos around facial outlines?	0.832	0.690	−0.142
Is this a black-and-white photo?	0.817	0.724	−0.093
Is this a purely monochromatic photo?	0.803	0.710	−0.092
Are the eyes looking in completely different directions?	0.859	0.780	−0.078
Is this a formal passport-style photo?	0.777	0.732	−0.045
Amplifying sensitivities			
Are the teeth completely covered by heavy braces?	0.767	0.822	+0.056
Are there digital graphics on the face?	0.808	0.859	+0.050
Is the background completely solid black?	0.822	0.860	+0.038
Is the face illuminated by colored lights?	0.820	0.848	+0.028
Does the facial skin look extremely waxy?	0.735	0.762	+0.027

Table 17: Xception (base AUC = 0.755). Top-5 negative and positive sensitivities, ranked by $|\Delta\mathrm{AUC}|$.

Attribute	**$\mathrm{AUC}_{\mathrm{no}}$**	**$\mathrm{AUC}_{\mathrm{yes}}$**	**ΔAUC**
Negative sensitivities			
Are there bright halos around facial outlines?	0.771	0.651	−0.120
Are the eyes looking in completely different directions?	0.779	0.685	−0.094
Is the background completely solid white?	0.765	0.703	−0.062
Does the face have distinct freckles?	0.750	0.692	−0.058
Is this a formal passport-style photo?	0.758	0.728	−0.030
Amplifying sensitivities			
Are there digital graphics on the face?	0.735	0.773	+0.038
Does the facial skin look extremely waxy?	0.684	0.720	+0.035
Is the face illuminated by colored lights?	0.756	0.791	+0.035
Is the facial skin extremely smooth?	0.754	0.789	+0.034
Are the lips painted with bright lipstick?	0.780	0.814	+0.034

Table 18: Per-image Hamming distance between Qwen3.5 and Gemma-4-26B attribute vectors (out of 60). Lower is better.

Group	**N**	**Mean**	**Median**
Random-Real	500	2.17	2
Random-Fake	500	2.86	2
Stratified-Real	500	4.03	3
Stratified-Fake	500	4.84	4
Overall	2,000	3.48	3

Table 19: Per-attribute Qwen–Gemma agreement rate: top 5 and bottom 5.

Attribute	**Agreement**
Highest agreement	
Is the person an extremely young baby?	0.996
Is this a black-and-white photo?	0.993
Are large headphones completely covering the ears?	0.993
Is this a purely monochromatic photo?	0.993
Is the background completely solid white?	0.993
Lowest agreement	
Are the ears cut off by the image border?	0.701
Does the background have a strong bokeh effect?	0.751
Does the image have black rotation borders?	0.752
Is the chin cut off by the image border?	0.825
Are there heavy shadows on the face?	0.833

Human annotation procedure. For a separate audit, two annotators independently reviewed the same 200 class-balanced ScaleDF images across all 60 attributes. Each batch displayed one attribute question and 200 images with opaque filenames; annotators could not see Qwen labels or authenticity labels. They checked images judged positive for the displayed attribute, and submitting the batch recorded unchecked images as negative. Figure 6 shows the selection interface.

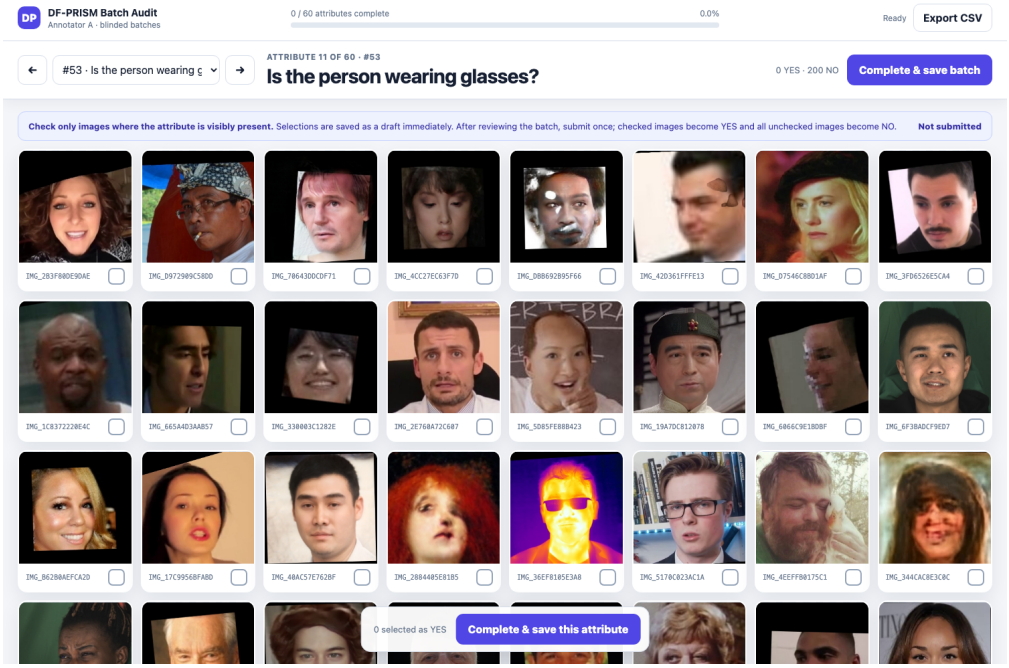

Figure 6: Blinded human attribute annotation interface. The initial, unsubmitted state is reconstructed using the original interface and blinded images; empty checkboxes are not completed negative judgments.

J Compute Resources

All experiments were conducted on a cluster of 16 NVIDIA H100 80 GB GPUs.

Attribute Annotation. Attribute annotation was performed using Qwen3.5 9B served via vLLM (Kwon et al., 2023). Each of the 60 profiling attributes was annotated in a separate inference pass over the full ScaleDF training split (≈5.85M images), achieving a sustained throughput of approximately 70 images/second per attribute. This yields a per-attribute annotation time of $5{,}850{,}000 \div 70 \approx 23$ hours, and a total of $60 \times 23 \approx 1{,}390$ GPU-hours for the complete attribute matrix. With 16 H100s running attribute jobs in parallel (up to 16 concurrent annotation jobs), the end-to-end wall-clock time for the annotation stage was approximately 90 hours.

Detector Inference. Profile construction requires running each of the 10 detector checkpoints over the annotated image set to obtain per-image prediction scores. The combined inference cost across all 10 detectors was approximately 130 GPU-hours.

Benchmark Evaluation. The 13 evaluation settings comprise approximately 1.02M images in total. Attribute annotation and detector inference were repeated for this evaluation set using the same pipeline, adding approximately $60 \times (1{,}020{,}000 \div 70)/3600 \approx 243$ GPU-hours for annotation and ≈23 GPU-hours for detector inference (scaled proportionally from the training-set cost).

Summary. The end-to-end DF-PRISM pipeline (profiling and evaluation) required approximately **5 days** wall-clock time on 16 H100 80 GB GPUs, totalling approximately **1,800 GPU-hours**. Table 20 provides a per-stage breakdown.

Table 20: Compute resource summary for the DF-PRISM profiling pipeline.

Stage	Hardware	Time
Attribute annotation (60 attrs × 5.85M images, profiling)	16× H100 80 GB	≈1,390 GPU-h
Detector inference (10 detectors × 5.85M images, profiling)	16× H100 80 GB	≈130 GPU-h
Attribute annotation (60 attrs × 1.02M images, evaluation)	16× H100 80 GB	≈243 GPU-h
Detector inference (10 detectors × 1.02M images, evaluation)	16× H100 80 GB	≈23 GPU-h
Total		**≈1,800 GPU-h**

K Visual Examples of Profiling Attributes

Figures 7–11 provide representative visual examples of the 60 profiling attributes organized by semantic category. Each example shows a single image in which the attribute is active. These figures serve as a qualitative visual glossary of the attribute vocabulary; they are not intended as quantitative evidence or detector failure-case analysis.

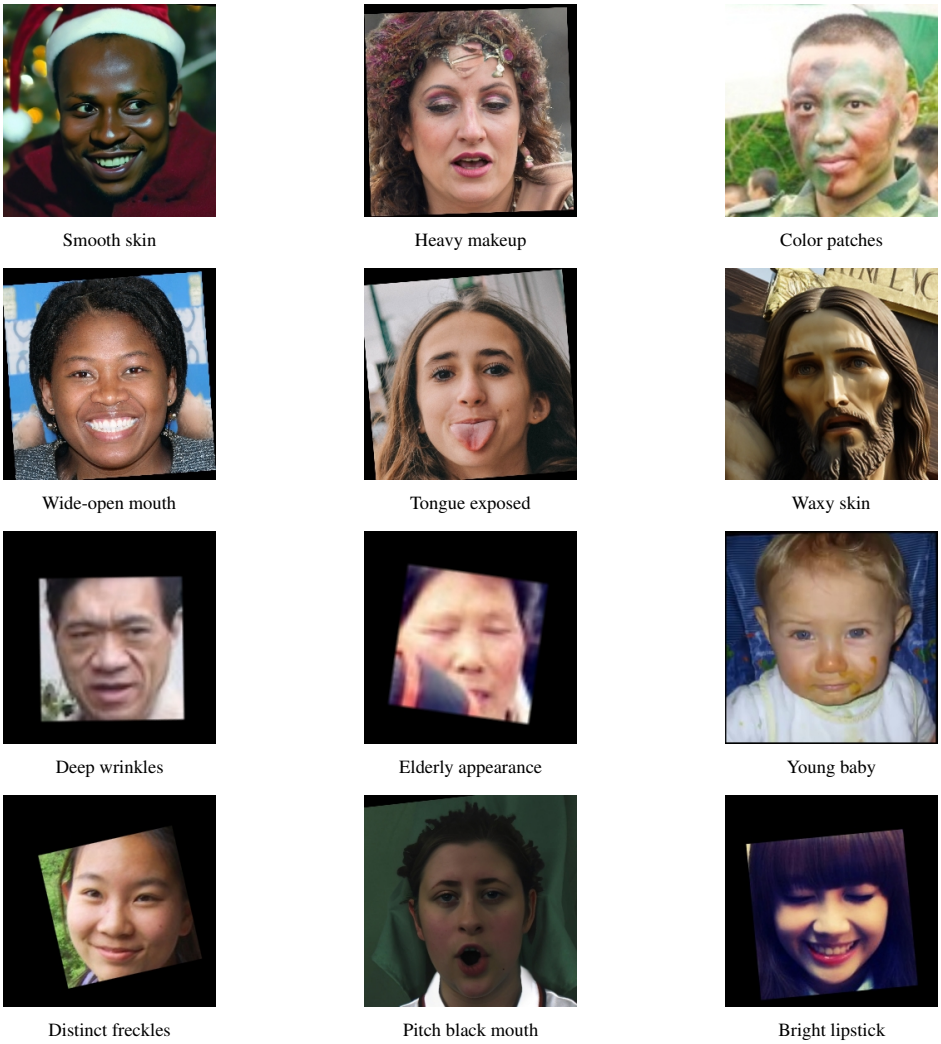

Figure 7: **Texture and Appearance.** These examples illustrate how the operational attribute vocabulary is instantiated in practice. They are intended as qualitative references for the annotation space, not as quantitative evidence.



Figure 8: **Background, Lighting, and Post-Processing.** Additional visual examples of profiling attributes (background, lighting, and post-processing).

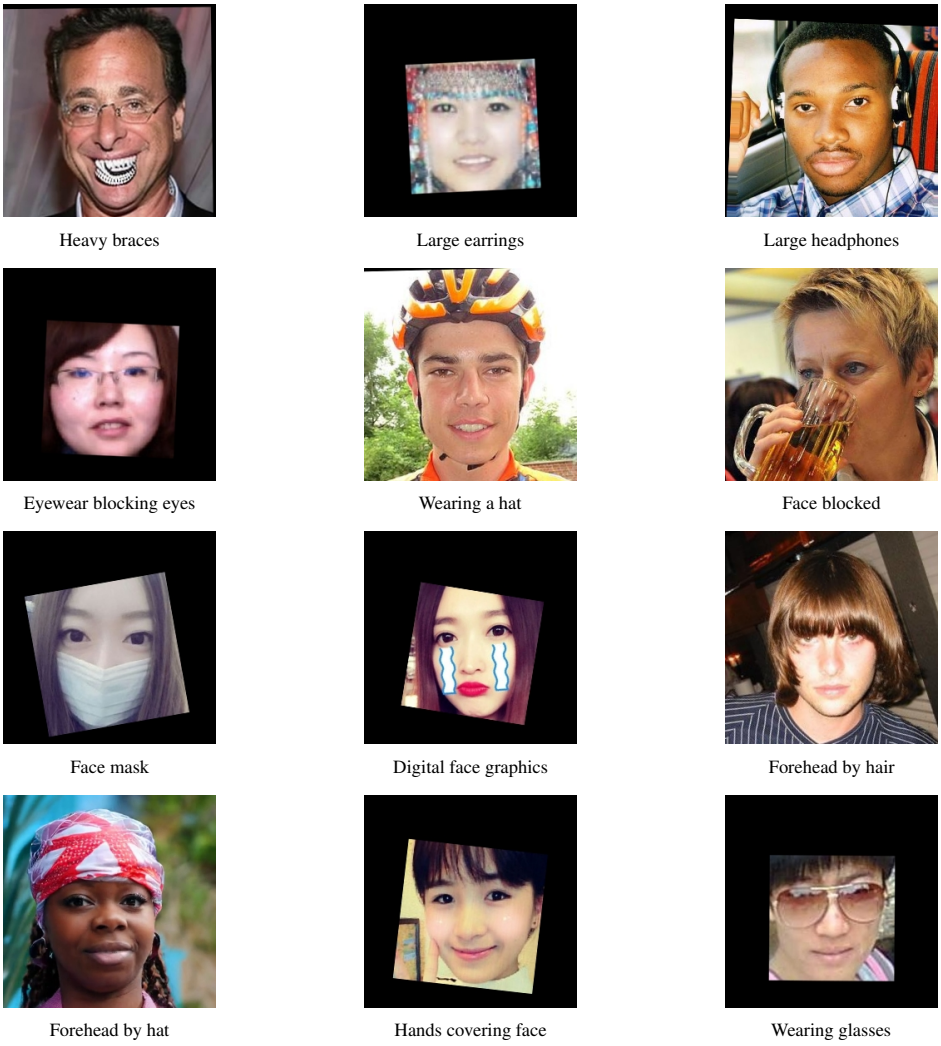

Figure 9: **Occlusion and Accessories.** Additional visual examples of profiling attributes (occlusion and accessories).

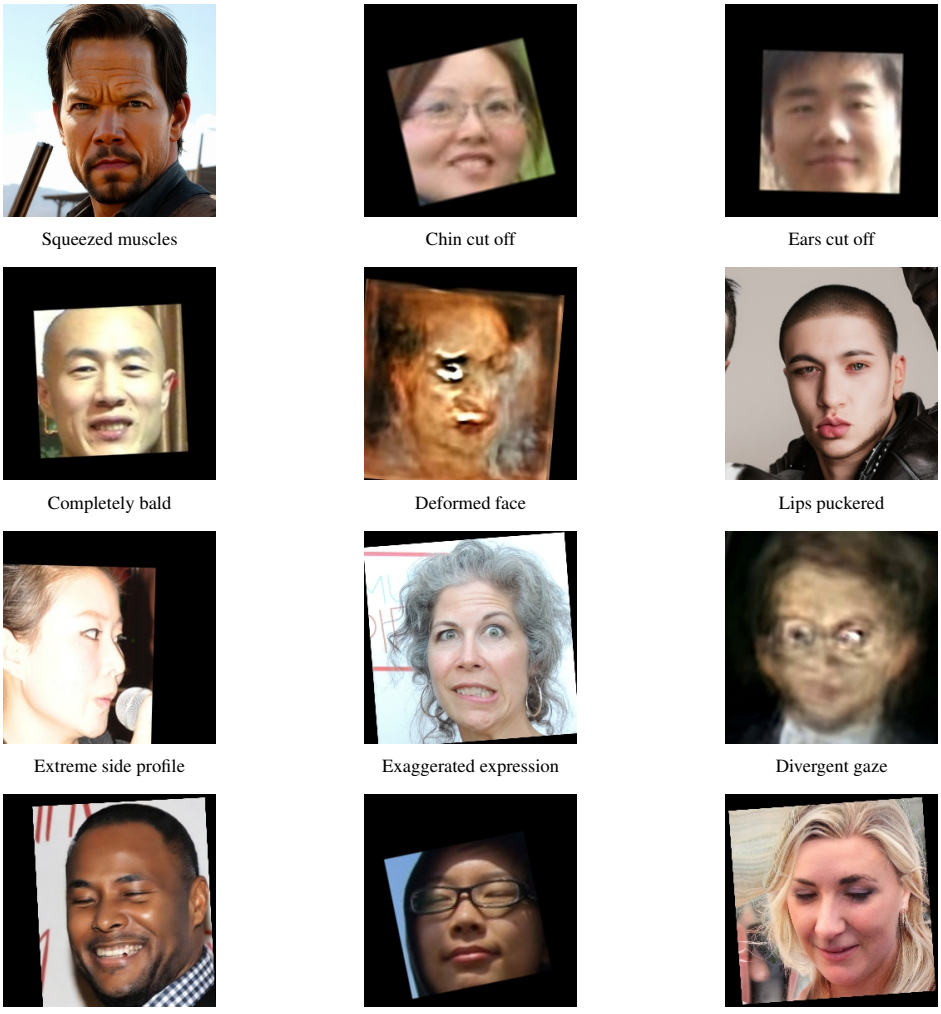

Figure 10: **Pose, Expression, and Geometry.** Additional visual examples of profiling attributes (pose, expression, and geometry).

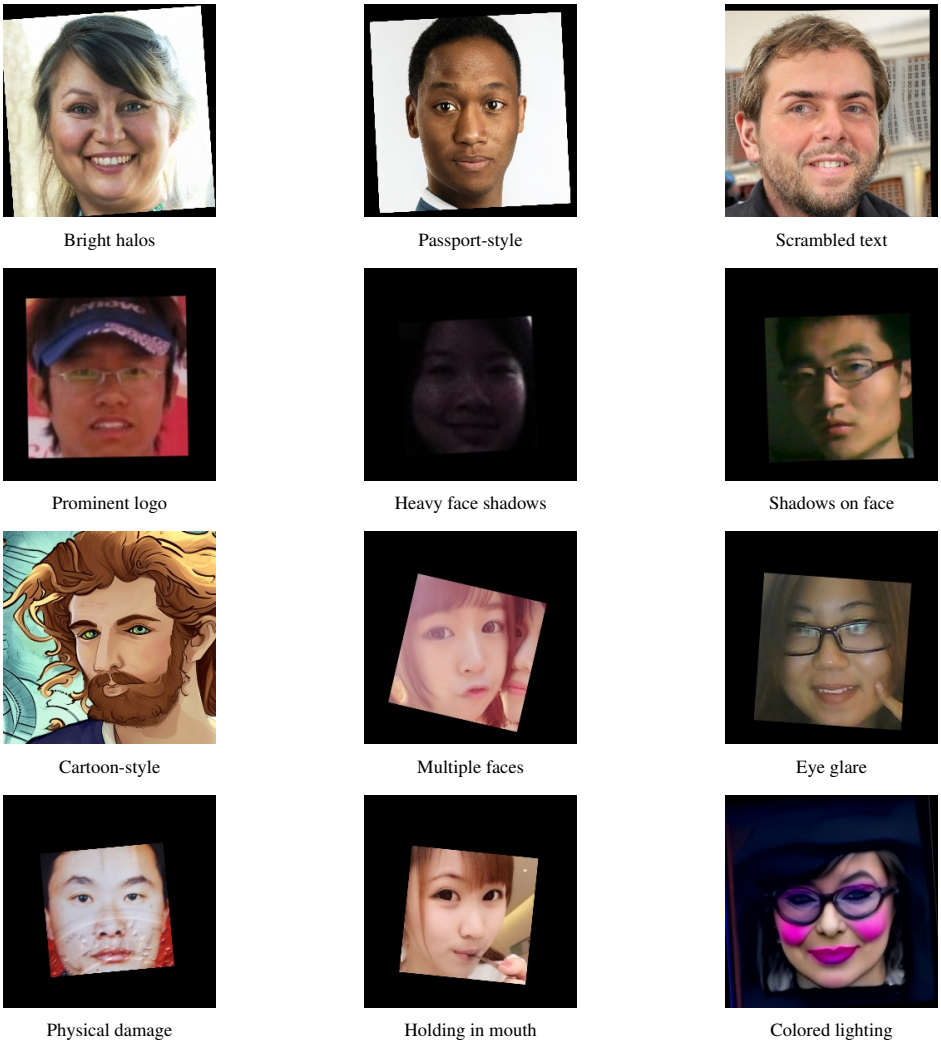

Figure 11: **Contextual.** Additional visual examples of profiling attributes (contextual).