Accepted to IEEE S&P 2027 (48th IEEE Symposium on Security and Privacy) in September 2026, in press, preprint available at https://doi.org/10.48550/arXiv.2610.23144



From GAN-Based Models to GPT-5: A Taxonomy-Guided Study of Human Performance in Detecting Deepfake Face Images

Zhiqing Guo*‖, Lixin Jia*‖, Luning Sun†‖, Xiaowei Chen‡, Guofeng Yu*
Liejun Wang***, Gaobo Yang§, David Stillwell†, Thorsten Holz‡ and Shujun Li¶**
*Xinjiang University, China †University of Cambridge, UK
‡Max Planck Institute for Security and Privacy (MPI-SP), Germany §Hunan University, China
¶Institute of Cyber Security for Society (iCSS), University of Kent, London and South East University Group, UK
‖Equivalent contribution **Corresponding co-authors: wljxju@xju.edu.cn (Liejun Wang), S.J.Li@kent.ac.uk (Shujun Li)

***Abstract*—Generative models for deepfakes have rapidly evolved from GAN-based architectures to diffusion models and multimodal large language models (MLLMs). However, empirical research on human performance in detecting deepfake face images has not kept pace. It remains unclear if (and how) detection performance varies across generative models and whether model-specific training improves detection more effectively than general-purpose training. In this paper, we propose the *Deepfake Face Artifact Taxonomy* (DFAT), a cross-model taxonomy of human-recognizable visual artifacts in deepfake face images, based on which we build both general-purpose and model-specific training materials. Through an online experiment ($n = 524$), we examined human performance in detecting deepfake face images and evaluated training efficacy across five representative generative models: three GAN-based, one diffusion-based, and one MLLM (GPT-5). Our results show that participants' detection accuracy was strongly model-dependent and that images generated by more recent and advanced models were not necessarily harder to detect. After training, participants achieved significant performance gains across most conditions, and their improvement varied by model. Notably, model-specific training generally yielded greater numerical gains than general-purpose training. In a five-month follow-up test ($n = 342$), we observed no significant decline in detection accuracy and no significant increase in perceived stress. Overall, our findings provide actionable insights for designing effective, sustainable, and targeted training programs to counter the growing threat posed by deepfakes.**

1. Introduction

Deepfake face images are frequently exploited for malicious purposes, including disinformation campaigns [1], non-consensual pornography [2], [3], online fraud [4], and cyberbullying [5]. The recent development of generative AI techniques, particularly multimodal large language models (MLLMs) [6], [7] and large vision models (LVMs) [8], has significantly escalated such threats. These modern deepfake generation models have lowered the technical barriers substantially, allowing adversaries to generate highly realistic deepfakes from textual and multimodal prompts, without the need for extensive expertise or computational resources.

Unfortunately, existing human-focused studies have not kept pace with the rapid development of modern deepfake generators. Most prior work on deepfake face images relies on relatively old GAN-based architectures, particularly StyleGAN2 [9]. Only one recent study [10] examined diffusion-based image generators (e.g., Midjourney V5 [11]) without considering any MLLMs. Cooke et al. [12] included both GAN-based and diffusion models, but failed to report model-specific detection performance. Without empirical evidence, it remains unclear whether and how human performance in detecting deepfake face images varies across different generative models.

Current approaches to mitigating deepfakes focus predominantly on automated detection [13], [14], [15], [16]. However, detector performance often degrades under distribution shifts to unseen generators and adversarial manipulation [17], [18]. These tools may also be unavailable in real-time or end-to-end encrypted settings. Even when they are available, users still need to determine whether they should trust these outputs and how to act on them. As a result, users are left to assess image authenticity on their own, making user-facing support for deepfake discernment an emerging topic in human-centered security [19].

Despite this need, human performance in detecting deepfakes remains comparatively understudied. A 2024 meta-analysis [20] identified 56 studies spanning multiple modalities (audio, image, text, and video). Constraining the scope to human-subject studies of deepfake face images, we identified only nine studies that reported quantitative performance for photorealistic facial deepfakes. Overall, these studies suggest that untrained users are poorly equipped to identify high-quality deepfakes, often performing at a near chance level [21], [22]. Although some studies reveal that feedback-based training and tutorials can improve human performance [23], [24], [25], [26], their effectiveness varies considerably across study designs, with several studies reporting largely negative results [27].

To address these gaps, we evaluate human performance in detecting deepfake face images generated by five representative models that span the current face-synthesis tra-

jectory: StyleGAN2, StyleGAN3 and GRAM-HD as GAN-based face generators, SD3.5 as an open-weight diffusion model, and GPT-5 as a widely accessible MLLM-based generator. To improve training effectiveness, we construct programs that are grounded in a systematic synthesis of human-recognizable visual artifacts across these generative models. In this study, we aim to address the following research questions (RQs):

RQ1: What human-recognizable visual artifacts characterize deepfake face images generated by different generative models?

RQ2: Does human performance in detecting deepfake face images vary across different generative models, and if so, how?

RQ3: How effective are taxonomy-guided training interventions in improving human performance in detecting deepfake face images, and is model-specific training more effective than general-purpose training?

RQ4: Is the performance improvement derived from the taxonomy-guided training interventions sustainable over a relatively long period of time?

This study focuses on photorealistic human face images, one of the most prevalent and socially consequential forms of deepfakes. We begin by examining real and deepfake face images generated by different models. Building on this systematic analysis, we propose the *Deepfake Face Artifact Taxonomy* (DFAT), which offers a structured and extensible framework for characterizing model-specific visual artifacts that can aid in deepfake detection. Guided by this taxonomy, we design two types of deepfake detection training interventions: general-purpose and model-specific. We then evaluate and compare their effectiveness through an online experiment comprising a main study ($n = 524$) and a follow-up study five months later ($n = 342$) assessing retention effects. This study makes the following key *contributions*:

- We propose DFAT, an extensible taxonomy of human-recognizable artifacts in deepfake face images generated by diverse models. DFAT supports systematic comparison of artifact profiles across deepfake generation models and provides a structured basis for developing training interventions.
- We present a cross-model baseline for human performance in deepfake detection and show that deepfakes generated by more recent and advanced models are not necessarily harder for humans to detect. This finding calls for a shift of future studies from focusing on more recent and advanced models to model-specific human-recognizable artifacts.
- We develop DFAT-guided training interventions and demonstrate their effectiveness in improving deepfake detection accuracy, with model-specific training outperforming general-purpose training. These results offer practical guidance for designing future educational interventions for deepfake detection.
- We provide longitudinal evidence on the retention of humans' deepfake detection performance through a follow-up test five months after the main experiment. The results reveal persistent improvement in detection accuracy and user confidence without increasing perceived stress, highlighting the sustainability of DFAT-guided training interventions.

2. Related Work

2.1. Deepfake Generation Models

Deepfake image generation now covers faces, objects, and scenes across several model families. GANs established high-fidelity image synthesis, with fully generative face synthesis becoming one of their most mature applications [33], [34]. Diffusion models and latent diffusion models made text-guided image generation easier to use widely [35], [36]. Recent systems, including MLLM-based interfaces and commercial image generators [11], [37], further allow users to generate realistic images using natural language inputs.

Human faces are a demanding target for image generation because they contain complex geometry, fine-grained textures, and rich expression variations. Face images also play a central role in social media, identity presentation, and online security, making realistic face synthesis one of the most visible forms of deepfake images. Deepfake face generation includes face manipulation, such as face swapping and editing [38], [39], as well as fully generative face synthesis. For fully generative face synthesis, the StyleGAN series remains a standard benchmark [9], [28], [40], and recent GAN-based approaches further improve visual quality and realism through new architectures and training strategies [29], [41], [42].

To operationalize this landscape for our empirical analysis, we select five representative deepfake models that cover the face-synthesis design space. We include StyleGAN2 [9] and StyleGAN3 [28] as widely used GAN-based face synthesis baselines, and GRAM-HD [29] as a recent GAN-based model for high-resolution face generation. Beyond GANs, we include SD3.5, an open-weight text-to-image diffusion model capable of synthesizing high-resolution face images at scale [30], [31]. We additionally include GPT-5 [32], OpenAI's image generation model accessed via the GPT-5 interface after its August 2025 release, to represent a widely deployed closed-weight model baseline. Table 1 summarizes the generators used for our work.

2.2. Automated Deepfake Detection

Automated detection has been extensively studied as a technical response to AI-generated images and broader synthetic media [18], [43]. Most approaches train deep models to distinguish real images from generated ones by learning statistical or visual discrepancies from labeled datasets. Within this broader area, deepfake face detection remains one of the most mature and widely studied branches [44]. Existing methods have examined visible synthesis and blending artifacts [45], [46], frequency-domain inconsistencies [47], and subtle trace amplification [48].

TABLE 1: Description of the five generative models used in our study.

Model	Venue/Vendor (Year)	Output Size(s)	Reason(s) of Selection
StyleGAN2 [9]	CVPR (2020)	1024×1024	The most widely studied high-fidelity GAN-based model
StyleGAN3 [28]	NeurIPS (2021)	1024×1024	The most recent variant of the StyleGAN family with improved alias-free generation
GRAM-HD [29]	ICCV (2023)	1024×1024	A GAN-based model more recent than StyleGAN3
Stable Diffusion 3.5 Large (SD3.5 for short) [30], [31]	Stability AI (2024)	Multiple sizes	The latest version of the Stable Diffusion open-weight text-to-image models supporting prompt-based generation
GPT image generation (GPT-5 interface) [32]	OpenAI (2025)	Multiple sizes	One of the best-performing and widely used MLLMs when we started our work in mid 2025

Under constrained experimental settings, some detection methods achieved very high accuracy [49].

Nevertheless, detection performance often degrades under realistic conditions due to distribution shifts [50], such as unseen deepfake generation models, model updates, image compression, or post-processing. In addition, automated detectors are not always available or accessible for everyday use. Even when such tools are available, users still need to interpret detector outputs and decide how to act. These limitations motivate complementary human-centered support for deepfake detection. Accordingly, our work focuses on identifying and organizing human-recognizable visual artifacts produced by deepfake generators, and on developing a taxonomy that supports humans in detecting deepfake faces.

2.3. Human Detection of Deepfakes

Empirical studies on human detection of deepfake content remain limited relative to automated detection. Existing work has examined synthetic media across modalities, including image, audio, text, and video [15], [20], [21]. In the image domain, prior studies have considered both face images and broader image categories such as landscapes, animals, food, and objects [10], [12]. Focusing on deepfake face images, we identified nine studies that reported quantitative human accuracy for photorealistic facial deepfakes [10], [12], [21], [22], [23], [24], [25], [26], [27] (see Table 4 in Appendix A). Despite differences in experimental design, these studies converge on the following two conclusions.

Unaided human performance is poor. Many past studies revealed that untrained participants tended to operate at a near-chance level on detecting high-quality deepfake faces [12], [21], [22], [25], [26], indicating limited spontaneous sensitivity to contemporary synthesis artifacts. For instance, Shen et al. [22] characterized baseline performance on StyleGAN2-generated faces as indistinguishable from random guessing. This perceptual limitation extends to large-scale settings involving image-face conditions, where Frank et al. [21] demonstrated that poor detection is a systemic failure rather than an artifact of specific participant pools or experimental designs. Such results validate the central premise that end users, despite being the primary targets of synthetic media, cannot reliably distinguish high-quality deepfakes without support.

Interventions may help improve human performance. It has been reported that interventions based on training or feedback can improve human performance [23], [24], [25], [26], but the improvements are modest, unstable across settings and studies, and strongly contingent on the specific training protocols and evaluation conditions employed. Feedback-based protocols serve as a primary mechanism for this improvement: Hulzebosch et al. [24] demonstrated that iterative feedback refines accuracy on GAN-generated faces, while Diel et al. [25] found these gains robust when feedback is paired with specific artifact attention. Nightingale et al. [26] further substantiated the utility of cue-based guidance, showing that artifact-focused tutorials on StyleGAN2 synthesis yield measurable discriminative improvements. Distinct from cognitive training, Boyd et al. [23] validated the integration of AI support systems, confirming that external algorithmic assistance provides a distinct pathway to augmented accuracy on deepfake image. Despite the positive evidence reported, the effectiveness of interventions remains inconsistent across studies. For instance, Bray et al. [27] observed negligible performance gains despite the provision of explicit knowledge regarding visible rendering artifacts, a result that decouples information availability from practical learnability. This finding implies that abstract awareness of artifacts does not automatically resolve into reliable perceptual heuristics. The literature thus characterizes training and assistance as promising yet volatile mechanisms, necessitating more systematic and actionable guidance to overcome the substantial variability observed across existing protocols.

Our analysis of the literature led to the discovery of some critical gaps. First, most studies focused on GAN-based models, especially StyleGAN2. Although some work studied multiple models including GAN-based models and diffusion models, model-specific effects have not been studied systematically. Therefore, if and how human performance and training effects vary across models remain unclear. Second, prior interventions often used artifacts as training cues, but these cues were selected without a systematic face-specific taxonomy. Consequently, the training and feedback interventions tested in prior work are less systematic.

Existing taxonomy-oriented work shows that organizing visual artifacts can support deepfake detection and related training materials. Kamali et al. [51] categorized observable cues in AI-generated photographs into anatomical, stylis-

tic, functional, physical, and sociocultural categories. Their follow-up work further characterized artifacts in diffusion-generated images through a large-scale human study [52]. These efforts show the value of structured artifact descriptions, but their scope mainly focuses on general AI-generated photographs and diffusion-model outputs. Their work does not cover model-specific and more fine-grained artifacts in deepfake face images.

Accordingly, we selected five representative generative models from three model families (GAN-based models, diffusion models, and MLLMs; see Table 1). We systematically characterize human-recognizable visual artifacts in fully generative face images and organize them into the Deepfake Face Artifact Taxonomy (DFAT), a training-oriented taxonomy for comparing artifacts across model families. DFAT provides a structured vocabulary of visual artifacts that can guide the construction of training materials and enable consistent analysis across models and participants.

3. Deepfake Face Artifact Taxonomy

This section describes our Deepfake Face Artifact Taxonomy (DFAT) with greater details, which addresses RQ1.

3.1. Taxonomy Construction Methodology

DFAT focuses on fully generative, photorealistic face images. We constructed DFAT by analyzing 1,000 1024×1024 images, including 500 real photographs and 500 deepfake images generated by the five models shown in Table 1. Each deepfake model contributed 100 images. Real images were sourced from the Flickr-Faces-HQ (FFHQ) dataset [40] to provide high-resolution reference images for comparison with synthetic outputs. To generate deepfake images, we used model-specific procedures:

- **GAN-based models (StyleGAN2, StyleGAN3, GRAM-HD)**: We used the pre-trained weights provided by the original authors. This setup reflects a common use case in which users directly deploy off-the-shelf generators to generate deepfake images.
- **One diffusion model (SD3.5)**: We used SD3.5 through a local deployment to generate photorealistic face images via text-to-image synthesis. The prompts described portraits and varied subject attributes and image details to improve realism and cover diverse facial appearances.
- **One MLLM (GPT-5)**: We used GPT-5 via the ChatGPT web interface to generate photorealistic face images. The prompts followed the same strategy, varying subject attributes and image details to improve realism and cover diverse facial appearances.

Model source links and all prompts are provided in the artifact (Appendix E). We used direct outputs from the selected models as the deepfake images for DFAT, without manual retouching, compositing, or region-level replacement. We then manually inspected the images to identify human-visible anomalies that could support visual detection. We first recorded recurring low-level observations, such as distorted eyes, blurred boundaries, or inconsistent accessories. We then grouped related observations into broader artifact categories and organized these categories into the three-level hierarchy shown in Figure 1. The resulting taxonomy captures four artifact domains, their artifact categories, and concrete visual manifestations.

3.2. Taxonomy Overview

DFAT organizes visual artifacts into four primary domains: **face-related**, **style-related**, **background-related**, and **accessory-related**. Each domain is divided into *artifact categories* (e.g., tooth rendering anomalies) and concrete *visual manifestations* (e.g., abnormal number or shape). Five annotators each labeled 100 images from one model, recording the artifact types present. One co-author reviewed all annotations and finalized any corrections through discussion with each annotator. Table 2 shows the prevalence levels of each artifact. These levels summarize the model-specific artifact profile observed in our sampled images and motivate the artifact cues used in the training materials.

3.2.1. Face-related Artifacts. Since the face is the primary focus of most models, these artifacts are often subtle and require careful observation across various facial features. We identified the following five categories of such artifacts.

Tooth rendering anomalies. While teeth may appear regular at first glance, models often struggle with the unique and complex shapes of individual teeth. This results in *abnormal number and shape*, such as three prominent front teeth or malformed dental arrangements, highlighting the difficulty in processing multiple similar anatomical structures.

Unnatural hair. Accurately rendering the complex structure and intricate illumination of thousands of independent hairs remains a challenge. This is manifested in *overly smooth hair textures* that lack depth and strand-level detail (e.g., common in SD3.5-generated images), and *hair fused with facial skin*, where strands unnaturally merge with the skin or exhibit anomalous clumping and orientations.

Unnatural facial skin. Synthesizing realistic skin with natural light response is a complex task. Artifacts include *overly smooth skin texture*, where the result lacks high-frequency details like pores (prevalent in SD3.5-generated images and occasionally observed in GPT-5-generated images), and *abnormal skin patterns*, where the texture resembles a painted surface or exhibits subtle architecture-related inconsistencies. The latter is surprisingly noticeable in StyleGAN3, where we identified a previously unreported artifact: the skin often contains strange oval/diamond-shaped local mini-textures. Such patterns are difficult to see at normal viewing size but become clearly recognizable after magnification (see Appendix D for examples).

Eye rendering anomalies. They primarily manifest as *unnatural pupil appearances*, where pupils may be irregular in shape, inconsistent in size between eyes, or exhibit ghosting.

Ear rendering anomalies. They typically involve the ear being *indistinguishable from hair*. For images with such

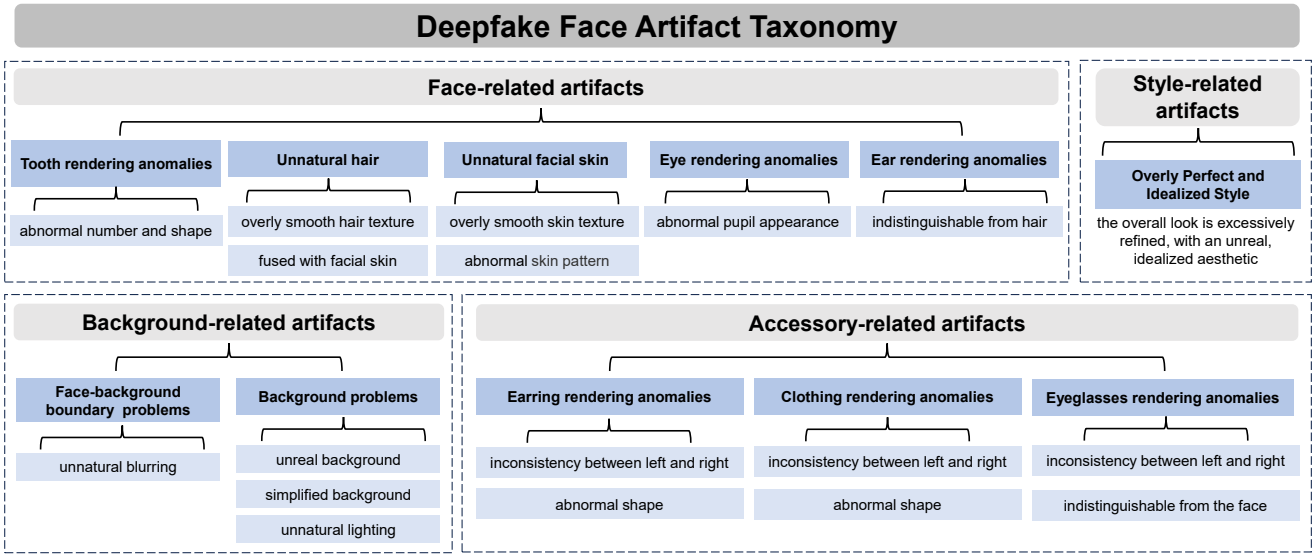


Figure 1: Illustration of the Deepfake Face Artifact Taxonomy (DFAT).

TABLE 2: **Prevalence of artifact types across deepfake models.** We categorize the occurrence frequency into four levels: **High** (>50%), **Medium** ($>$ 20% and $\leq$ 50%), **Low** (1–20%), and **–** (not observed, i.e., 0%).

Artifact Type	StyleGAN2	StyleGAN3	GRAM-HD	SD3.5	GPT-5
Tooth rendering anomalies (abnormal number/shape)	Low	Medium	Medium	–	–
Unnatural hair (overly smooth texture)	–	–	Low	Medium	Medium
Unnatural hair (fused with facial skin)	Medium	High	Medium	Medium	Low
Unnatural facial skin (overly smooth texture)	–	–	–	High	Low
Unnatural facial skin (abnormal skin pattern)	Low	High	High	High	–
Eye rendering anomalies (abnormal pupil appearance)	–	–	Low	Low	–
Ear rendering anomalies (indistinguishable from hair)	Low	Medium	Medium	Low	–
Overly perfect and idealized style	–	–	–	High	High
Face-background boundary (unnatural blurring)	Medium	Medium	High	High	–
Background problems (unreal background)	Medium	Medium	Medium	Low	–
Background problems (simplified background)	Medium	Medium	Low	High	Medium
Background problems (unnatural lighting)	–	–	Low	Low	–
Earring anomalies (left/right inconsistency)*	High	High	High	Medium	–
Earring anomalies (abnormal shape)*	Medium	Medium	High	Low	–
Clothing anomalies (left/right inconsistency)	Medium	Medium	Low	Low	–
Clothing anomalies (abnormal shape)	Medium	Medium	Medium	Low	–
Eyeglasses anomalies (left/right inconsistency)*	High	High	High	–	–
Eyeglasses anomalies (indistinguishable from face)*	Medium	Medium	Medium	–	–

*For accessory-related artifacts, the level indicates the proportion of images exhibiting at least one such artifact out of all images with one or more relevant accessories (not out of all images).

artifacts, the model failed to separate contours, causing the ear to appear blurred or fused with adjacent hair strands.

3.2.2. Style-related artifacts. Unlike localized glitches, style-related artifacts pertain to the overall aesthetic composition and stylistic signature. Diffusion models and MLLMs like SD3.5 and GPT-5 are often over-optimized so that the generated images have a look and feel of being **overly perfect and idealized**. This results in a distinct, often unrealistic visual signature characterized by hyper-realism and "idealized" outputs that deviate from the natural variations found in genuine photography.

3.2.3. Background-related Artifacts. Background anomalies typically arise from generative designs that prioritize facial regions over peripheral content. Furthermore, training datasets often contain diverse and unstructured background information, making it challenging for models to learn semantically consistent representations of the background. There are two categories.

Face-background boundary problems. These occur at the interface between the subject (face, hair or shoulders) and

the background. The primary manifestation is *unnatural blurring*, where the face appears excessively sharp relative to the environment, or hair edges exhibit uneven blurring that is inconsistent with realistic depth-of-field effects.

Background problems. This category encompasses three distinct manifestations: (1) *unreal background*, where the model generates meaningless or physically impossible textures; (2) *simplified background*, where models (notably SD3.5) produce overly simplistic or blurry regions that lack the intricate details of real environments; and (3) *unnatural lighting*, characterized by illumination inconsistencies within the background or between the subject and the scene, such as illogical light sources or absent shadows.

3.2.4. Accessory-related Artifacts. Deepfake models often have difficulties with rendering accessories to maintain semantic coherence, therefore leading to noticeable inconsistency issues that serve as crucial identification clues. There are three categories we observed.

Earring rendering anomalies. Earrings are small but highly diverse in design and material, often featuring complex reflective properties. Anomalies include *inconsistency between left and right* (differences in size, shape, or color) and *abnormal shapes* that do not conform to realistic designs or appear unnaturally attached to the earlobe.

Clothing rendering anomalies. Due to the immense variety in fabric materials and textures, clothing often exhibits *left-right inconsistency* in symmetrical parts like collars or shoulder patterns. Additionally, *abnormal shapes* may occur, such as impossible seams or textures that are inconsistent with known fabrics.

Eyeglasses rendering anomalies. As precisely structured objects, eyeglasses frequently exhibit *left-right inconsistency* in frame geometry and lens size. In some cases, they are *indistinguishable from the face*, where the frames unnaturally merge with the facial skin, indicating a failure to enforce clear object boundaries.

3.3. From DFAT to Training Materials

The taxonomy served as the basis for designing our training materials that were used to train participants where to look and what visual cues to inspect for deepfake images. We selected artifacts using three criteria: frequency in the sampled images, visual salience, and human recognizability. These criteria prioritized cues that could be shown visually and practiced consistently, including background consistency, overall style, facial texture, and accessory symmetry. Prior research has shown that targeted guidance can improve deepfake detection performance [23], [24], [25]. Building on this evidence, our training materials use DFAT as a structured vocabulary for teaching human-visible artifacts across different deepfake models.

3.3.1. Selecting Learnable Artifact Cues. DFAT showed that different deepfake models share several broad artifact domains, such as background consistency, overall style, facial texture, and accessory symmetry. We used these shared domains to design general-purpose training material that teaches participants a common vocabulary and inspection routine across models. At the same time, the manifestations of these artifacts vary by model and differ in how easily they can be recognized. For example, SD3.5 and GPT-5 images often exhibit global cues such as an overly polished or idealized visual style, while StyleGAN3 images may contain abnormal skin-texture patterns that become recognizable only under magnification. We therefore designed separate model-specific training materials to focus on the most recognizable cues associated with each deepfake model.

3.3.2. Constructing the Training Materials. We designed two sets of short instructional materials: one general-purpose set covering all five selected deepfake models, and one model-specific set for each model. The general-purpose set includes a five-minute video and ten interactive web-based practice questions. It was designed to cover a broader range of artifact domains across models and teach participants a shared vocabulary and inspection routine for deepfake faces. Each model-specific set includes a one- to three-minute video and five interactive practice questions focused on artifacts associated with the corresponding model, based on its artifact profile shown in Table 2. Having these two sets allowed us to compare whether broad artifact awareness and model-specific guidance can have different effects on improving human performance on deepfake detection (RQ3).

For each practice question, the system displayed an explanation of the correct answer to help participants connect the visual cue to the detection judgment. We chose short videos as the training format because they have been found effective for educating on security-related topics [53], [54], [55]. The videos follow a global-to-local inspection routine: they first guide viewers to inspect image-level cues, such as the background and overall style, and then ask them to zoom in on localized cues around the face and accessories. This ordering was intended to help participants form a consistent inspection routine rather than memorize isolated features.

4. Online Experiment

Building on the DFAT-guided training materials, we designed and conducted an online experiment to evaluate how human performance in detecting deepfakes and identifying subjective states varies across different models and training conditions. The main experiment addresses **RQ2** and **RQ3** by measuring detection performance for the five deepfake models before and after DFAT-guided training (general-purpose vs. model-specific). To address **RQ4**, we invited participants back five months later and tested retention and transfer on newly sampled deepfake images they had not seen before. This section describes the study design, procedure, stimuli, participants, measures, and statistical analyses.

4.1. Study Design

We used a mixed design with two between-subject factors and two within-subject factors. The between-subject

factors were training type (general-purpose vs. model-specific) and deepfake model condition (StyleGAN2, StyleGAN3, GRAM-HD, SD3.5, or GPT-5). Participants were randomly assigned to one training group and one model condition. The within-subject factors were testing session (pre-training vs. post-training) and presentation format (single-image vs. side-by-side).

The two presentation formats captured different detection settings. In the **single-image** format, participants judged one face image in isolation, reflecting scenarios such as social media profile photos or shared photographs where users usually lack an authentic reference. In the **side-by-side** format, participants compared one real and one deepfake image and were asked to identify the real one. This format captures reference-assisted detection, such as educational quizzes or platform-assisted settings where a verified image is shown next to a suspicious one [56]. Pairwise judgments are also widely used in subjective image-quality assessment [57], [58], [59]. Each participant completed both formats before and after training, with ten questions per format in each testing session. The follow-up study retained each participant's original training group and generative model.

We measured detection performance as the proportion of correct judgments in each task condition. For the single-image subtask, we also computed deepfake-image detection accuracy and real-image detection accuracy to examine whether training affected responses to deepfake and real images differently. After each testing session, participants reported task-specific confidence and stress levels on a five-point scale. These measures supported auxiliary analyses related to their subjective states before and after training.

4.2. Participants

Participants were recruited from *Prolific*, an online platform widely used by researchers for human experiments. We restricted eligibility to adults who were fluent in English to ensure comprehension of the task description. Before recruiting participants, we conducted a power analysis to calculate the target sample size for the planned mixed design. Using G*Power with detection accuracy as the primary outcome, the analysis assumed a small effect ($f = .10$), $\alpha = .05$, power $(1 - \beta) = .95$, 10 between-subject groups, and four repeated measurements, yielding a target sample size of 440 participants. Participants were informed about the purpose, structure, and approximate duration of the study before providing informed consent through the experiment website. To ensure data quality, we excluded participants who failed attention checks, had missing responses, or completed the tests in less than 6 minutes (see Appendix B).

A total of 608 *Prolific* users participated in the experiment. After applying the exclusion criteria, 524 valid participants remained for analysis: 278 in the general-purpose training group and 246 in the model-specific group, distributed across the five model conditions. On average, it took each participant 32.3 minutes (general-purpose) and 29.5 minutes (model-specific) to complete the main study. Based on demographic profiles, participants were categorized by gender (Male, Female), ethnicity (White, Black, Asian), and age group (Table 5 in Appendix B reports more details).

Five months later, 342 participants from the original valid sample (181 from the general-purpose group and 161 from the model-specific one) completed the follow-up study.

4.3. Procedure

4.3.1. Main Study. The main study had four phases: consent and demographics, pre-training test, DFAT-guided training with practice questions, and post-training test.

Consent and demographics. After reading the consent form, participants could either proceed or leave the study. They reported demographic information, including age, gender, and ethnicity. Before the first test, the system explained the task interface and response requirements.

Pre-training test. Participants completed 20 baseline questions before receiving any training:

- **Single-image** (10 questions): Participants viewed one image and judged whether it was a real photograph or a deepfake.
- **Side-by-side** (10 questions): Participants viewed two images, one real and one deepfake, and identified which image was real.

After the questions, participants reported their confidence and perceived stress using the self-assessment questions listed in Appendix C.2.

Training. Participants received either general-purpose or model-specific DFAT-guided training according to their assigned group. The training included a video and interactive practice questions. For each practice question, the system explained the reasons for the correct answer.

Post-training test. The post-training session used the same task structure as the pre-training session, with newly sampled images to reduce memory effects. Participants reported their confidence and perceived stress after completing the test. They also provided feedback on the training regarding perceived usefulness, memorable content, and open-ended comments (see Appendix C.3 for the list of questions).

4.3.2. Follow-up Study. Five months after the main study, we invited all participants in the final sample to complete a follow-up study. Without viewing the training materials again, participants repeated the post-training test with exactly the same materials as last time, including both single-image and side-by-side tasks. They then completed 20 additional single-image questions with newly sampled real and deepfake images. These additional trials tested whether previously learned cues could be transferred to unseen images. Subsequently, participants reported their confidence and perceived stress and completed a feedback questionnaire about visual cues, memorable training content, and open-ended comments (see Appendix C.4).

4.4. Test Stimuli

For the main study, each participant completed a total of 40 questions across the pre- and post-training tests. The

stimulus set consisted of 30 deepfake images created by the assigned model and 30 real photographs sourced from the FFHQ dataset [40]. All stimuli were 1024×1024 images. Participants assigned to the same model were tested on the same image set. Across participants, we randomized the presentation order and interleaved the two subtasks (single-image and side-by-side) to reduce order effects and task-mode biases. The pre-training and post-training tests used non-overlapping stimuli, so that participants did not see the same test image twice across sessions.

For the follow-up study, the retention test used the same task structure and image set as the post-training test in the main study. The additional single-image trials used newly sampled real and deepfake images from the corresponding model condition to assess transfer to unseen images.

4.5. Statistical Analyses

Before examining training effectiveness, we confirmed that the two training groups did not differ in demographic distribution (Gender: $\chi^2(1) < 0.01$, $p = .979$; Ethnicity: $\chi^2(2) = 1.30$, $p = .521$; Age: $\chi^2(2) = 4.84$, $p = .089$). Because accuracy scores were non-normally distributed (Shapiro–Wilk, $p < .05$), we used non-parametric analyses: Aligned Rank Transform (ART) ANOVA [60] for factorial designs and Wilcoxon signed-rank tests [61] for paired comparisons. We report partial η_p^2 and rank-biserial correlation (r_{rb}) as effect sizes, respectively. Unless stated otherwise, we apply Holm correction within each family of related post hoc or paired comparisons reported in Section 5.

5. Results

Building on the DFAT that addresses **RQ1**, this section reports results for **RQ2–RQ4**: model-dependent detectability, training effectiveness, and five-month retention of detection and subjective-state outcomes. Table 3 summarizes descriptive statistics of pre- and post-training detection accuracy in different conditions. We also report participants' feedback that contextualizes how participants perceived, used, and remembered the training cues.

5.1. Baseline Detection Performance

5.1.1. Baseline Detectability Is Model-Dependent. We use pre-training accuracy to assess **RQ2** because it captures unaided detection performance before training materials could have affected participants' judgments. A two-way ART ANOVA on pre-training accuracy showed a significant main effect of deepfake model ($F(4, 519) = 64.02$, $p < .001$, $\eta_p^2 = 0.33$). The image presentation format also had a significant main effect ($F(1, 519) = 35.21$, $p < .001$, $\eta_p^2 = 0.064$), with the side-by-side setting yielding higher overall accuracy. The deepfake model $\times$ format interaction was not significant ($F(4, 519) = 1.12$, $p = .347$, $\eta_p^2 = 0.009$), suggesting that the relative detectability of the models was broadly consistent across presentation formats.

TABLE 3: Descriptive statistics for detection accuracy.

Training materials	Format	Deepfake model	Detection accuracy: mean (SD)		N
			Pre-Training	Post-Training	
General-purpose	Side-by-side	StyleGAN2	62.0% (19.8%)	74.1% (18.4%)	54
		StyleGAN3	67.5% (25.8%)	81.6% (20.9%)	55
		GRAM-HD	79.2% (22.4%)	86.8% (15.9%)	60
		SD3.5	92.7% (12.9%)	93.2% (12.1%)	60
		GPT-5	80.8% (21.2%)	85.3% (18.7%)	49
	Single image	StyleGAN2	56.5% (22.8%)	67.6% (17.6%)	54
		StyleGAN3	65.3% (21.2%)	72.0% (19.2%)	55
		GRAM-HD	72.8% (18.7%)	77.3% (16.6%)	60
		SD3.5	88.0% (15.5%)	81.2% (12.8%)	60
		GPT-5	75.1% (22.5%)	75.7% (18.1%)	49
Model-specific	Side-by-side	StyleGAN2	62.0% (21.4%)	74.6% (17.1%)	50
		StyleGAN3	66.5% (25.6%)	89.2% (13.4%)	51
		GRAM-HD	77.8% (18.6%)	93.7% (9.5%)	49
		SD3.5	97.2% (5.7%)	98.8% (3.3%)	50
		GPT-5	89.1% (18.0%)	96.1% (10.4%)	46
	Single image	StyleGAN2	58.0% (17.5%)	68.4% (18.1%)	50
		StyleGAN3	65.1% (23.6%)	79.2% (15.7%)	51
		GRAM-HD	70.0% (19.7%)	83.3% (14.9%)	49
		SD3.5	89.4% (11.9%)	97.4% (6.3%)	50
		GPT-5	82.0% (18.6%)	93.9% (8.8%)	46

Figure 2 shows the baseline patterns of detection performances for the five models: images generated by SD3.5 and GPT-5 were easier to detect, images generated by GRAM-HD fell in the middle, and images generated by StyleGAN2 and StyleGAN3 were more difficult. Holm-corrected post hoc comparisons supported this broad ordering across presentation formats. The adjacent differences among the upper three groups were reliable in both formats: SD3.5 had higher baseline accuracy than GPT-5 ($p_{\mathrm{adj}} \leq .0019$), GPT-5 had higher baseline accuracy than GRAM-HD ($p_{\mathrm{adj}} \leq .032$), and GRAM-HD had higher baseline accuracy than StyleGAN2 ($p_{\mathrm{adj}} < .001$). The boundary between GRAM-HD and StyleGAN3 depended on presentation format: the difference was reliable in the side-by-side format ($p_{\mathrm{adj}} = .005$), but did not survive correction in the single-image format ($p_{\mathrm{adj}} = .074$). These results align with predictions based on DFAT: higher baseline accuracy for SD3.5 and GPT-5 is consistent with global cues, such as an overly polished or idealized visual style, being easier for untrained observers to notice; lower baseline accuracy for StyleGAN-based models was linked to texture-level and localized artifacts that are considered harder to identify without explicit guidance.

Two secondary analyses further bounded this baseline pattern. First, side-by-side references improved baseline accuracy for four of the five deepfake models. The format advantage was largest for SD3.5 ($p < .001$, $r_{\mathrm{rb}} = 0.74$) and was also significant for GPT-5, GRAM-HD, and StyleGAN2; StyleGAN3 was the only exception ($p = .398$, $r_{\mathrm{rb}} = 0.11$). Second, demographic effects were limited. Age was the only variable with a consistent main effect across both presentation formats (both ps $< .001$), with participants aged 18–24 achieving higher baseline accuracy than older groups, especially for GPT-5, GRAM-HD, and SD3.5. Gender and ethnicity showed no significant main effects (all ps $\geq .076$). Although we observed one deepfake model $\times$ gender interaction in the side-by-side format ($p < .001$, $\eta_p^2 = 0.04$), it was model-specific and did not indicate a

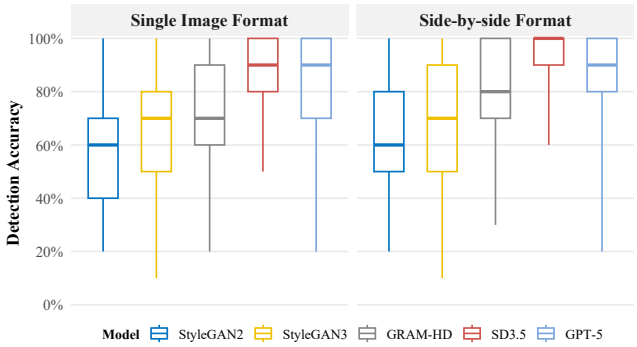


Figure 2: Baseline detection accuracy across deepfake models and presentation formats. Box plots display median, interquartile range, and full range.

consistent gender advantage.

Finding 1 (RQ2): Baseline detection performance varied across deepfake models with distinctive human-recognizable artifacts. Deepfakes generated by more recent and advanced models were not necessarily harder to detect, e.g., participants detected SD3.5 and GPT-5 images more accurately than StyleGAN2 and StyleGAN3 ones, and StyleGAN2 images were harder than StyleGAN3 ones. Side-by-side references improved baseline accuracy, but model-dependent differences remained across presentation formats.

5.2. Training Effects on Detection Performance

DFAT-guided training improved detection in most conditions, but the size of the gain depended on the deepfake model and the training material. We quantified training gains as post-training accuracy minus pre-training accuracy for each participant and compared gains across deepfake models, training-material types, and presentation formats.

5.2.1. Training Gains Are Uneven Across Models. Holm-corrected Wilcoxon signed-rank tests showed significant accuracy gains in 13 of the 20 experimental conditions. Figure 3 summarizes the model-level gain patterns: more consistent gains were observed for GAN-based models, whereas gains for SD3.5 and GPT-5 were more conditional. StyleGAN3 showed the largest improvement with model-specific training (side-by-side: $+22.8\%$, $p < .001$, $r_{\mathrm{rb}} = 0.87$). For SD3.5, baseline accuracy was already high ($\geq 88\%$), and significant gains appeared only in the single-image format with model-specific training ($+8.0\%$, $p = .001$, $r_{\mathrm{rb}} = 0.96$). General-purpose training was associated with a decrease for SD3.5 in the single-image condition (-6.8%, $p = .015$, $r_{\mathrm{rb}} = -0.53$), consistent with general artifact cues being less well matched to SD3.5 images. For GPT-5, significant gains appeared with model-specific training, while general-purpose training produced no significant improvement.

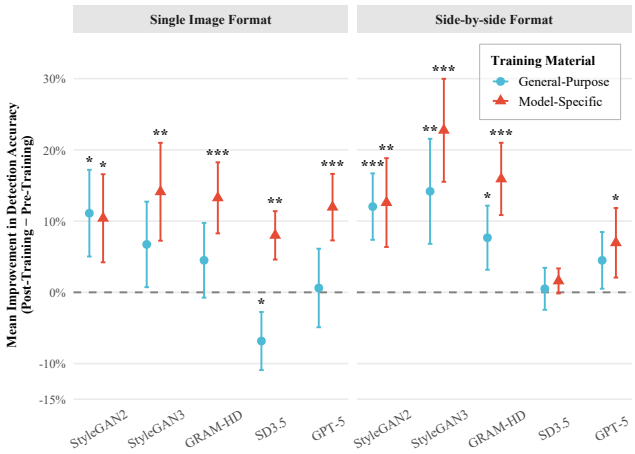


Figure 3: Mean improvement in detection accuracy (post-training minus pre-training) by deepfake model, training material, and presentation format. Error bars show 95% confidence intervals; significance markers denote Holm-corrected Wilcoxon signed-rank tests for pre–post changes.

Two-way ART ANOVAs showed that the gains depended on deepfake model ($ps < .05$). StyleGAN3 showed the largest average improvement (14.5%). This improvement was similar to the gains for StyleGAN2 and GRAM-HD. It was larger than the gains for SD3.5 in most conditions ($p_{\mathrm{adj}} \leq .016$), and larger than the GPT-5 gain only under model-specific training in the side-by-side format ($p_{\mathrm{adj}} = .0145$). These results suggest that participants benefited more consistently from training on GAN-related artifacts, whereas gains for SD3.5 and GPT-5 were more conditional, partly because their baseline accuracy was already high.

5.2.2. Model-Specific Training Helped Most in Single-Image Subtasks. Model-specific training provided the clearest advantage when participants could not rely on a side-by-side reference. Two-way ART ANOVAs comparing gains by training-material type showed significant training-material type $\times$ time (pre- and post-training) interactions in three of the ten model-format combinations: GRAM-HD in the single-image format ($F(1, 107) = 5.95$, $p = .016$, $\eta_p^2 = 0.053$), SD3.5 in the single-image format ($F(1, 108) = 65.74$, $p < .001$, $\eta_p^2 = 0.378$), and GPT-5 in the single-image format ($F(1, 93) = 8.03$, $p = .006$, $\eta_p^2 = 0.080$).

For GAN-based models, direct comparisons of performance gains between general-purpose and model-specific training were not statistically significant after Holm correction (all adjusted p values $> .10$). Descriptively, model-specific training showed numerical advantages for StyleGAN3 (+8.6% in side-by-side and +7.4% in single-image) and GRAM-HD (+8.3% in side-by-side and +8.8% in single-image). These differences were not significant in direct gain comparisons, possibly reflecting limited sample sizes and high variance in participant performance.

The training-material difference was clearest for SD3.5

and GPT-5 in the single-image format. For SD3.5, model-specific training outperformed general-purpose training by 14.8% ($p_{\mathrm{adj}} < .001$, $r_{\mathrm{rb}} = 0.58$), with a 6.8% performance decrease under general-purpose training and an 8.0% increase under model-specific training. Similarly, GPT-5 benefited more from model-specific training in the single-image condition (difference in gains $= +11.3\%$, $p_{\mathrm{adj}} = .022$, $r_{\mathrm{rb}} = 0.35$), whereas general-purpose training produced almost no improvement (+0.6%).

5.2.3. Format and Demographic Effects Are Limited. Presentation format did not substantially change the training-gain pattern. Two-way ART ANOVAs conducted separately for each deepfake model and training-material type showed significant Time × Format interactions in three of the ten conditions. Holm-corrected post hoc comparisons showed significant format-dependent gain differences only for SD3.5. Under general-purpose training, side-by-side presentation yielded larger gains than the single-image format ($p_{\mathrm{adj}} = .017$, $r_{\mathrm{rb}} = 0.51$); under model-specific training, the pattern reversed, with larger gains in the single-image format ($p_{\mathrm{adj}} = .008$, $r_{\mathrm{rb}} = -0.81$). Gains were comparable across formats for all other deepfake models.

Demographic variables showed no consistent relationship with training gains. ART ANOVAs tested interactions between age, gender, ethnicity, and training gains across deepfake models, training material types, and presentation formats. Across 60 tests, only six interactions reached $p < .05$, and these effects did not form a consistent pattern across demographic variables or task conditions. Thus, although age predicted baseline accuracy, we found little evidence that training gains systematically differed across demographic groups. Such seemingly negative results deliver a positive signal: when training worked, it worked across different demographic groups.

Finding 2 (RQ3): DFAT-guided training improved participants' detection accuracy in most conditions, and model-specific training yielded substantive advantage over general-purpose training in the single-image format. The size of the improvement varied across deepfake models and training types: GAN-based models showed more consistent gains, while SD3.5 and GPT-5 improved most clearly under model-specific training.

5.3. Training Effects on Detection Behavior and Subjective States

Aggregate training gains show whether overall accuracy improved, but they do not reveal whether gains came from better recognition of deepfake images, better recognition of real images, or both. To characterize how training changed single-image judgments, we decomposed accuracy by image type and analyzed confidence and perceived stress as subjective correlates of detection performance.

5.3.1. Training Shifted Sensitivity Toward Deepfake Images. DFAT-guided training improved deepfake-image accuracy more consistently than real-image accuracy. As shown in Figure 4, deepfake-image accuracy increased in nine of the ten conditions after Holm correction. The only non-significant case was SD3.5 under general-purpose training, where baseline accuracy was already above 90%. Real-image accuracy changed less consistently than deepfake-image accuracy. Under general-purpose training, real-image accuracy decreased for SD3.5 ($p_{\mathrm{adj}} < .001$) and for StyleGAN3 and GPT-5 (both $p_{\mathrm{adj}} = .048$). Under model-specific training, we observed no Holm-corrected decreases in real-image accuracy. Overall, training increased sensitivity to deepfake cues, and general-purpose training increased false positives on real images for a subset of deepfake models.

ART ANOVAs showed significant image type × time interactions for all five deepfake models under general-purpose training (ps $< .01$), but only for StyleGAN2 and StyleGAN3 under model-specific training (ps $< .001$). Post-training comparisons showed a consistent asymmetry: deepfake-image accuracy was higher than real-image accuracy in all conditions (ps $< .05$). These results suggest that training primarily increased participants' sensitivity to deepfake images, while real-image judgments remained stable except under general-purpose training.

5.3.2. Confidence and Stress Level Change After Training. We measured task-specific confidence and stress self-reported by participants after each testing session (5-point scales). In the main study ($n = 524$), model-specific training increased confidence (pre-training: $M = 3.41$, post-training: $M = 3.78$; Holm-corrected $p_{\mathrm{adj}} < .001$). General-purpose training showed a smaller confidence change that did not survive Holm correction (pre-training: $M = 3.49$, post-training: $M = 3.57$; $p_{\mathrm{adj}} = .143$).

Self-reported stress level increased after general-purpose training (pre-training: $M = 2.16$, post-training: $M = 2.38$; $p_{\mathrm{adj}} < .001$). Model-specific training showed no Holm-corrected stress increase (pre-training: $M = 2.09$, post-training: $M = 2.00$; $p_{\mathrm{adj}} = .143$). As a whole, the results suggest that training can increase participants' confidence, while its effects on self-reported subjective states vary by training type. Considering the training can improve detection performance in most conditions, the overall effect of training is therefore positive in most cases or neutral in a few cases.

Finding 3 (RQ3): DFAT-guided training effectively improved participants' performance in detecting deepfake images, with detection accuracy increased in nine out of ten single-image conditions. Under general-purpose training, real-image detection accuracy decreased for certain models, while this decrease was not found under model-specific training. Model-specific training increased confidence without increasing stress, whereas general-purpose training produced a small but significant stress increase.

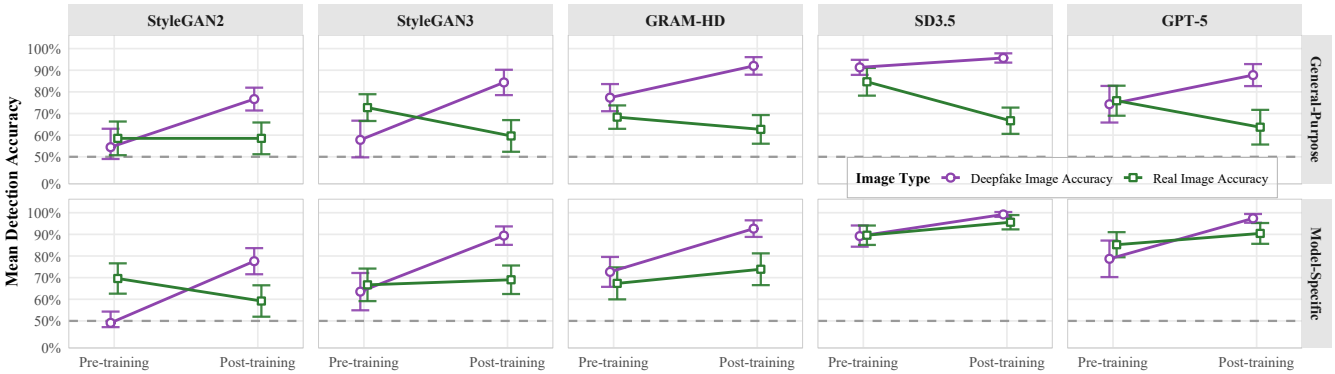


Figure 4: Pre–post changes in deepfake-image and real-image detection accuracy in the single-image task. Points show mean accuracy; error bars indicate 95% confidence intervals.

5.4. Training Effects After Five Months

Five months after the main experiment, 342 participants completed the follow-up study. To answer **RQ4**, we compared immediate post-training and follow-up results for detection accuracy, newly sampled single-image trials, and self-reported confidence and stress levels.

5.4.1. Detection Gains Showed Little Decay. The retention test repeated the post-training test structure, allowing paired comparisons under the same deepfake-model condition and presentation format. As shown in Figure 5, follow-up accuracy generally remained close to immediate post-training accuracy and above the corresponding pre-training baseline.

Newly sampled single-image trials showed a similar descriptive pattern. Because these trials used real and deepfake images that were not matched to the stimuli from the main experiment, we interpret them as descriptive evidence of transfer rather than as part of the paired retention test.

For the formal retention analysis, we used paired Wilcoxon signed-rank tests to compare retention-test and immediate post-training accuracy for each combination of deepfake model, training-material type, and presentation format. Mean changes from post-training to follow-up were small, ranging from -7.3% to $+6.4\%$. After Holm correction, none of the 20 paired comparisons showed evidence of a significant decline in detection accuracy. The largest numerical decrease occurred for StyleGAN2 under general-purpose training in the side-by-side format (-7.3%; post-training: $M = 76.3\%$, retention: $M = 69.0\%$), whereas the largest numerical increase occurred for SD3.5 under general-purpose training in the single-image format ($+6.4\%$; post-training: $M = 79.8\%$, retention: $M = 86.2\%$). Neither change was significant after correction.

5.4.2. Confidence and Stress Levels Remained Stable. We compared subjective-state ratings between the immediate post-training test and the five-month follow-up. Confidence showed no Holm-corrected decline in the general-purpose group (post-training: $M = 3.49$, follow-up: $M = 3.34$; $p_{\text{adj}} = .071$). The model-specific group showed a modest decline (post-training: $M = 3.67$, follow-up: $M = 3.41$; $p_{\text{adj}} = .001$). The self-reported stress level did not show a Holm-corrected increase in either training group (general-purpose: post-training $M = 2.46$, follow-up $M = 2.28$, $p_{\text{adj}} = .089$; model-specific: post-training $M = 2.11$, follow-up $M = 2.20$, $p_{\text{adj}} = 1.000$). These results reconfirmed that applying learned cues after five months did not increase stress.

Finding 4 (RQ4): The results show that the effects of DFAT-guided training persisted over five months. Detection accuracy did not decline significantly in any paired comparison, and the new image tests showed a similar pattern. Self-reported stress showed no increase at follow-up, while confidence exhibited limited decay for those who received model-specific training.

5.5. Participant Feedback

In total, 520 (out of 524) participants in the main study and 341 (out of 342) participants in the five-month follow-up provided feedback after completing the test. We analyzed these responses to understand how participants perceived the training, which visual artifacts they used after training, and which ones remained salient five months later.

Most participants perceived the training as useful. In the main study, 89.9% of the participants reported that the training improved their ability to identify deepfake face images. Participants valued both training components: 83.4% considered the training video useful, 68.3% considered the practice tasks useful, and 52.9% selected both. Among the artifacts participants relied on during the tests, the most frequently reported were unnatural facial skin (71.9%), background problems (69.7%), accessory rendering anomalies (68.3%), overly perfect and idealized style (66.2%), and unnatural hair (63.4%), which are aligned with our DFAT.

Several cues remained salient after five months. Five months later, participants still reported most frequently using unnatural facial skin (78.4%) and overly perfect and idealized style (74.3%) as cues. Many participants also retained

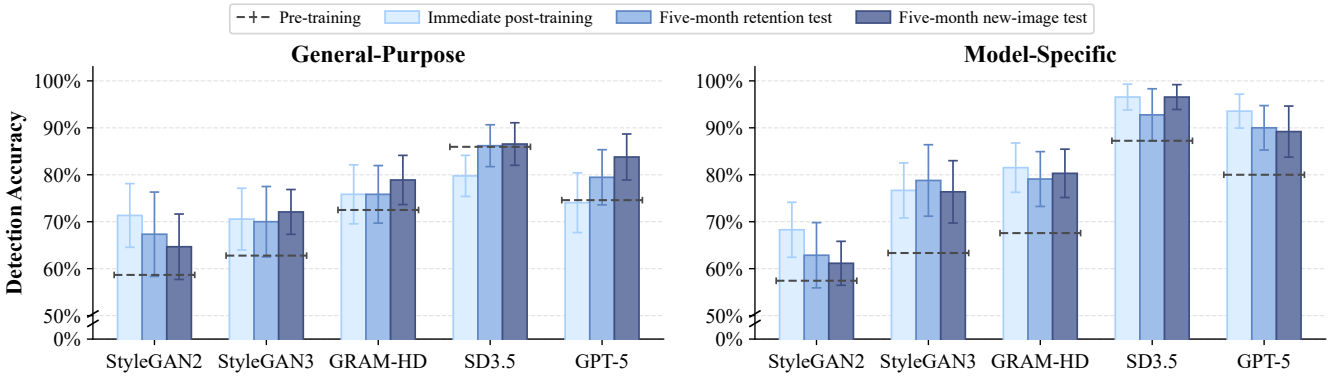


Figure 5: Post-training detection accuracy across models and training material types: the main experiment (immediately after training), the follow-up test and the new-image test (five months later). Error bars indicate standard errors; annotations show changes from immediate post-training to five-month retention.

some training memory, with 62.3% remembering at least some content and 26.9% remembering most of it or very well. Among those who described what they remembered, 73.9% mentioned artifact cues, 65.1% mentioned global or style cues, and 42.6% mentioned local or skin-related cues. These responses suggest that the effect of the DFAT-guided training is sustainable over a relatively long period of time.

Training was described as clear, useful, and informative. After excluding empty entries and replies such as "N/A" or "nothing", two co-authors conducted keyword-based content analysis on 147 main-study and 97 follow-up responses. In the main study, 59.2% of the responses were positive or described the study as clear, and 25.9% described the training as useful or informative. Another 13.6% pointed to the difficulty of the task, often noting that the face images appeared realistic. A similar pattern was observed in the follow-up study: 45.4% were positive about the training and its clarity, and 9.3% mentioned realistic-looking images and difficulty of distinguishing them from real faces. Overall, the feedback proves that our deepfake training materials were well-received and valued by the participants.

6. Discussion

6.1. Key Findings

Human performance in detecting deepfakes differs across generative models. Existing literature shows that untrained users often struggle to identify high-quality deepfake faces, with several GAN-based studies reporting near-chance performance [12], [21], [22]. These findings suggest that deepfake detection is challenging for humans. Our study revealed two key insights: this difficulty varies across generative models and deepfakes generated by more recent and advanced models may not necessarily be harder to detect. Such results can be well explained based on our DFAT, which identifies distinct visible artifacts produced by the generative models that show varying levels of human-recognizability. Local and subtle artifacts are especially difficult for untrained users to notice, resulting in low detection performance for relevant models as reported in prior works. This finding contributes to the existing literature and highlights the need for cross-model evaluation of human performance in detecting deepfakes.

Model-specific deepfake training produces better outcomes than general-purpose training. It is known that feedback, tutorials, AI support, and artifact-based guidance can improve detection, although their effects vary across settings [23], [24], [25], [26], [27]. Recent studies have also shown that human-recognizable artifacts in AI-generated images can be identified and categorized [51], [52]. Our results reveal that matching training content with the target generative model produces better training outcomes. This is likely due to the fact that the general-purpose training included mixed artifacts from different models, potentially increasing participants' cognitive load and undermining the training effectiveness. This finding provides practical guidance for future development of model-specific deepfake training interventions.

Artifact-based training has lasting value. Prior studies involving deepfake training mainly evaluated immediate post-training effects [20]. It remains unclear whether users retain learned artifact cues and maintain improved detection performance over a longer period of time. Our follow-up study addresses this gap: after five months, detection accuracy remained stable, while confidence level showed only limited decline. Participant feedback also shows that many users could still recall and use the artifact cues learned during the training five months ago. These results suggest that artifact-based training can yield sustainable improvement in deepfake detection.

6.2. Practical Implications

User training remains an important layer of defense. Platforms can use automated deepfake detectors to identify likely AI-generated face images and display labels or warnings to users. Prior work shows that labels are not

always effective; their effect depends on design, user trust, and user interpretation [62], [63]. In practice, images are often screenshotted, cropped, compressed, or reposted to other platforms. During this process, labels may be lost, and image sources may become hard to verify. Users still need to judge whether online images are trustworthy on social networks, often with limited support to verify the source or context [64]. Our work shows that user training can complement platform defenses by helping users make more evidence-based judgments when external cues are missing.

DFAT provides concrete inspection guidance. Many users know that AI images may contain errors, yet they lack clear guidance on where to look and how to interpret visible anomalies. DFAT gives concrete inspection dimensions for training: error-prone regions, local texture, overall style, and contextual consistency. These dimensions allow training materials to move beyond general warnings and point users to specific visual cues. This structure makes deepfake training easier to teach and easier for non-expert users to remember. It also aligns with recent work showing that targeted deepfake literacy interventions can improve users' ability to reason about manipulated media [19]. DFAT therefore offers a practical way to organize platform prompts, educational materials, and content moderation guidelines.

Model-specific training can reduce misjudgment. Different generative models produce different types of visual artifacts. Overly broad training may lead users to treat a small set of common artifacts as general signs of AI generation, which can increase mistakes on real images. Model-specific training helps users focus on more relevant visual cues and allows training materials to be updated as generative models change. This matters for deployment because user education should raise awareness of fake images while limiting unnecessary suspicion of real ones. Prior work on uncertainty and confidence in decision support also shows that user-facing information can shape trust and confidence under uncertain judgments [65], [66]. DFAT supports this goal by grounding user judgment in concrete model-specific visual evidence.

6.3. Extending DFAT

DFAT can help estimate detection difficulty across models. Different generative models leave different visual traces, which can affect how easily users can detect them. Prior work showed that artifact types and photorealism shape human detection of AI-generated images [52]. Our results are consistent with this pattern: models with clear global style problems were easier to detect without training, while models whose errors mainly appeared in local texture required training to educate users about them. This suggests that DFAT can help researchers use artifact patterns to form early expectations about which models may be easier or harder for users to detect and what artifacts may be easier or harder to learn. These expectations still need validation through user studies, and actual human performance should be confirmed in real detection tasks.

DFAT can serve as an updatable analysis framework. Generative models are constantly evolving. Recent work has shown new methods for removing visual imperfections and improving scene consistency in AI-generated images [67], [68]. A fixed list of artifacts will become outdated quickly. DFAT is designed to be an easily extensible taxonomy. As new models appear, researchers can add new artifact types, refine existing categories, and update user training accordingly. In this way, DFAT can grow from a one-time tool into a dynamic framework for future research and education activities. Future work can also build DFAT-guided training into interactive modules or assistive interfaces that guide users during image inspection [69], [70].

6.4. Limitations

While our work studied five generative models, AI-based image generation techniques continue to change quickly, requiring constant updates to the DFAT and training materials. Future work should examine more models and track how artifacts change as generative methods evolve.

Emphasizing controlled comparisons, our online experiment did not cover real-world use cases. To ensure generalizability, larger and more diverse samples are warranted. Although the single-image and side-by-side tasks helped isolate user detection behaviors, face images often appear in social feeds, messaging threads, or news stories. In these settings, users may be affected by context, purpose, and time pressure, which call for more complicated experiment designs in future research. Participant dropout may also introduce selection bias into the five-month retention findings.

DFAT currently focuses on fully generated, photorealistic face images, because this type of deepfakes is among the most abused ones in real-world environment affecting identity trust, social interactions, and online safety. Future research is encouraged to investigate other types of deepfake images and media types as the artifacts involved will differ and the effects of training may change. To facilitate future research, DFAT can be easily extended to support other generative models and incorporate corresponding artifacts across diverse deepfake detection tasks [18], [20].

7. Conclusion

In this paper, we conducted a comprehensive evaluation of human performance in detecting deepfake face images across five representative generative models following three different architectures. To systematically characterize visual anomalies, we introduced the Deepfake Face Artifact Taxonomy (DFAT), based on which we designed two types of training interventions (general-purpose and model-specific). Our online experiment produced new insights on human performance being strongly model-dependent and images generated by some more recent and advanced models being easier to detect by humans than their older less advanced versions. We further demonstrated that DFAT-guided training materials significantly improved detection accuracy, especially when they were model-specific. The improved accuracy did not come with decreased user confidence or increased stress. We also observed that such training effects

persisted after five months, indicating strong sustainability of the training interventions.

Based on the empirical evidence reported, we argue that efforts to raise public awareness should shift from generic alerts to concrete, model-specific and artifact-focused interventions. As deepfake models continue to correct visual imperfections, visual inspection alone may not suffice. Instead of relying solely on human perception, future defenses should integrate technical scaffolding in the human detection process. We advocate implementing a "Reference Mandate" in user interfaces that provides authentic baselines for side-by-side comparison. We also encourage the development of active web browser and app plugins that can guide user attention to harder-to-recognize (easier-to-miss) anomalies. As a promising approach to complementing technical solutions, taxonomy-guided public education offers a path to securing the information ecosystem against increasingly hyper-realistic deepfakes.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (grant numbers 62302427 and 62462060), China's Central Government Guides Local Science and Technology Development Fund (grant number ZYYD2026ZY21), and the Tianshan Talent Training Program (grant number 2022TSYCLJ0036). We also thank the five annotators for contributing to DFAT annotations.

References

[1] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media + Society*, vol. 6, no. 1, 2020. [Online]. Available: https://doi.org/10.1177/2056305120903408

[2] Furizal, A. Ma'arif, H. Maghfiroh, I. Suwarno, D. Prayogi, Kariyamin, S. Lonang, and A.-N. Sharkawy, "Social, legal, and ethical implications of ai-generated deepfake pornography on digital platforms: A systematic literature review," *Social Sciences & Humanities Open*, vol. 12, pp. 101 882:1–101 882:25, 2025. [Online]. Available: https://doi.org/10.1016/j.ssaho.2025.101882

[3] T. Li, K. M. Reyes, L. Jimenez, K. S. Nam, and D. Y. Wohn, "Understanding educators' perceptions of ai-generated non-consensual intimate imagery," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2026. [Online]. Available: https://doi.org/10.1145/3772318.3790584

[4] L. S. F. Lin, "Examining the role of deepfake technology in organized fraud: Legal, security, and governance challenges," *Frontiers in Law*, vol. 4, pp. 6–17, 2025. [Online]. Available: https://doi.org/10.6000/2817-2302.2025.04.02

[5] S. Alexander, "Deepfake cyberbullying: The psychological toll on students and institutional challenges of ai-driven harassment," *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, vol. 98, no. 2, pp. 36–50, 2025. [Online]. Available: https://doi.org/10.1080/00098655.2025.2488777

[6] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, M. Cornia, and R. Cucchiara, "The revolution of multimodal large language models: A survey," in *Findings of the Association for Computational Linguistics: ACL 2024*. ACL, 2024, pp. 13 590–13 618. [Online]. Available: https://doi.org/10.18653/v1/2024.findings-acl.807

[7] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, pp. nwae403:1–nwae403:20, 2024. [Online]. Available: https://doi.org/10.1093/nsr/nwae403

[8] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024. [Online]. Available: https://doi.org/10.1109/TPAMI.2024.3369699

[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 8110–8119. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00813

[10] Z. Lu, D. Huang, L. Bai, J. Qu, C. Wu, X. Liu, and W. Ouyang, "Seeing is not always believing: Benchmarking human and model perception of AI-generated images," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25 435–25 447, 2023.

[11] "Midjourney," Online model, 2025, accessed: 2025-11-17. [Online]. Available: https://www.midjourney.com/

[12] D. Cooke, A. Edwards, S. Barkoff, and K. Kelly, "As good as a coin toss: Human detection of AI-generated content," *Communications of the ACM*, vol. 68, no. 10, pp. 100–109, 2025. [Online]. Available: https://doi.org/10.1145/3729417

[13] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 7:1–7:41, 2021. [Online]. Available: https://doi.org/10.1145/3425780

[14] A. Dixit, N. Kaur, and S. Kingra, "Review of audio deepfake detection techniques: Issues and prospects," *Expert Systems*, vol. 40, no. 8, pp. e13 322:1–e13 322:19, 2023. [Online]. Available: https://doi.org/10.1111/exsy.13322

[15] J. Pu, Z. Sarwar, S. M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, and B. Viswanath, "Deepfake text detection: Limitations and opportunities," in *Proceedings of the 2023 IEEE Symposium on Security and Privacy*. IEEE, 2023, pp. 1613–1630. [Online]. Available: https://doi.org/10.1109/SP46215.2023.10179387

[16] A. Uchendu, S. Venkatraman, T. Le, and D. Lee, "Catch me if you GPT: Tutorial on deepfake texts," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*. ACL, 2024, pp. 1–7. [Online]. Available: https://doi.org/10.18653/v1/2024.naacl-tutorials.1

[17] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A survey on speech deepfake detection," *ACM Computing Surveys*, vol. 57, no. 7, pp. 165:1–165:38, 2025. [Online]. Available: https://doi.org/10.1145/3714458

[18] J. Wang, Z. Li, C. Zhang, J. Chen, Z. Wu, L. S. Davis, and Y.-G. Jiang, "Fighting malicious media data: A survey on tampering detection and deepfake detection," *Proceedings of the IEEE*, vol. 113, no. 3, pp. 287–311, 2025. [Online]. Available: https://doi.org/10.1109/JPROC.2025.3576367

[19] D. Geissler, C. Robertson, and S. Feuerriegel, "Designing effective digital literacy interventions for boosting deepfake discernment," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2026. [Online]. Available: https://doi.org/10.1145/3772318.3790428

[20] A. Diel, T. Lalgi, I. C. Schröter, K. F. MacDorman, M. Teufel, and A. Bäuerle, "Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers," *Computers in Human Behavior Reports*, vol. 16, pp. 100 538:1–100 538:13, 2024. [Online]. Available: https://doi.org/10.1016/j.chbr.2024.100538

[21] J. Frank, F. Herbert, J. Ricker, L. Schönherr, T. Eisenhofer, A. Fischer, M. Dürmuth, and T. Holz, "A representative study on human detection of artificially generated media across countries," in *Proceedings of the 2024 IEEE Symposium on Security and Privacy*. IEEE, 2024, pp. 55–73. [Online]. Available: https://doi.org/10.1109/SP54263.2024.00159

[22] B. Shen, B. RichardWebster, A. O'Toole, K. Bowyer, and W. J. Scheirer, "A study of the human perception of synthetic faces," in *Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2021. [Online]. Available: https://doi.org/10.1109/FG52635.2021.9667066

[23] A. Boyd, P. Tinsley, K. Bowyer, and A. Czajka, "The value of AI guidance in human examination of synthetically-generated faces," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 5, pp. 5930–5938, 2023. [Online]. Available: https://doi.org/10.1609/aaai.v37i5.25734

[24] N. Hulzebosch, S. Ibrahimi, and M. Worring, "Detecting CNN-generated facial images in real-world scenarios," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2020, pp. 642–643. [Online]. Available: https://doi.org/10.1109/CVPRW50498.2020.00329

[25] A. Diel, A. Bäuerle, and M. Teufel, "Inability to detect deepfakes: Deepfake detection training improves detection accuracy, but increases emotional distress and reduces self-efficacy," Preprint, 2024. [Online]. Available: https://doi.org/10.31219/osf.io/muwnj

[26] S. J. Nightingale and H. Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, pp. e2 120 481 119:1–e2 120 481 119:3, 2022. [Online]. Available: https://doi.org/10.1073/pnas.2120481119

[27] S. D. Bray, S. D. Johnson, and B. Kleinberg, "Testing human ability to detect 'deepfake' images of human faces," *Journal of Cybersecurity*, vol. 9, no. 1, pp. tyad011:1–tyad011:18, 2023. [Online]. Available: https://doi.org/10.1093/cybsec/tyad011

[28] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/076ccd93ad68be51f23707988e934906-Abstract.html

[29] J. Xiang, J. Yang, Y. Deng, and X. Tong, "GRAM-HD: 3D-consistent image generation at high resolution with generative radiance manifolds," in *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision*. IEEE, 2023, pp. 2195–2205. [Online]. Available: https://doi.org/10.1109/ICCV51070.2023.00209

[30] Stability AI, "Introducing Stable Diffusion 3.5," News release, 2024, accessed: 2025-08-03. [Online]. Available: https://stability.ai/news/introducing-stable-diffusion-3-5

[31] ——, "stabilityai/stable-diffusion-3.5-large," Online model at Hugging Face, 2024, accessed: 2025-08-03. [Online]. Available: https://huggingface.co/stabilityai/stable-diffusion-3.5-large

[32] OpenAI, "GPT-5 is here," Web page, 2025, accessed: 2025-09-07. [Online]. Available: https://openai.com/gpt-5/

[33] A. Kammoun, R. Slama, H. Tabia, T. Ouni, and M. Abid, "Generative adversarial networks for face generation: A survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 94:1–94:37, 2022. [Online]. Available: https://doi.org/10.1145/3527850

[34] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proceedings of the 2018 International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk99zCeAb

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the 2022 IEEE/CVF conference on computer vision and pattern recognition*. IEEE, 2022, pp. 10 684–10 695. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01042

[36] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *Proceedings of the 12th International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=di52zR8xgf

[37] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," Preprint arXiv:2204.06125, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.06125

[38] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 2003–2011. [Online]. Available: https://doi.org/10.1145/3394171.3413630

[39] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8789–8797. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00916

[40] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 4401–4410. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00453

[41] L. Chai, M. Gharbi, E. Shechtman, P. Isola, and R. Zhang, "Any-resolution training for high-resolution image synthesis," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*. Springer, 2022, pp. 170–188. [Online]. Available: https://doi.org/10.1007/978-3-031-19787-1_10

[42] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. de Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3D generative adversarial networks," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 16 123–16 133. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01565

[43] Q. Xu, W. Mu, J. Li, T. Sun, and X. Jiang, "Advancements in ai-generated content forensics: A systematic literature review," *ACM Comput. Surv.*, vol. 58, no. 3, 2025. [Online]. Available: https://doi.org/10.1145/3760526

[44] Z. Guo, L. Wang, W. Yang, G. Yang, and K. Li, "LDFnet: Lightweight dynamic fusion network for face forgery detection by integrating local artifacts and global texture information," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 1255–1265, 2023. [Online]. Available: https://doi.org/10.1109/TCSVT.2023.3289147

[45] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2018, pp. 46–52. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Li_Exposing_DeepFake_Videos_By_Detecting_Face_Warping_Artifacts_CVPRW_2019_paper.html

[46] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, pp. 5001–5010. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00505

[47] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*. Springer, 2020, pp. 86–103. [Online]. Available: https://doi.org/10.1007/978-3-030-58610-2_6

[48] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 2185–2194. [Online]. Available: https://doi.org/10.1109/CVPR46437.2021.00222

[49] Z. Ba, Q. Liu, Z. Liu, S. Wu, F. Lin, L. Lu, and K. Ren, "Exposing the deception: Uncovering more forgery clues for deepfake detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, pp. 719–728, 2024. [Online]. Available: https://doi.org/10.1609/aaai.v38i2.27829

[50] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 2382–2390. [Online]. Available: https://doi.org/10.1145/3394171.3413769

[51] N. Kamali, K. Nakamura, A. Chatzimparmpas, J. Hullman, and M. Groh, "How to distinguish ai-generated images from authentic photographs," Northwestern University, WorkingPaper, Jun. 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2406.08651

[52] N. Kamali, K. Nakamura, A. Kumar, A. Chatzimparmpas, J. Hullman, and M. Groh, "Characterizing photorealism and artifacts in diffusion model-generated images," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3706598.3713962

[53] X. Chen, V. Distler, C. Gordon, Y. Yao, and Z. Teuber, "Empowering parents to support children's online security and privacy: Findings from a randomized controlled trial," in *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2025, pp. 1934–1948. [Online]. Available: https://doi.org/10.1145/3719027.3765214

[54] S. Kaushik, Y. Yao, P. Dewitte, and Y. Wang, ""How I know for sure": People's perspectives on solely automated decision-making (SADM)," in *Proceedings of the 17th Symposium on Usable Privacy and Security*. USENIX, 2021, pp. 159–180. [Online]. Available: https://www.usenix.org/conference/soups2021/presentation/kaushik

[55] B. M. Berens, M. Mossano, and M. Volkamer, "Taking 5 minutes protects you for 5 months: Evaluating an anti-phishing awareness video," *Computers & Security*, vol. 137, pp. 103 620:1–103 620:19, 2024. [Online]. Available: https://doi.org/10.1016/j.cose.2023.103620

[56] J. West and C. Bergstrom, "Which face is real?" 2019, university of Washington, Calling Bullshit project. Accessed: 2025-12-18. [Online]. Available: https://www.whichfaceisreal.com/

[57] Z. Zhang, W. Sun, Y. Zhou, J. Jia, Z. Zhang, J. Liu, X. Min, and G. Zhai, "Subjective and objective quality assessment for in-the-wild computer graphics images," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 4, pp. 96:1–96:22, 2023. [Online]. Available: https://doi.org/10.1145/3631357

[58] S. Mohammadi and J. a. Ascenso, "Predictive sampling for efficient pairwise subjective image quality assessment," in *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 2023, pp. 6676–6684. [Online]. Available: https://doi.org/10.1145/3581783.3612443

[59] Y. Li, S. Wang, X. Zhang, S. Wang, S. Ma, and Y. Wang, "Quality assessment of end-to-end learned image compression: The benchmark and objective measure," in *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, 2021, pp. 4297–4305. [Online]. Available: https://doi.org/10.1145/3474085.3475569

[60] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures," in *Proceedings of the 2011 SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 143–146. [Online]. Available: https://doi.org/10.1145/1978942.1978963

[61] R. F. Woolson, "Wilcoxon signed-rank test," in *Wiley Encyclopedia of Clinical Trials*. John Wiley & Sons, Inc, 2007. [Online]. Available: https://doi.org/10.1002/9780471462422.eoct979

[62] D. Gamage, D. Sewwandi, M. Zhang, and A. K. Bandara, "Labeling synthetic content: User perceptions of label designs for ai-generated content on social media," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3706598.3713171

[63] S. Höltervennhoff, J. Ricker, M. M. Raphael, C. Schwedes, R. Weil, A. Fischer, T. Holz, L. Schönherr, and S. Fahl, ""that's another doom i haven't thought about": A user study on ai labels as a safeguard against image-based misinformation," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2026. [Online]. Available: https://doi.org/10.1145/3772318.3791006

[64] C. Geeng, S. Yee, and F. Roesner, "Fake news on Facebook and Twitter: Investigating how people (don't) investigate," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 655:1–655:14. [Online]. Available: https://doi.org/10.1145/3313831.3376784

[65] J. Zhou, S. Z. Arshad, S. Luo, and F. Chen, "Effects of uncertainty and cognitive load on user trust in predictive decision making," in *Human-Computer Interaction – INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25-29, 2017, Proceedings, Part IV*. Springer, 2017, pp. 23–39. [Online]. Available: https://doi.org/10.1007/978-3-319-68059-0_2

[66] S. Z. Arshad, J. Zhou, C. Bridon, F. Chen, and Y. Wang, "Investigating user confidence for uncertainty presentation in predictive decision making," in *Proceedings of the 2015 Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. ACM, 2015, pp. 352–360. [Online]. Available: https://doi.org/10.1145/2838739.2838753

[67] Z. Yu and C. S. Chan, "Yuan: Yielding unblemished aesthetics through a unified network for visual imperfections removal in generated images," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, pp. 9716–9724, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i9.33053

[68] Y. Yuan, X. Wang, Y. Sheng, P. Chennuri, X. Zhang, and S. Chan, "Generative photography: Scene-consistent camera control for realistic text-to-image synthesis," in *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2025, pp. 7920–7930.

[69] L. Poretski and A. Tang, "Press A to jump: Design strategies for video game learnability," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, 2022, pp. 155:1–155:26. [Online]. Available: https://doi.org/10.1145/3491102.3517685

[70] M. Schemmer, P. Hemmer, M. Nitsche, N. Kühl, and M. Vössing, "A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2022, pp. 617–626. [Online]. Available: https://doi.org/10.1145/3514094.3534128

Appendix A. Detailed Review of Prior Related Studies

Table 4 compares our work with nine closely related empirical studies, which include a photorealistic face-image condition, report quantitative human detection accuracy, and use controlled experimental settings. The table shows that prior work mainly evaluated older GAN-based models or did not report model-specific human performance, and that existing training interventions provide limited evidence on model-specific training or longer-term retention.

Appendix B. Data Quality and Incentives

Data quality control. To ensure data quality, we applied the following screening criteria:

TABLE 4: Overview of prior work on human performance in detecting deepfake images, categorized by training strategy.

Training strategy	Study	Stimulus type(s)	Human performance	Deepfake model(s) (year)
Deepfake knowledge	Bray et al. [27]	Image (face)	Not significant (avg. 59.7% to 64.3%)	StyleGAN2 (2020)
	Nightingale et al. [26]	Image (face)	Improved (avg. 48.2% to 59.0%)	StyleGAN2 (2020)
Feedback training & AI support	Boyd et al. [23]	Image (face)	Improved fake-image accuracy (avg. 51.9% to 65.2%); no model-specific accuracy reported	ProGAN (2019), StyleGAN (2019), StyleGAN2 (2020), StyleGAN2-ADA (2020), StyleGAN3 (2021), StarGANv2 (2020)
	Hulzebosch et al. [24]	Image (face)	Improved (avg. 67.6% to 72.9%)	StarGAN (2018), GLOW (2018), ProGAN (2019), StyleGAN (2019)
	Diel et al. [25]	Image (face)	Improved (avg. 44.2% to 64.7%)	StyleGAN2 (2020)
No training	Shen et al. [22]	Image (face)	Avg. 49.1%	StyleGAN2 (2020)
	Frank et al. [21]	Image (face) and other modalities	Avg. 45.9%–49.9%	StyleGAN2 (2020)
	Lu et al. [10]	Image (face, landscape, plant, animal, object etc.)	Avg. 64.3%–66.4% (faces)	Midjourney V5 (2023)
	Cooke et al. [12]	Image (face, animal, food, landscape) and other modalities	Avg. 46.6% (faces)	StyleGAN (2019), StyleGAN2 (2020), StyleGAN3 (2021), Stable Diffusion v2.1 (2022), DALL-E 2 (2022)

TABLE 5: Demographic breakdown of valid main-study participants ($n = 524$). Age is also reported as mean (SD).

Characteristic		Count (n)	Percentage (%)
Gender	Female	267	51.0
	Male	257	49.0
Ethnicity	White	198	37.8
	Black	163	31.1
	Asian	163	31.1
Age	18–24 years	197	37.6
	25–54 years	202	38.5
	55+ years	125	23.9
	Mean (SD)	*37.41 (16.05)*	

- **Attention checks**: The main study included four attention-check trials, one in each pre-/post-training $\times$ single-image/side-by-side testing session. The five-month follow-up study included three attention-check trials across its testing sessions. These trials used easy-to-recognize images, and we excluded participants who failed at least two attention checks in either study.
- **Time filtering**: We recorded the completion time for the full session. Submissions completed in $<$6 minutes were treated as invalid due to insufficient effort.
- **Data integrity**: We removed submissions with missing responses, inconsistent self-reported demographic information, or an unfinished procedure.

After applying these quality-control measures, 524 participants remained for the main-study analysis and 342 for the five-month follow-up analysis.

Participant incentivization. Participants received £5.00 for completing the main study and £3.00 for completing the five-month follow-up study, consistent with Prolific payment guidelines. We also awarded performance-based bonuses within each training group and model condition. In the main study, participants in the top 10% received an additional £10.00, and those ranked between 10% and 30% received an additional £5.00. In the follow-up study, participants in the top 10% received an additional £6.00, and those ranked between 10% and 30% received an additional £3.00.

Participant demographics. Table 5 summarizes the demographic characteristics of the 524 valid participants.

Appendix C. Platform and Questionnaires

This appendix illustrates the experimental platform and lists the questionnaire items used in the main study and the follow-up study.

C.1. Platform User Interface

Figure 6 shows representative interfaces from the experiment platform, including the single-image test, side-by-side test, and practice-task feedback interfaces. In the practice-task feedback interface, participants received the correct answer and an explanation after submitting their response.

C.2. Confidence and Perceived-Stress Questions

After each testing session, participants answered two self-assessment questions about confidence and perceived stress. These questions supported the analyses of subjective states in Section 5.3 and Section 5.4. Both used a 5-point Likert scale: *Very low; Low; Moderate; High; Very high.*

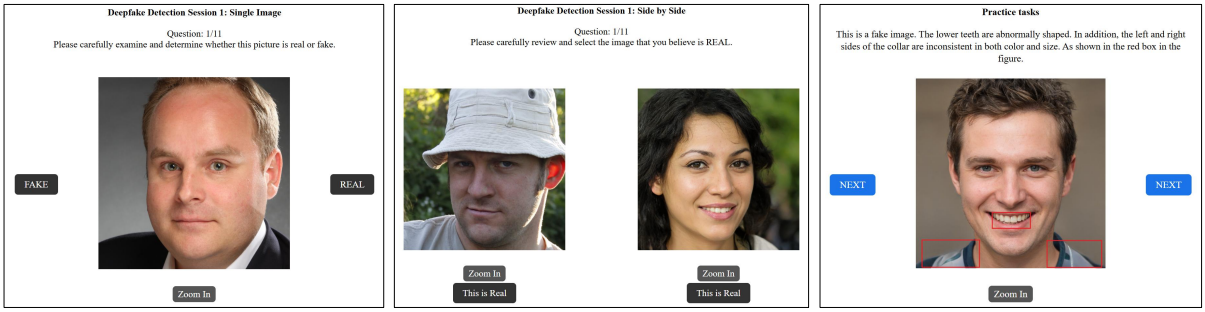


Figure 6: Screenshots of the experimental platform testing interfaces.

Q1. *Overall, how confident did you feel about your answers in the session you just completed?*
Q2. *Overall, how stressed did you feel during the session you just completed?*

C.3. Main-Study Feedback Questionnaire

The following feedback questions were shown after the post-training tests in the main study.
Q1. *Do you agree that deepfake technology is a significant societal problem?*
Options: Strongly disagree; Disagree; Neutral; Agree; Strongly agree.
Q2. *Do you agree that the training session was helpful for improving your ability to identify deepfake images?*
Options: Strongly disagree; Disagree; Neutral; Agree; Strongly agree.
Q3. *What aspects of the training do you find particularly useful?*
Options (select all that apply): Training video; Practice tasks; Others (open-ended text).
Q4. *What are the main clues you now look for when identifying deepfake images?*
Options (select all that apply): Background problems; Unnatural hair; Unnatural facial skin; Accessory rendering anomalies; Overly perfect and idealized style; Others (open-ended text).
Q5. *Do you have any other feedback or comments?*
Response format: Open-ended text.

C.4. Follow-Up Feedback Questionnaire

The following feedback questions were shown after the five-month follow-up tests.
Q1. *What are the main clues you look for when identifying deepfake images?*
Options (select all that apply): Background problems; Unnatural hair; Unnatural facial skin; Accessory rendering anomalies; Overly perfect and idealized style; Others (open-ended text).
Q2. *Do you still remember the training content from the previous study?*
Options: Not at all; A little; Some of it; Most of it; Very well.
Q3. *If you remember the training at all, what aspects of it do you still remember?*
Response format: Open-ended text.
Q4. *Do you have any other feedback or comments?*
Response format: Open-ended text.

Appendix D. Examples of StyleGAN3 Texture Artifacts



Figure 7: Examples of skin texture artifacts in StyleGAN3-generated images.

Some model-specific artifacts require close inspection to identify. In StyleGAN3-generated images, we observed recurring abnormal skin-texture patterns. These patterns are subtle at normal viewing size but become visible under magnification. Figure 7 shows two representative examples. In the zoomed-in crops, the skin surface lacks natural pore structure and instead exhibits structured local patterns with oval or diamond-like shapes. Additional annotated examples are available in the artifact repository.

Appendix E. Artifact Availability

The DFAT corpus and annotations, generation prompts, study materials, de-identified data, analysis scripts, and experimental website source code are available at https://doi.org/10.5281/zenodo.22843044.

Appendix F.
Ethics Considerations

The experiments that recruited human participants from Prolific received ethics approval from the Departmental Ethics Review Group at Cambridge Judge Business School, University of Cambridge, UK (reference number: 25-46). Participants provided informed consent before taking part in the experiments, where they were asked to identify real and deepfake face images. They were compensated for their time at a rate following Prolific recommendations and some of them received additional performance-based bonuses. We screened all face images to exclude inappropriate or disturbing content and minimize potential discomfort during the online experiment. Real images were sourced from the FFHQ dataset. We collected Prolific IDs for payment and follow-up recruitment and removed these identifiers from the publicly released study data. This work did not identify vulnerabilities or involve testing live third-party systems. The DFAT and associated training materials are intended to support education, evaluation, and human-centered security research.

To examine the prevalence of artifact types across deepfake models, five annotators who were not co-authors were recruited to label the deepfake images generated by different models. This task was led by the first two authors, whose institution (Xinjiang University, China) did not require a research ethics review process for such low-risk studies involving human participants. No personal or sensitive data was collected for this task, nor were the participants asked to do heavily repetitive work. All annotators did the work voluntarily and no financial compensation was involved.

Appendix G.
Meta-Review

The following meta-review was prepared by the program committee for the 2027 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

G.1. Summary

This paper presents a multi-month spanning study on human performance in detecting deep-fake images and the efficacy of different types of training in that period. As part of this study, the paper systematically studied the artifacts present in deep-fake images and created a taxonomy of those artifacts. Based on the artifacts, the study then conducted an online user-study ($N = 524$) where the study focused on different factors that may influence the training efficacy, such as generalized and specialized training, pre-and-post training tasks, and a follow-up study. Through the study, the paper demonstrated that the taxonomy guided approach for training the users on detecting deep fake images is useful and lasting even after months.

G.2. Scientific Contributions

- Creates a New Tool to Enable Future Science
- Provides a New Data Set For Public Use
- Provides a Valuable Step Forward in an Established Field

G.3. Reasons for Acceptance

1) The paper creates a taxonomy that reflects the large enumeration of artifacts and offers an effective approach for updating training materials, thus enabling future science.
2) The reviewers found the paper well organized, experimentally detailed, and rigorous with human-subject studies, thus providing a valuable step towards making progress in understanding human performance in detecting deep-fake images.
3) The open-sourced artifact is well organized and documented to make it suitable for public use and to support future research.