Characterizing Photorealism and Artifacts in Diffusion Model-Generated Images

Negar Kamali
Computer Science
Northwestern University
Evanston, Illinois, USA
negar.kamali@u.northwestern.edu

Karyn Nakamura
Kellogg School of Management
Northwestern University
Evanston, Illinois, USA
karynnakamura68@gmail.com

Aakriti Kumar
Kellogg School of Management
Northwestern University
Evanston, Illinois, USA
aakriti.kumar@kellogg.northwestern.edu

Angelos Chatzimparmpas
Department of Information and Computing Sciences
Utrecht University
Utrecht, Netherlands
a.chatzimparmpas@uu.nl

Jessica Hullman
Computer Science
Northwestern University
Evanston, Illinois, USA
jhullman@northwestern.edu

Matthew Groh
Kellogg School of Management
Northwestern
Evanston, Illinois, USA
matthew.groh@northwestern.edu

Abstract

Diffusion model-generated images can appear indistinguishable from authentic photographs, but these images often contain artifacts and implausibilities that reveal their AI-generated provenance. Given the challenge to public trust in media posed by photorealistic AI-generated images, we conducted a large-scale experiment measuring human detection accuracy on 450 diffusion-model generated images and 149 real images. Based on collecting 749,828 observations and 34,675 comments from 50,444 participants, we find that scene complexity of an image, artifact types within an image, display time of an image, and human curation of AI-generated images all play significant roles in how accurately people distinguish real from AI-generated images. Additionally, we propose a taxonomy characterizing artifacts often appearing in images generated by diffusion models. Our empirical observations and taxonomy offer nuanced insights into the capabilities and limitations of diffusion models to generate photorealistic images in 2024.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Human computer interaction (HCI)**.

Keywords

photorealism, diffusion models, generative AI, synthetic media, deepfakes, misinformation

ACM Reference Format:
Negar Kamali, Karyn Nakamura, Aakriti Kumar, Angelos Chatzimparmpas, Jessica Hullman, and Matthew Groh. 2025. Characterizing Photorealism and Artifacts in Diffusion Model-Generated Images. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3706598.3713962

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713962

1 Introduction

The capabilities of diffusion models to generate photorealistic images of people are beginning to contribute to disinformation and erode trust in the media [17]. For example, in March 2023, realistic AI-generated images of world leaders went viral on social media, showing Pope Francis wearing what appeared to be a designer puffer jacket, Donald Trump getting arrested, and Vladimir Putin standing behind prison bars [2]. These exemplar images may appear both provocative and realistic at first glance, but they are far from perfectly photorealistic; they contain distortions of hands and faces, implausible grasping of objects, and shadows that do not match the objects that appear to cast them. These distortions are not unique to these particular images but are pervasive in diffusion model-generated images produced by text–to–image tools such as Midjourney (the source of these fake images of world leaders), Stable Diffusion by Stability AI, and Firefly by Adobe [40]. While it is possible to generate images that seem indistinguishable from photographs, many diffusion model-generated images still leave behind human-identifiable artifacts. This raises an open research question for human–computer interaction: What drives human perception of photorealism in images generated by diffusion models?

We approach this question by conducting a large–scale, online experiment where we collect data on human participants' accuracy in identifying whether images are AI-generated or real. We measure photorealism following a psychophysics approach [95] that defines photorealism as human discrimination performance. Accuracy scores are inversely associated with photorealism: a high accuracy score indicates low photorealism, whereas a low accuracy score indicates high photorealism. By defining photorealism based on discrimination performance, we avoid the speculation and subjectivity of asking participants questions like "Is the image photorealistic?" [47] and "Could these images be taken with a camera?" [86].

By comparing human detection accuracy across a diverse set of images, we can evaluate the contexts that influence the continuum of photorealism. Past research has demonstrated that GAN-generated human portraits can be indistinguishable from real portrait images [60]. However, open questions on context remain: How often do portrait images appear indistinguishable from real portraits? How does scene complexity across styles of photographic portraiture (e.g., single-subject close-up, single-subject full body, posed group, and candid group) influence aggregate measures of photorealism? How accurately can people identify real images across scene complexities? What kind of artifacts arise in diffusion model-generated images and how are the presence of those artifacts related to photorealism? How does display time of an image influence measures of photorealism? How does human curation of AI-generated image stimuli influence measures of photorealism?

We approach each of these questions in turn and then consider an interventional question: How should an AI literacy guide categorize artifacts and implausibilities that emerge in photorealistic AI-generated images to promote attention to and communicate these visual cues? We also consider a flipped version of that question: How do we help people avoid falsely identifying a real image as AI-generated? This question differs from the first because it focuses on how people can identify when to trust what they see. This is important because there has already been a case where a politician has incorrectly, publicly claimed that an authentic photograph of his opponent was AI-generated [3, 63]. Enhancing human skill at distinguish real from AI-generated remains important because technical platform–level solutions (e.g. watermarking and machine learning classification) lack robustness and are susceptible to error when images are slightly modified via cropping, compression, and other edits. AI literacy guides have the potential to help humans stay abreast of the capabilities and limitations of AI to better navigate assessing the authenticity of images.

Our contributions toward answering these research questions are fourfold: First, we contribute a taxonomy of artifacts and implausibilities in diffusion model-generated images of humans along five dimensions: (1) anatomical implausibilities: representations of human anatomy that deviate from realistic or common forms; (2) stylistic artifacts: visual elements that often appear in images generated by diffusion models like shiny or plastic textures; (3) functional implausibilities: design or structural flaws that would make an object or system unlikely to function as intended in the real world; (4) violations of physics: instances where an object or scenario defies the laws of physics; and (5) sociocultural implausibilities: representations of people that are unlikely to reflect the norms of a culture or society. Second, we conduct a large–scale digital experiment and present an empirical evaluation of photorealism – as measured by human detection accuracy – across images from three state-of-the-art diffusion models, varying photographic contexts, types of artifacts in an image, and randomized display time of an image. This evaluation offers insight into the capabilities and limitations of diffusion models to produce photorealistic images. Third, we provide empirical evidence revealing the influence of human curation on the level of photorealism in images generated by diffusion models. This finding reveals the importance of human curation and the limits of state-of-the-art diffusion models to producing photorealistic images, and create consistent and believable narratives via an automated deluge of fake images that are indistinguishable from real photographs. Fourth, we release a public dataset of 749,828 responses from 50,444 participants on 599 images to enable the replications of our results and further research on photorealism in diffusion models; replication data, code, and stimuli can be found at the following link: https://github.com/negarkamali/Replication-for-Characterizing-Photorealism-2025/.

2 Background

2.1 Limitations of machine learning approaches to detect AI-generated images

Machine learning models for detecting AI-generated images are brittle and lack robustness to simple data transformations. Corvi et al. [10] compare four different machine learning approaches to deepfake detection and demonstrate that recropping and compression – simple modifications common on social media – lead to drops in accuracy such that the classifiers are nearly just as good as random guessing. Dong et al. [13]reveal the ease with which spectral artifacts used in the identification of GAN-generated images can be mitigated via blurring and resizing, demonstrating a noticeable decrease in accuracy under basic modifications. Cozzolino et al. [11] demonstrate that post–processing images by random–cropping, resizing, and compression lead to a drop in AI-generated image detection from 90% accuracy to 85% accuracy. The fundamental problem is that machine learning models for deepfake detection lack robustness to context shift, out–of–distribution data and adversarial perturbations [23, 28, 35, 80].

How an image is generated influences the ability of deepfake detection classifiers to accurately identify it as AI-generated. Classifiers trained to detect GAN-generated images tend to fail to detect diffusion model-generated images. For example, the approach to detecting GAN-generated images based on frequency spectra [6, 14, 21, 54, 91, 94] and inconsistencies in head poses and facial landmark positions [58, 88, 89], do not generalize to detecting images generated by diffusion models [62]. GAN-trained detection models miss these patterns because they have learned patterns for identifying GAN-generated images [70, 79]. Likewise, it is possible to learn the statistical regularities in diffusion model-generated images but these regularities are not invariant to image post-processing. [4, 52, 82, 85, 87].

Moreover, machine learning models' lack of robustness for detection is exacerbated by the changing architectures of generative AI models [48, 57]. Vision transformers [62, 68] and multi–architecture training [15, 39, 67] show promise for enhancing the detection of AI-generated images, but adversarial attacks and large architectural changes in generative models continue to affect robustness of detection.

Figure 1 highlights the increasing complexity of AI-generated images over the past decade. The changing architectures and increasing photorealism pose a challenge for both humans and machines to distinguish real from AI-generated images. However, humans and machines are fundamentally different. For example, humans can critically reason about an image's elements and its context [81]. On the other hand, machine learning classifiers for detecting AI-generated images often oversimplify image authenticity as a question of real versus fake and ignore the critical reasoning about component parts and sub–questions that an ordinary person or digital forensics expert may consider when evaluating an image's authenticity [37].

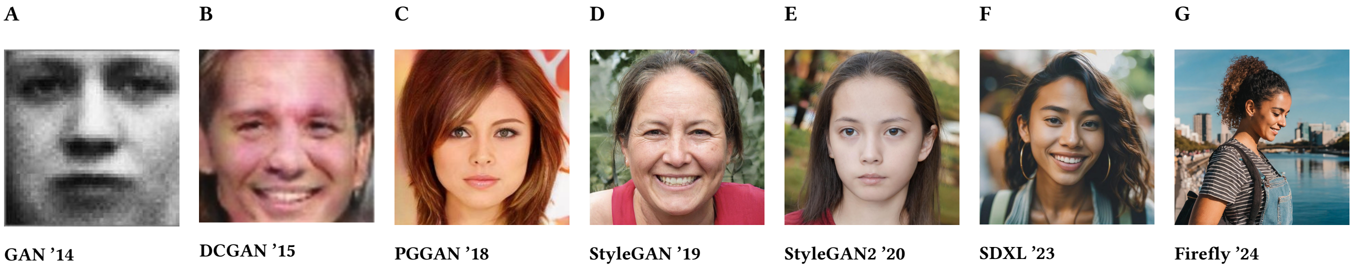


Figure 1: Exemplar images of photorealism across a range of generative models. Examples of AI-generated images from 2014 to 2024 [1, 22, 41–43, 66, 69, 92].

2.2 Human perception and evaluation of AI-generated media

In response to the increasing realism of AI-generated media, researchers have been examining the degree to which humans can distinguish between authentic and AI-generated media. For example, researchers found that GAN-generated images of faces are indistinguishable from real face portraits [44, 60]. However, for video deepfakes, humans are much better than random guessing [24], which may in part be due to humans’ specialized ability to process the temporal elements of faces [24, 73]. Researchers found that text–to–speech voices were rated as lower in quality and clarity than human voices in 2020 [9] but have reached the point where research participants cannot tell the difference between short 20-second recordings of AI-generated voices and authentically recorded voices [5].

Recent research has identified specific cues and heuristics that people use to evaluate AI-generated media. For example, cues such as recording settings in the detection of text-to-speech audio [30] and speaking patterns in political deepfake videos [25]. However, two studies found that participants rarely attributed their judgments to specific visual features [29, 84], and in one of these deepfake studies, researchers found that participants are noticing the artifacts but rarely linking these to manipulation [84]. With respect to AI-generated text, research has highlighted that people tend to use flawed heuristics when attempting to distinguish AI-generated text from human–written text, like associating grammatical errors with AI-generation [38].

Social context also plays a significant role in both what diffusion models generate [50] and how people form beliefs about AI-generated images and their content. For example, researchers have found detection ability is influenced by shared identity between the viewer and subject of the content [56]. Furthermore, researchers have found that white AI-generated faces were disproportionately judged as human more frequently than their real counterparts [55]. GAN-generated faces in portrait images were often perceived as more trustworthy than real faces [60], and as a result, people were less likely to question their authenticity [44]. In instances where AI-generated images are linked to misinformation, researchers find that labeling AI-generated images and the associated content as “potentially misleading" instead of simply “AI-generated" had a stronger influence on curtailing participants’ self–reported intentions to share misinformation [16, 83].

Researchers have approached a number of methods for measuring photorealism perceived by humans. For example, prior research has examined photorealism with carefully worded questions such as “Is the image photorealistic?” [47], “Does the image look like a real photo or an AI-generated image?” [46, 64] and “Whether this image could be taken with a camera?”[86]. These questions are useful for assessing participants’ subjective opinions but do not capture the human ability to distinguish real images from fake images and can potentially suffer from demand characteristic bias. Another approach has been to characterize photorealism by examining the features that can influence realism, such as aesthetics and semantically meaningful content of an image [65]. A third approach involves simply defining images as photorealistic if they are rendered with computer graphics software [51]. In this paper, we approach photorealism from the psychophysics perspective, examining participants’ objective performance at distinguishing real images from fake images [95].

2.3 Categorizing artifacts and implausibilities in diffusion model-generated images

Previous research on earlier versions of diffusion models categorized the kinds of qualitative failures of diffusion model-generated images as distorted body parts, impossible geometry, physics violations, illogical relationships in a scene, and noise [7]. In addition to obvious issues with hands, feet, eyes, and teeth, research at the intersection of digital foresnics and AI-generated images shows details such as corneal reflections [34] and irregular pupil shapes [26] can also be artifacts. Likewise, violations of physics like implausible shadows, lighting, and perspective errors [19, 20, 72] often occur in diffusion model generated images that otherwise appear photorealistic.

3 Methods

We develop a detailed taxonomy of visual features, qualities, and artifacts that offer cues that an image is AI-generated or not following a three-step process based on the taxonomy development method proposed by Nickerson et al. [59]. We began by drafting an initial version of the taxonomy based on a review of visual features previously identified in AI literacy resources, academic literature, and online discussions (Section 3.1). We then employed two parallel processes to develop the taxonomy: iteratively generating and curating a dataset of 599 images to showcase the taxonomy artifacts (Section 3.2) and conducting an online, crowdsourced experiment using these curated images to assess human detection ability (Section 3.3). Third, we integrated participant feedback—both accuracy

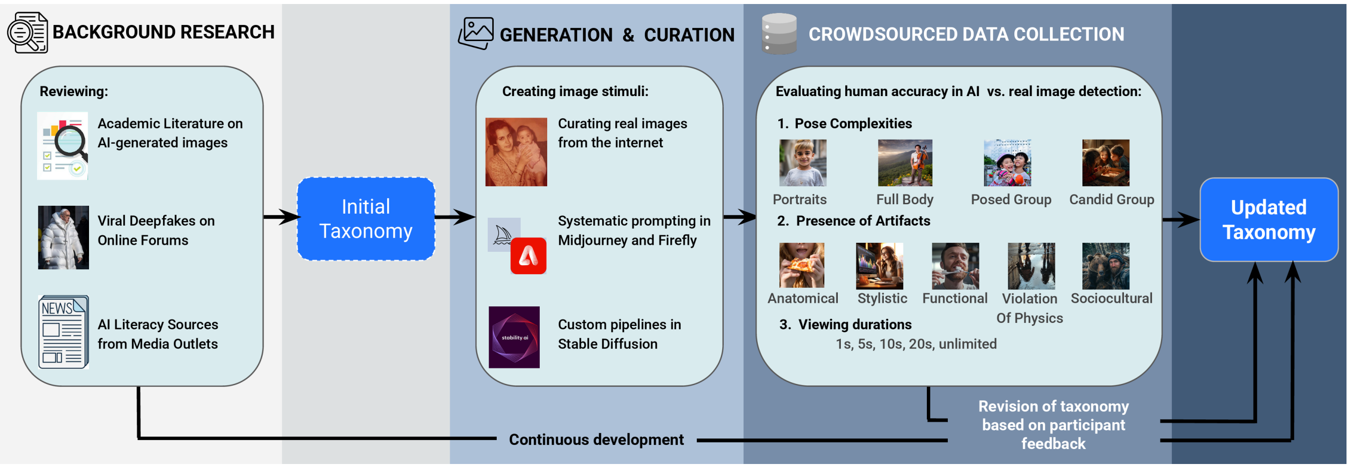

BACKGROUND RESEARCH
Reviewing:
Academic Literature on AI-generated images
Viral Deepfakes on Online Forums
AI Literacy Sources from Media Outlets
Initial Taxonomy
GENERATION & CURATION
Creating image stimuli:
Curating real images from the internet
Systematic prompting in Midjourney and Firefly
Custom pipelines in Stable Diffusion
CROWDSOURCED DATA COLLECTION
Evaluating human accuracy in AI vs. real image detection:
1. Pose Complexities
Portraits
Full Body
Posed Group
Candid Group
2. Presence of Artifacts
Anatomical
Stylistic
Functional
Violation Of Physics
Sociocultural
3. Viewing durations
1s, 5s, 10s, 20s, unlimited
Updated Taxonomy
Revision of taxonomy based on participant feedback
Continuous development

Figure 2: Overview of the taxonomy development process. In the background research stage, we reviewed existing literature on visible features of AI-generated images from a wide range of sources. This included academic literature, practitioner perspectives in AI literacy articles, and discussions on the photorealism of AI-generated images online. From these features, we developed an initial taxonomy of artifacts. In the Generation and Curation stage, we used our taxonomy of artifacts to create a dataset of 599 images. Of these images, 149 were real photographs curated from the internet, and 450 were generated in Midjourney, Firefly, and Stable Diffusion through extensive iteration with photorealistic image generation techniques. We used the dataset of images for an online crowdsourced experiment where we evaluated participant accuracy in identifying AI-generated images. We iteratively refined the taxonomy based on results from the experiment and continued monitoring new literature on AI-generated images as generative models evolved.

metrics and thematic comments—back into the taxonomy, allowing real-world human detection behaviors to inform the final categorization.

While the taxonomy development was guided by data, we acknowledge that subjectivity is inherent in the categorization process. To mitigate this subjectivity and ensure methodological rigor, multiple team members independently identified recurring patterns and artifacts during image generation and curation, and we reconciled any differences through structured discussion until stable, consistently observed phenomena emerged. In Figure 2, we show an overview of the taxonomy development process.

3.1 Initializing the Taxonomy

In addition to reviewing academic literature discussed in Section 2, we surveyed traditional and social media discussions about distinguishing AI-generated images. These included AI literacy resources on how to identify AI-generated content in media (see Figure S1), online discussions of AI-generated images in response to viral deepfakes, and popular posts discussing photorealism on online forums for AI image creators (Reddit channels such as r/Midjourney and r/StableDiffusion) to initialize the taxonomy. These sources highlighted several visual cues, including (1) anatomical implausibilities such as pupil dilation and misaligned eyes [61, 74, 76], teeth [12], hair [12, 61], fingers, and alignment of body parts [12, 61, 76]; (2) irregular reflections and shadows [12, 18, 74]; (3) unnatural color balances [12, 61]; (4) a mismatch in textures and styles within an image [12, 18, 61]; (5) garbled or nonsensical text [76]; (6) photoshoot-like perfection and overly cinematic scenarios [77].

While some prior research has suggested that AI-generated face images can be indistinguishable from real ones [35, 60], more complex scenes, such as group photos have not been thoroughly explored. We address this by introducing a detailed categorization of *scene complexity* across all images. We identified four distinct scene types that capture varying levels of detail within an image:

- **Portraits (Single-Subject Close-Up)**: An image featuring a single individual, typically focusing on the face and torso. The individual is the primary focus, often set against a blurred or minimal background. Portraits have relatively low scene complexity.
- **Full-Figure (Single-Subject Full Body)**: An image featuring a single individual whose entire body is visible along with the surrounding environment. These images exhibit moderate scene complexity, as they include more details than portraits, such as the person's posture and interaction with their setting.
- **Posed Group**: An image featuring multiple people posing for the camera in a structured manner. These images involve higher scene complexity due to the presence of multiple subjects, their interactions, and the added challenge of capturing each person accurately.
- **Candid Group**: An image of multiple people captured in candid moments. These images often feature intricate interactions between people and their environments, representing the highest level of scene complexity.

3.2 Stimuli Generation and Curation

We created a dataset of 599 images. This image set included 149 real photographs curated from the internet, from which we derived the scenarios for 450 images that we generated using AI. Of the

AI-generated images, 207 were generated in Midjourney, 133 in Firefly, and 110 in Stable Diffusion.

Drawing on techniques shared on online forums and articles, we developed strategies to generate photorealistic images. We first curated real photographs and then experimented extensively with the three AI-generation tools to create over 3000 images that depict similar scenarios as the real images. The final dataset represents a selection of images from this larger set that we judged to be not immediately identifiable as AI-generated at first glance. This selection enabled us to focus on more challenging cases, better assess participants' ability to detect subtle artifacts, and enhance the relevance of our taxonomy in real-world scenarios.

3.2.1 Curating Real Photographs. We sourced real photographs from online platforms, selecting them to represent diverse scenarios (e.g., diverse cultural settings with celebrity and non-celebrity figures engaging in common and uncommon activities) across the four dimensions of scene complexity Section 3.1. We established these categories to curate a diverse range of real photographs and ensure our dataset accurately captures how the features in our taxonomy may manifest and be perceived in both real and AI-generated images. We verified that these images were real photographs by confirming details like the creation date, photographer, and publisher. We include a complete list of image sources and verification details in the following link: https://github.com/negarkamali/Replication-for-Characterizing-Photorealism-2025/. We used these real images to inform the prompts to generate images using AI tools.

3.2.2 Generating Images using AI tools. Based on our curated set of real images, we generated images in Midjourney V5 and V6, Adobe Firefly Image 2, and Stable Diffusion to depict similar scenarios. In Midjourney and Firefly, we started the image generation process by creating a simple prompt describing the scenario. We then progressively refined the prompts by adding details about the quality of the image using keywords known to enhance image quality and resolution from the sources mentioned in Section 3.1. Our prompts followed the basic structure of: "[Subject description] [action or pose], [context or setting description], [clothing or appearance details], [image quality and style attributes], [camera or film type if applicable]." If the images were insufficiently realistic, then further details were added to the end of the above prompt such as: [specific details unique to the scenario], ['high resolution', 'hyper-realistic', 'megapixel', etc.]. The sequence of images in Figure 3 shows the progression of a prompt and the resulting image qualities as more details and keywords are added in Midjourney V6.

We also generated images inspired by real scenes of human interactions found in publicly available news sources and online media, ensuring they maintained a similar context and zoom level to real photographs. For example, we used a real reference image of a Ukrainian soldier getting married from New York Magazine [53].

In Stable Diffusion, we developed custom pipelines in order to generate images that were more realistic than the outputs of the original models. Using SD1.5 [71] and SDXL [66] as the base models, we used techniques such as merging fine-tuned portrait models and combining outputs of different models to reduce obvious artifacts and generate highly photorealistic images, particularly for portraits. We also experimented with generating the same poses in various styles in order to isolate the impact of certain categories of artifacts, as shown in Figure 4. We used ControlNets [93] to maintain consistent poses while altering other elements such as models, seed, and prompt scheduling. Additionally, we used Low-Rank Adaptation (LoRAs) [33] to introduce realistic imperfections like wrinkles and shorten the depth of field to produce iPhone-style images. We further refined images by implementing pipelines that regenerate artifacts in the hands and faces. A refining pipeline is shown in Appendix S2.

3.3 Crowdsourced Experiment

3.3.1 Image Stimuli. Across the entire experiment timeline, we collected 749,828 responses to whether 1083 images are AI-generated or real from 50,444 participants. Across the experiment timeline, we added and removed stimuli for two reasons. First, we included higher quality and more diverse images over the course of the experiment as new tools for controlling diffusion models became available (e.g., ControlNets and LORAs), and we identified prompt engineering techniques for producing more photorealistic images. Second, we split the experiment into two phases based on how we selected the diffusion model-generated images. In the first and main phase of the experiment, the stimuli were 149 real photographs and 450 most photorealistic images that our research team could generate with diffusion models. By comparison, the 482 stimuli in the second phase were based on generating 11 or more images for each of the 39 text prompts without curation. This second phase enables us to identify the effect of human curation (the selection bias involved in our research team selecting the most photorealistic AI-generated images) relative to no human curation on how accurately participants can distinguish AI-generated images.

3.3.2 Experimental Design. To ensure the quality of results, we implemented two measures. First, participants were shown an attention check image that was clearly AI-generated. Those who failed to identify this image correctly were excluded from the analysis. Second, we included an optional checkbox allowing participants to indicate if they recognized an image from outside the experiment, allowing us to filter out responses influenced by prior familiarity with the image.

In the initial version of the experiment from February to May 2024, we prioritized presenting unseen images to participants rather than maintaining a balanced ratio of real and AI images. In the next version of the experiment, from May to June 2024, we used stratified random sampling to select stimuli, ensuring participants saw real images 50% of the time and AI-generated images 50% of the time. We repeated the analysis both with and without the data from the initial version of the experiment and did not find significant changes in the accuracy distributions. Details of this comparison are provided in Appendix S1 and S4. To ensure that newly added images to the stimuli set as described in Section 3.3.1 were adequately represented in participant responses, we implemented an up-sampling strategy that prioritized showing images that were labeled fewer than 100 times.

After participants responded to five images, we randomized the display time of each subsequent image to one of the following conditions: unlimited time, 20 seconds, 10 seconds, 5 seconds, and 1 second. Participants were informed of the time limit at the start of each time-restricted trial by an on-screen message (e.g., "You

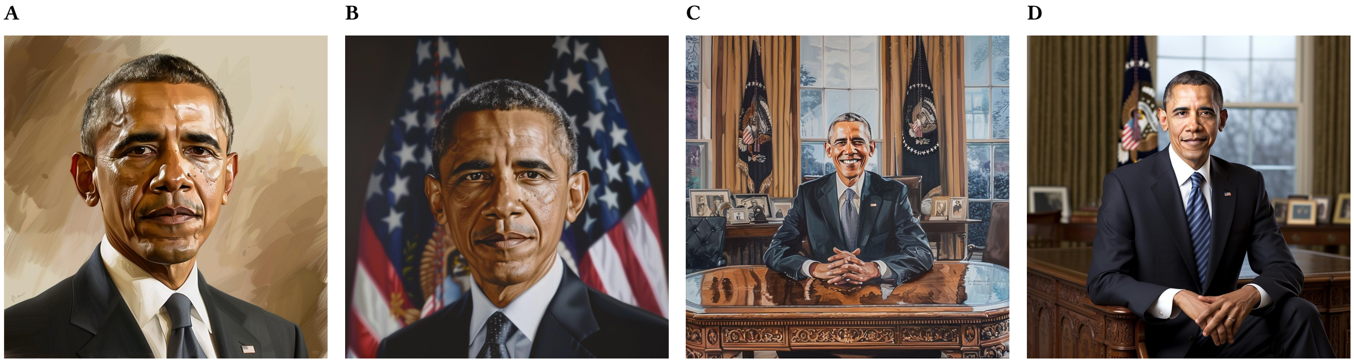


Figure 3: Images of Barack Obama generated in Midjourney V5. Images were created by progressively adding details to the prompt shown below each image: **A.** "Portrait of Barack Obama." **B.** "Portrait of Barack Obama, hyperrealistic, megapixel." **C.** "Portrait of Barack Obama, sitting in his Oval Office, smiling, hyperrealistic, megapixel." **D.** "A portrait of Barack Obama sitting in the Oval Office, smiling, wearing a suit and tie, shot on Kodak, hyperrealistic, grainy, official portrait."

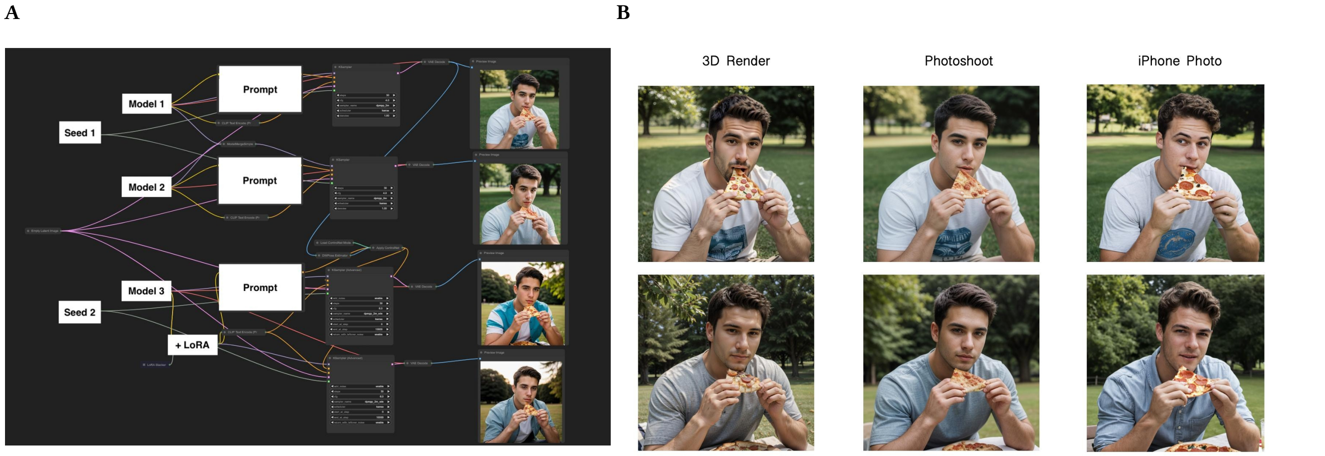


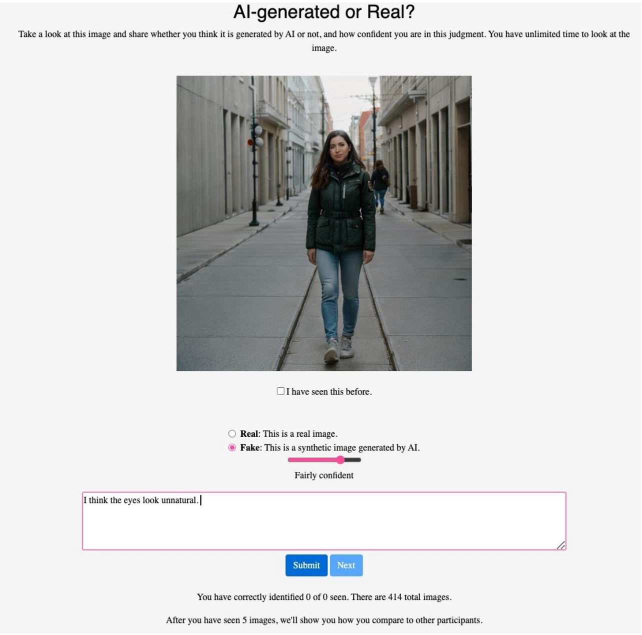
Figure 4: Stable Diffusion pipeline and outputs of varied styles from the same pose and prompt. A. Four pipelines for generating four variations of the prompt "photo of a 25 year old man eating a slice of pizza, outside on the grass in a park, sunny, plain clothes." **B.** A sample of the variations that we labeled as having the style of a "3D Render", "Photoshoot", or "iPhone photo."

have 20 seconds to view this image") and were instructed to click a button to reveal the image and begin the countdown.

3.3.3 Participants. We collected data through a public website (detectfakes.kellogg.northwestern.edu) where people could test their ability to detect AI-generated images. The website remained accessible throughout our taxonomy development, allowing us to gather responses as we updated image stimuli to reflect improvements in generation models. In total, 50,444 unique participants contributed 749,828 observations. According to Google Analytics, participants who visited our website came from 165 countries; the five countries with the most participants were United States, South Korea, United Kingdom, India, and Germany. We did not collect additional demographic data or other data on participants.

3.3.4 Image-level and Participant Level Analyses. We define accuracy as a binary measure of whether a participant selected the correct label (Real/AI-generated) for an image. We aggregated accuracy at two levels: image-level accuracy and participant-level accuracy. For image-level accuracy measurements described in Sections 5.1, 5.3, 5.5, 5.4 and 5.8, we aggregated and averaged the binary responses (0 for "real" and 1 for "AI-generated") provided by participants for each image. Image-level accuracy was calculated as the mean of correct identifications across various factors contributing to photorealism, which are described in each section. For participant-level accuracy measurements described in Section 5.2, we calculated each participant's accuracy by averaging their correct identifications across all viewed images.

We present descriptive statistics to summarize our findings, focusing on mean accuracies and their associated 95% confidence intervals (CIs) obtained through non-parametric bootstrapping [36]. We use these measures to describe trends and patterns in the data without using statistical significance to dichotomize effects. However, readers can apply that interpretation to the CIs if they desire.



AI-generated or Real?

Take a look at this image and share whether you think it is generated by AI or not, and how confident you are in this judgment. You have unlimited time to look at the image.

□ I have seen this before.

○ **Real**: This is a real image.

● **Fake**: This is a synthetic image generated by AI.

Fairly confident

I think the eyes look unnatural.

You have correctly identified 0 of 0 seen. There are 414 total images.

After you have seen 5 images, we'll show you how you compare to other participants.

Figure 5: A screenshot of the experiment website interface.

For a qualitative analysis of the 34,675 optional comments provided on our website, we utilized GPT-3.5-turbo to extract ten recurring themes and map the comments to our taxonomy categories based on the types of artifacts and patterns reported by participants.

3.3.5 Ethics. This research complied with all relevant ethical regulations. The Northwestern University Institutional Review Board (IRB) determined that it met the criteria for exemption from further review. The study’s IRB identification number is STU00220627.

4 Taxonomy of Artifacts in AI-generated Images

Our taxonomy organizes artifacts and implausibilities that may appear in AI-generated images into five high-level categories that are described in further detail in a how-to guide for identifying diffusion model-generated images[40].

1. Anatomical Implausibilities: This category refers to artifacts that appear in the depiction of people within an image. These include unlikely artifacts in individual body parts, like hands with extra or missing fingers as shown in Figure 6–1B, or the disproportionately long woman’s neck in Figure 6–1A. They also include artifacts in facial features such as an unnaturally empty gaze, overly shiny eyes, overlapping of the teeth and mouth, and unlikely proportions or configurations of limbs. In images of multiple people, this includes merged body parts and inconsistent proportions of body parts across different people. Anatomical implausibilities also include biometric artifacts such as size, shape, contours, and proportions of specific facial features if the person in the image is known. These biometric features include eyes, nose structure, mouth edges, interpupillary distance, ear shape and positioning, as well as distinctive markers like moles, dimples, and scars [45]

2. Stylistic Artifacts: This category refers to qualities of entire images or inconsistencies of those qualities within an image. This includes images of people that are waxy (Figure 6–2A), glossy (Figure 6– 2C), shiny, and appear perfect like a model doing a photoshoot. These characteristics often appear in plastic–like skin and excessively soft hair. Additionally, this category includes noticeably cinematic, picturesque, and dramatic images that often appear in artistic photographs like Figure 6–2B. Stylistic artifacts also include inconsistencies between different subjects or parts of an image. This may appear as smudge-like distortions at the edges of different components or differences in resolution that make these parts look like they are cut out from different scenes.

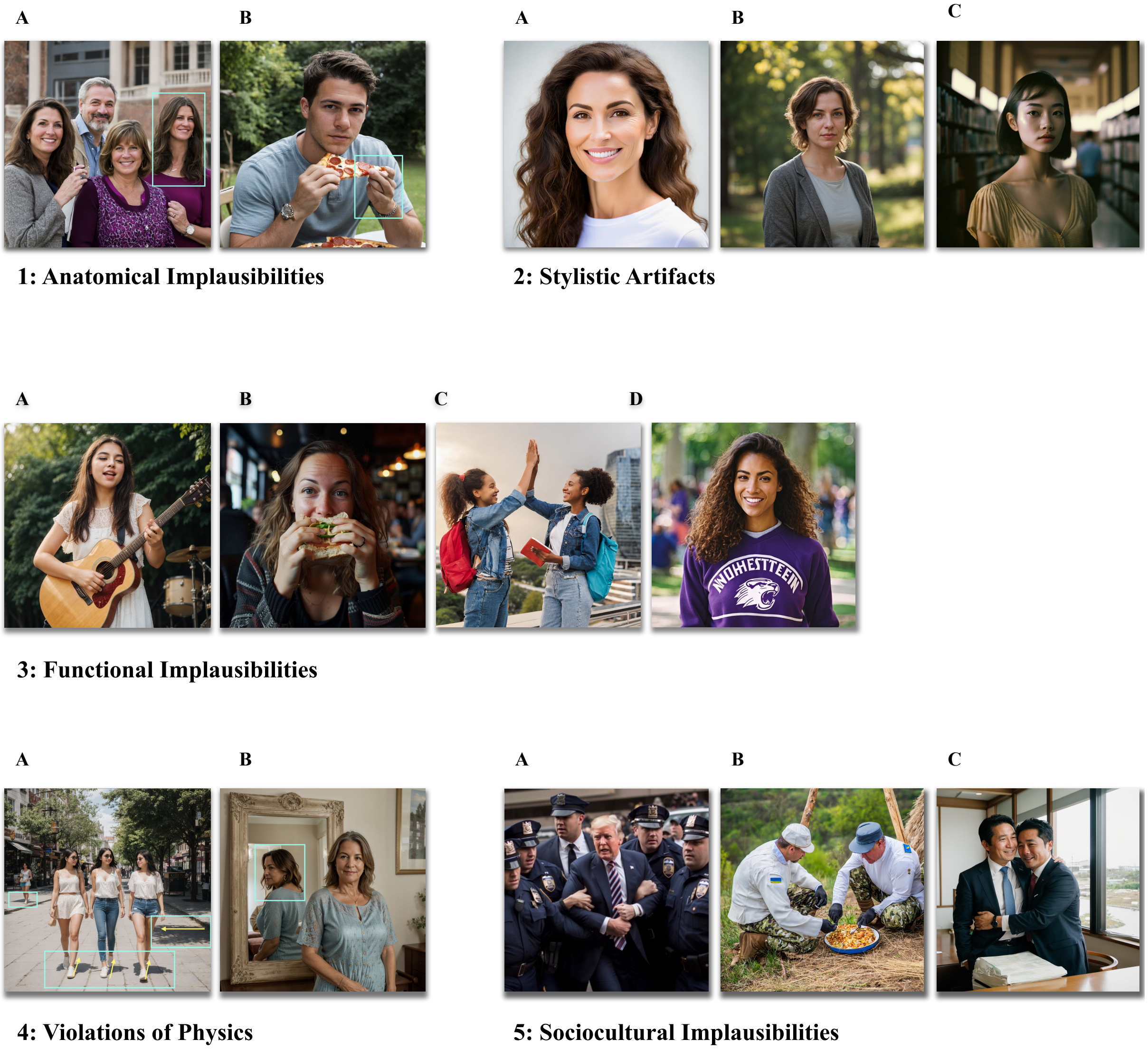
3. Functional Implausibilities: Functional implausibilities result from a lack of understanding of the fundamental logic of real–world mechanical principles. This includes implausibilities in the objects themselves, their placement within the environment, and how the people in the image may be holding or using these objects, such as the woman holding a sandwich sideways in Figure 6–3B. Objects may also appear unable to function, like the loose strings of the guitar in Figure 6–3A, or placed in a way that they cannot function. Functional implausibilities also include distortion in fine details of the image. The image may present atypical designs in details like the print, buttons, and buckles on pieces of clothing, as seen in a backpack strap merging into a denim jacket in Figure 6–3C. Functional implausibilities also include errors in text, such as distorted or unconventional glyphs and odd spelling errors as seen in Figure 6–3D.

4. Violations of Physics: This category addresses inconsistencies in the image content that violate the expected logic of physical reality. Examples include shadows pointing in diverging directions, as shown in Figure 6–4A, or shadows that do not correspond to their light sources. Additionally, reflections on surfaces like water, mirrors, or shiny objects may appear misaligned with their surroundings, as illustrated in Figure 6–4B. Violations of physics also include depth and perspective issues, like warping and trajectories that do not align with the rest of the image. These distortions can also occur in real photographs, for example as seen with fish-eye lens distortions.

5. Sociocultural Implausibilities: This category includes scenarios that are socially inappropriate and unlikely to be seen in the real world, such as people wearing bathing suits at a funeral and a selfie with a bear. Violations of social and cultural norms could also be more subtle, present in details specific to certain cultures like Figure 6– 5B and 5C attempting to depict Ukrainian and Japanese cultures, respectively. Historical inaccuracies and fake images of public figures in unlikely settings like 5A of Figure6 are also examples of sociocultural implausibilities.

5 Accuracy in Distinguishing AI-generated Images from Real Photographs

In the main phase of the experiment, we collected 539,749 responses on 599 images from 37,568 participants from February 5, 2024 to June 22, 2024. Sections 5.1 through 5.7 focus on data from the main phase of the experiment. The second phase of the experiment started on June 22 and ended on August 30, with 83,577 responses on 482 images from 3,787 participants. Sections 3.3.1 and 5.8 describe



1: Anatomical Implausibilities

2: Stylistic Artifacts

3: Functional Implausibilities

4: Violations of Physics

5: Sociocultural Implausibilities

Figure 6: Categories of Artifacts in AI-Generated Images. This figure presents representative examples of common artifacts found in AI-generated images across five categories: **1. Anatomical Implausibilities: A.** Stable Diffusion image of a group of people where one woman has an abnormally long neck. **B.** Stable Diffusion image of a man eating pizza where his left fingers appear anatomically implausible. **2. Stylistic Artifacts: A.** Firefly image of a woman with a waxy texture. **B.** Stable Diffusion image with a cinematized style. **C.** Midjourney image of a woman with glossy skin. **3. Functional Implausibilities: A.** Stable Diffusion image where the guitar strings are not taut. **B.** AI-generated image of a woman holding a sandwich in an unlikely way. **C.** Firefly image where the strap on the red backpack merges into the denim jacket. **D.** AI-generated image of a woman wearing a shirt with incomprehensible text. **4. Violations of Physics: A.** Stable Diffusion image where the shadows fall in inconsistent directions. **B.** Stable Diffusion image of a woman standing in front of a mirror in which her reflection is inconsistent with the direction of her face. **5. Sociocultural Implausibilities: A.** Midjourney image of Donald Trump being restrained [32]. **B.** Firefly image depicting Ukrainian servicemen dressed in white shirts and hats that are not commonly part of the uniform. The flags on their shirts are different, and on the right serviceman, the flag is positioned awkwardly on their back and not their arm. **C.** Stable Diffusion image of an unlikely scenario of two Japanese businessmen hugging in a professional setting.

the influence of human curation of the stimuli on how accurately participants identify the stimuli as AI-generated or real.

The design of our experiment involves several important design choices. First, we selected the three models of Midjourney, Firefly, and Stable Diffusion as the diffusion models. Second, we crafted prompts to produce realistic outputs across various pose categories and content types. Third, we curated 450 images from over 3000 images generated to use as image stimuli in the experiment. These images were selected to maximize realism while also representing different visual artifacts and implausibilities. Inevitably, these design choices on models, prompts, and stimuli introduce some selection bias into the experiment.

Additionally, we implemented two exclusion criteria that should be considered when interpreting our results. First, for all the analyses in Section 5, we excluded observations where participants checked the box on the website "I have seen this before". These observations, which account for 2% of the total observations, were excluded because of the strong possibility that participants who had previously seen the images were already aware of whether they were fake or real. For these observations marked as having been seen before, 38% of these observations were on AI-generated stimuli and 62% were on real images. The image most frequently reported as 'seen before' is a real portrait of Martin Luther King Jr, which was one of the few real images of a well-known celebrity included in the experiment.

Second, in line with our goals of studying detection ability on images for which there was some ambiguity, we excluded all images where participants' accuracy suggested very little ambiguity. We operationalized this as accuracy above 90%.

These exclusion criteria remove all observations on 68 fake images and 4 real images, which represent 14% of observations from the entire experiment.

In the human-coded analysis of artifacts discussed in Section 5.4, we apply an additional exclusion criterion to make the coding tractable. Specifically, we exclude all images accurately identified in more than 80% of observations. This exclusion criterion focuses the analysis on the most challenging images by excluding the most egregious distortions that lead to low photorealism (i.e., high participant accuracy).

5.1 Overall Accuracy

In the main study, participants correctly identified AI-generated images and authentic photographs in 76% and 74% of observations, respectively. Accuracy varied substantially across images. Prior to implementing our accuracy-based exclusion described above, we found that for AI-generated images, accuracy ranged from 32% to 99%. Similarly, accuracy on real photographs ranged from 28% to 92%. Figure 7 shows the distribution of accuracy in both AI-generated and real images with example images selected from the top, bottom, and middle deciles of each distribution. At the image level, the mean accuracy for identifying AI-generated and real images was 76% (95% CI:[74,77]) and 74% (95% CI:[72,76]), respectively.

Despite our efforts to minimize obvious artifacts, some images - particularly non-portraits - were challenging to generate without noticeable artifacts. As a result, participants achieved nearly 100% accuracy on a few AI-generated images with obvious features. We present examples of these images in Figure 8. In contrast to AI-generated images, real photographs rarely contain definitive artifacts and visual cues often seen in AI-generated images, which limits participants from achieving near-perfect accuracy on real photographs.

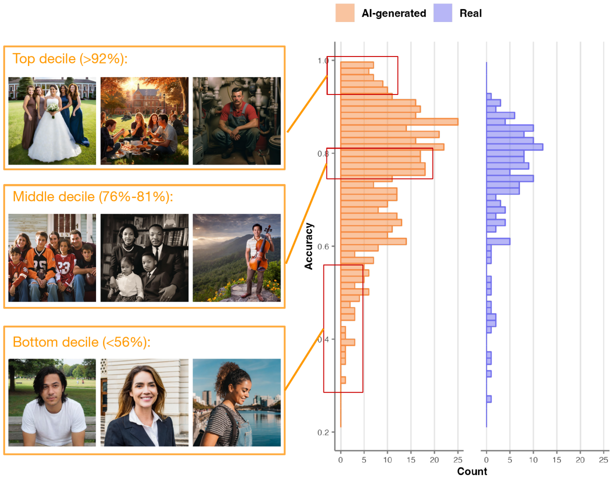


Figure 7: Distribution of accuracy scores for real and AI-generated images with example images representing different accuracy levels.

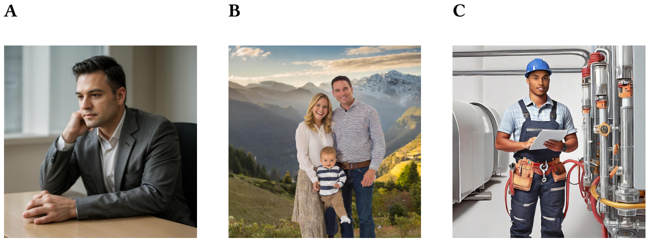


Figure 8: Examples of obviously AI-generated images and their corresponding accuracy. A. AI-generated portrait with 92% accuracy. **B.** AI-generated posed group image with 95% accuracy. **C.** AI-generated full-body image with 99% accuracy.

5.2 Participant Level Accuracy

Given the organic nature of participants' engagement with this experiment, we did not impose restrictions on the number of images a participant saw. Most participants in this study provided responses to at least seven images, but some participants only provided a single response, and one participant provided 502 responses.

The vast majority of participants (75%) saw 16 or fewer images. Figure 9A and B present the distribution of participant–level accuracy by number of viewed images.

In order to compare participant performance and avoid issues that arise with differential attrition, we focus on the first ten images seen by participants who saw at least 10 images, which includes 152,050 observations from 15,205 participants. First, we note that 34% of these participants achieved 90% accuracy or higher on the first ten images seen. If the AI-generated images were perfectly photorealistic such that the human ability to distinguish is no higher

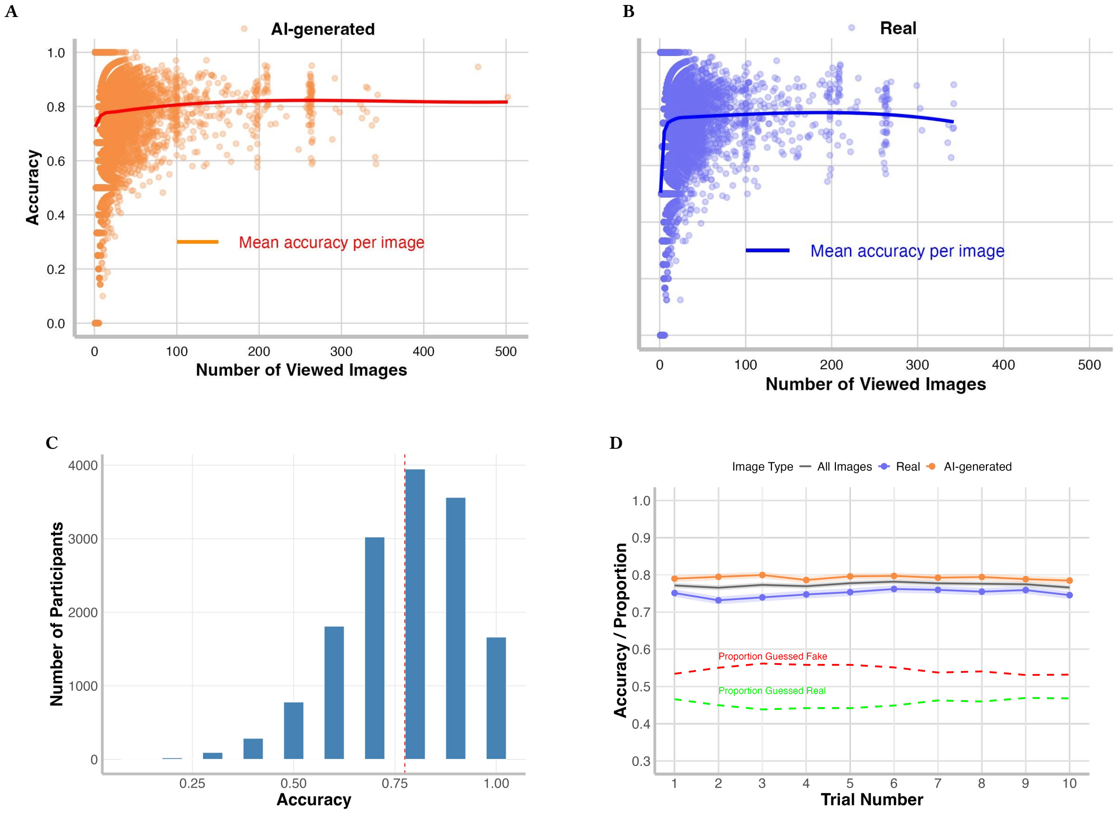


Figure 9: Participant-Level Accuracy and Learning Trends. **A.** Scatterplot of participant-level accuracy for AI-generated images. **B.** Scatterplot of participant-level accuracy for real images. **C.** Histogram showing the distribution of accuracy across the first ten images seen by participants who viewed at least 10 images. **D.** Learning curve illustrating accuracy trends and classification biases when detecting AI-generated and real images.

than random guessing, then we would have expected only 1% of participants to achieve this threshold of accuracy (assuming random guessing at 50% accuracy, with participants evaluating 10 images each, achieving at least 9 out of 10 correct responses would occur with a probability of approximately 1.07%, based on the binomial probability distribution). Figure 9C shows the distribution of accuracy across the first ten images seen by participants who saw at least 10 images.

In Figure 9D, we present accuracy rates by the number of images seen. We find that on average, participants begin the experiment by disproportionally identifying images as fake in 63% of observations. Notably, this bias is reduced after only a few images.

5.3 Accuracy by Scene Complexity

We find that on average, participants' accuracy increases as scene complexity increases. For example, we find that 16% of portraits appear in the bottom decile of accuracy scores (representing the highest level of photorealism), whereas only 3% of AI-generated posed group images appear in the bottom decile. Figure 10 presents the distribution of accuracy for each category, separately for real and AI-generated images. For AI-generated images, the mean accuracy was 72.7% (95% CI: [72.4, 72.9]) for portraits, 77.2% (95% CI: [76.8, 78.6]) for full body, 76.2% (95% CI: [75.8, 76.7]) for posed groups, and 73.4% (95% CI: [73, 73.8]) for candid groups. For real images, the accuracy was 71.1% (95% CI: [70, 71.4]) for portraits, 75.5% (95% CI: [75.1, 75.8]) for full body, 76.7% (95% CI: [76.3, 77]) for posed groups, and 74.8% (95% CI: [74.4, 75.1]) for candid groups.

As exemplified in Figure 10C, we note that portraits, relative to the other levels of scene complexity, typically have less detail, simpler and more standardized poses, more blurred backgrounds, and fewer available cues than full-body or group images.

5.4 Accuracy by Presence of Artifacts

In order to analyze accuracy by artifact type, we annotated images with diffusion model artifact categories from the taxonomy based on a three-step process. First, four co-authors independently annotated all 218 images with accuracy below 80%, identifying artifacts and providing detailed explanations for their annotations. Second, each of these annotations was reviewed and edited by two additional co-authors. Third, a fifth co-author reviewed all annotations for consistency. Figures 12A–C and 12D–F provide examples of how we annotated images, displaying the identified artifact categories, the reasoning behind their identification, and the associated detection accuracy for each image. During this process, we observed that the three main artifact types—anatomical implausibilities, stylistic artifacts, and functional artifacts—each appeared in nearly a third of the images we annotated. In contrast, violations of physics and sociocultural implausibilities were less common, appearing in only

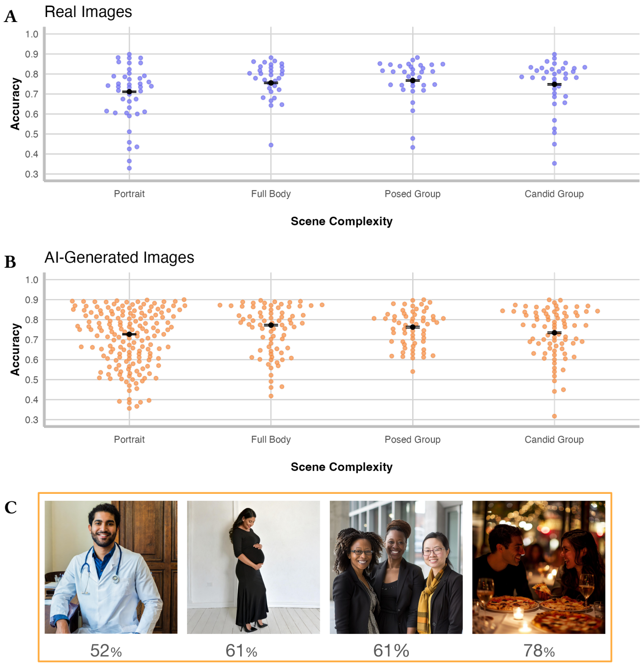


Figure 10: Scene complexity: Accuracy of real and AI-generated images by scene complexity levels. Beeswarm plots of image-level accuracy for each dimension of scene complexity with bootstrapped 95% confidence intervals. We exclude images identified with above 90% accuracy in this analysis. **A.** Real images **B.** AI-generated images **C.** AI-generated images across scene complexities.

20 and 12 images, respectively. In light of this distribution of artifacts, Figure 11A presents the distribution of accuracy scores across images containing at least the three listed artifact types.

Based on our annotations of artifacts in images, we find participants are less accurate on images with functional implausibilities than images with anatomical implausibilities or stylistic artifacts. The mean accuracy on images with at least one functional implausibility, one anatomical implausibility, and one stylistic artifact is 64.1% (95% CI: [63.8, 64.5]), 65% (95% CI: [64.6, 65.4]), and 64.9% (95% CI: [64.5, 65.3]), respectively. While the accuracy on images with functional implausibilities is lower than on images with other implausibilities and artifacts, the mean accuracy scores are similar. However, this similarity in means masks the differences in the distribution of accuracy scores, as shown in Figure 11A. We find that images with participant accuracy scores in the 40–60% range (which represent images approaching indistinguishability between real and AI-generated) make up 32.8% of images annotated with functional implausibilities compared to 21.4% and 22.4% of images annotated with anatomical implausibilities and stylistic artifacts, respectively.

We find that images that we annotated as containing multiple artifacts can still appear photorealistic enough to make detection difficult for most people. Artifacts vary in levels of visibility, as shown in Figure 12A–C. While Figure 12A and C contain stylistic artifacts, they are far more apparent in Figure 12B, which is reflected in its higher detection accuracy. Despite Figure 12A and C containing multiple artifact categories, they had low detection accuracy, suggesting that the presence of multiple artifacts does not necessarily make images easier to identify and that artifact visibility is also a contributing factor.

The visibility of artifacts is highly variable, and Figure 12D–F present examples highlighting this variability. The anatomical implausibility in the fingers in image Figure 12D is very noticeable, whereas the functional implausibilities in the tennis racket and shirt design of Figure 12F are more subtle. The corresponding accuracy scores for these images— 62% for Figure 12E and 54% for Figure 12F —reinforce the observation that anatomical artifacts tend to be more easily detected, while functional implausibilities often require closer attention and familiarity with depicted objects. The stylistic artifacts in the cinematization of Figure 12E and plastic-like skin texture fall in between, further showing the spectrum of detectability across different artifact categories and visibility.

5.5 Accuracy by Randomized Display Time

By randomizing the display time of images in this experiment, our results support evaluating how viewing duration influences participants' accuracy. We find that longer viewing times improve performance. With just 1 second of display time, participants are 72% accurate (95% CI=[71.6, 72.5], 95% CI=[71.3, 72.2]) on AI-generated and real images, respectively. With 5 seconds of display time, accuracy increases to 77% (95% CI=[77.0, 77.8], 95% CI=[76.6, 77.4]) for both AI-generated and real images, respectively. While accuracy on real images appears to plateau by 5 seconds of display time, accuracy on AI-generated images increases up to 80% (95% CI=[79.6, 80.4]) at 10 seconds and 82% (95% CI=[81.2, 81.9]) at 20 seconds. Figure 11B presents the distribution of accuracy scores across display time conditions. Across the observations where display time was randomized, we find that the proportion of AI-generated images that are identified below random chance decreases from 43% when participants only have 1 second to view the image to 30%, 25%, 17%, and 17% when participants have 5, 10, 20 seconds, and unlimited time to view the image.

In some images, AI artifacts can be noticed with a quick glance, but for others, careful attention to detail is necessary to spot the artifact. Figure 13 presents three images that require careful attention, as evidenced by the fact that most participants mark as real when they are limited to seeing the image for a second but fake once they take into account the details of the scene.

Accuracy across all artifact types improved with increased display time. As shown in Figure 11C, participants showed higher accuracy when images were displayed for longer time (anatomical artifacts: 63% at 5 seconds vs. 59% at 1 second; stylistic artifacts: 63% at 5 seconds vs. 60% at 1 second; functional artifacts: 60% at 5 seconds vs. 55% at 1 second). For all artifacts, there is a significant improvement in detection accuracy when increasing display time from 5 seconds to unlimited.

In Figure 11C, we observe that participants improved the most in identifying functional artifacts, with an 18% improvement from 1 second to unlimited viewing time. In comparison, anatomical and stylistic artifacts showed smaller improvements of 11% each over the same time interval. Unlike anatomical and stylistic implausibilities that can be identified at first glance, functional artifacts often require a closer look and familiarity with the elements in the image

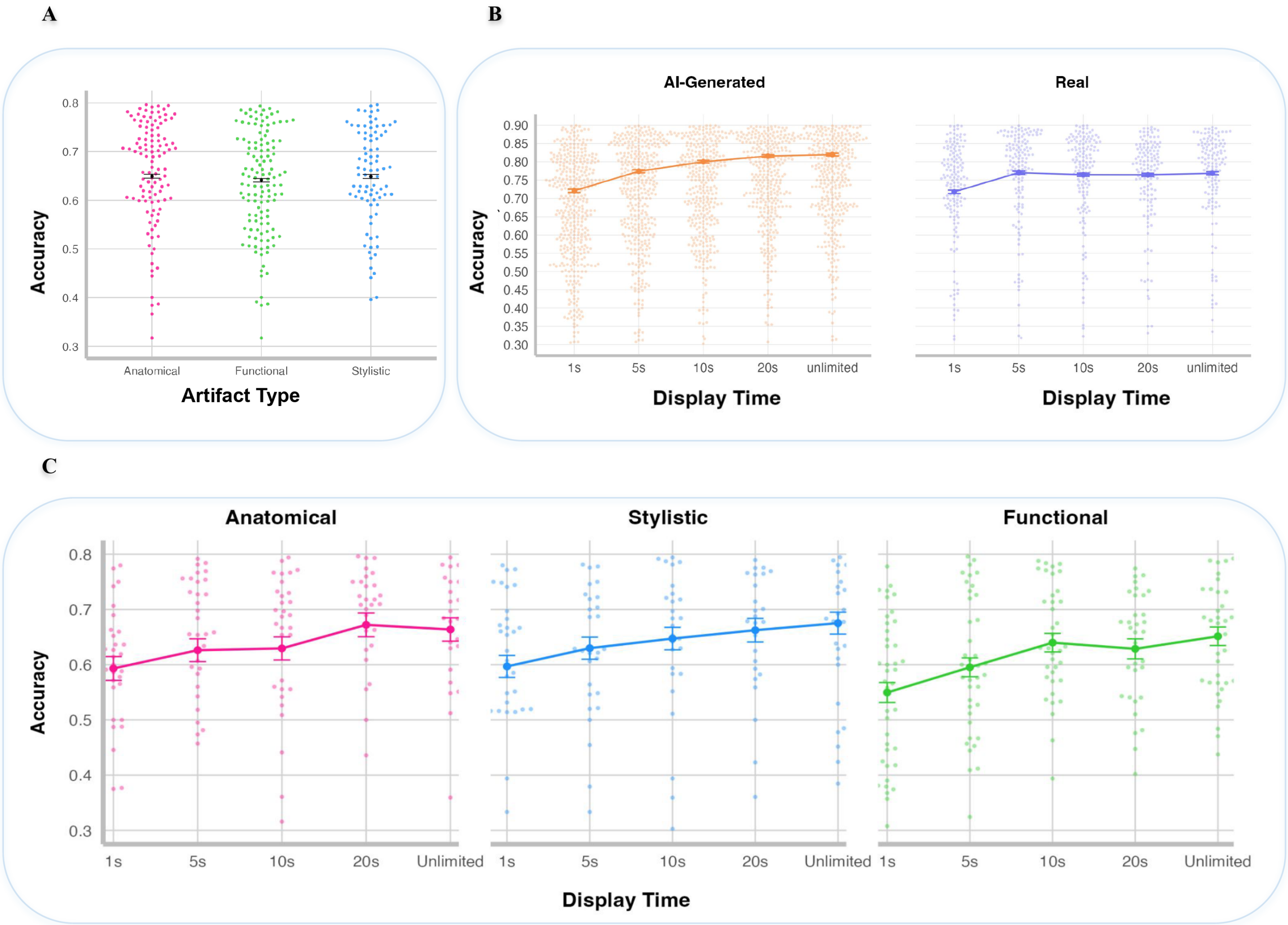


Figure 11: Accuracy by artifact types and display times A. Distribution of accuracy scores by artifact type for images with at least one artifact. **B. Mean Accuracy Over Display Time.** Change in mean accuracy across different display time assignments (1 second, 5 seconds, 10 seconds, 20 seconds, and unlimited) with 95% confidence intervals and bee swarm plots of image accuracy for AI-generated and real images. **C. Mean Accuracy Over Time for Different Artifact Types.** Change in mean accuracy across different time assignments (1 second, 5 seconds, 10 seconds, 20 seconds, and unlimited) with 95% confidence intervals and bee swarm plots of image accuracy for images with Anatomical (pink), Functional (green), and Stylistic (blue) artifacts. The x–axis shows the display time intervals, and the y–axis shows accuracy.

as they often appear in parts of the image that are not the main subject.

5.6 Qualitative Analysis of Participant Comments

We collected 34,675 comments from participants who filled out the optional text input box asking participants: "If you think this is AI-generated, please explain why." In order to identify themes from these 34,675 comments, we prompted GPT-3.5 Turbo to identify 10 main themes across these comments. GPT-3.5 Turbo responded with the following ten themes, which we manually reviewed and refined to mitigate the ambiguities and generalization typical of large language models [75]: (1) Image quality focusing on the overall appearance, smoothness, and sometimes unrealistic perfection of image elements; (2) Facial and anatomical inconsistencies where participants pointed to irregularities in eyes, mouths, noses, skin texture, expressions, and general human anatomy; (3) Anatomical and functional anomalies such as deformities, misplaced body parts, and irregularities in objects or environments; (4) Lighting and environmental inconsistencies including unnatural lighting, inconsistent shadows, and reflections; (5) Digital manipulation indicators suggesting suspicions of AI-generation or digital alteration; (6) Biometric discrepancies particularly unnatural or imperfect body parts like hands and fingers; (7) Uncanny valley perceptions where images almost looked human but had subtle unnatural features that caused discomfort; (8) Contextual incongruities such as unrealistic scenarios and mismatched social elements; (9) Physical anomalies highlighting illogical physical interactions within the images; and (10) holistic authenticity assessment making overall judgments based on a combination of multiple cues and inconsistencies. Based on these ten main themes, we prompted GPT-3.5 to label each comment with one of the ten themes. Figure 15 illustrates

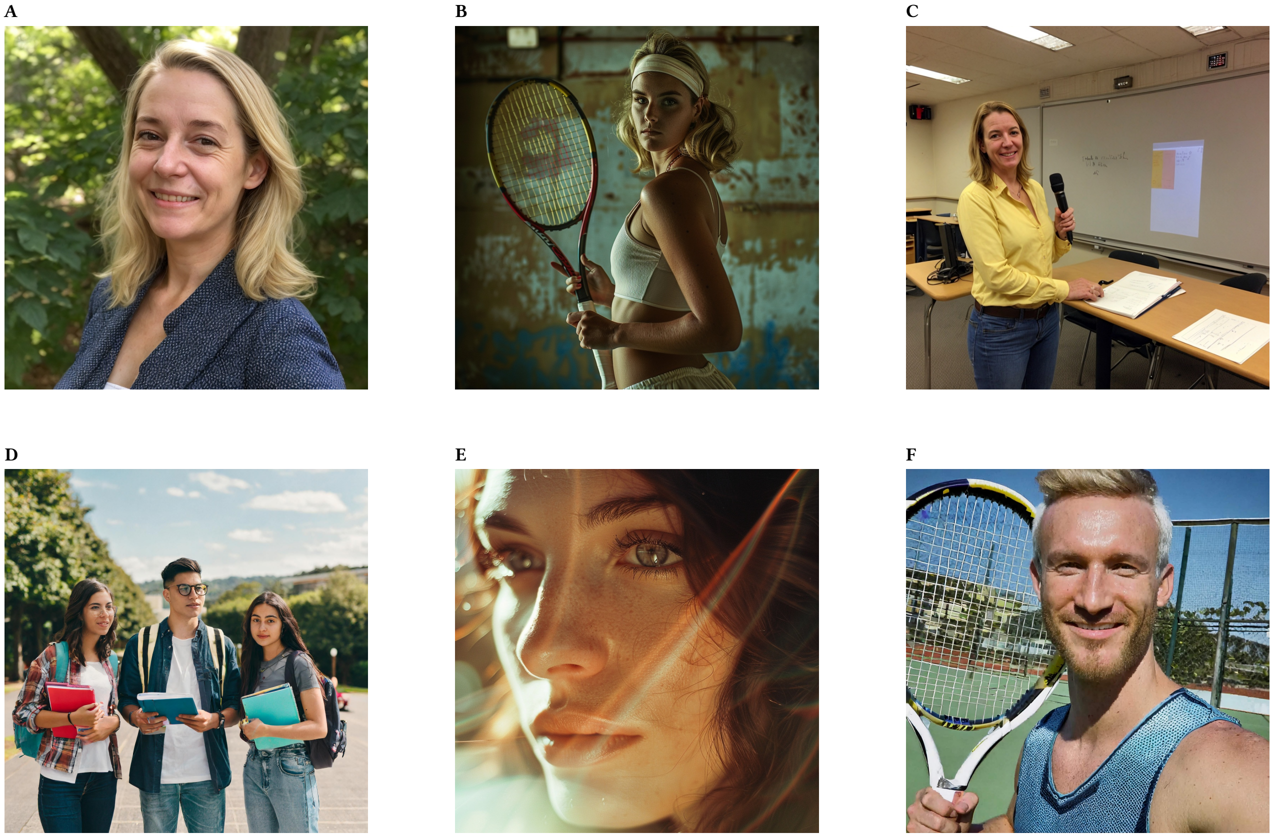


Figure 12: Examples of images with varying artifact visibility. Top row (A–C): Example images showcasing stylistic and functional artifacts with varying visibility. **A.** A subtle stylistic artifact in the soft and wispy textures of the woman’s hair and a minor functional implausibility in the atypical design of her shirt collar (Accuracy: 47%). **B.** An obvious stylistic artifact due to the overall cinematization of the image (Accuracy: 73%). **C.** A combination of multiple artifacts, including anatomical implausibilities in the woman’s hand, functional implausibilities in the table shape and wall panels, and a stylistic artifact in the soft texture of the woman’s face (Accuracy: 38%). **Bottom row (D–F):** Images with anatomical, stylistic, and functional artifacts of varying visibility. **D.** Anatomical implausibilities in the fingers of the three students (Accuracy: 84%). **E.** A stylistic artifact in the cinematized look and plastic-like texture of the woman’s skin (Accuracy: 62%). **F.** No obvious anatomical or stylistic artifacts, but closer inspection reveals functional implausibilities: the tennis racket is asymmetrical, its strings are not taut, and the shirt has irregularly shaped designs with glitch-like inconsistencies (Accuracy: 54%).

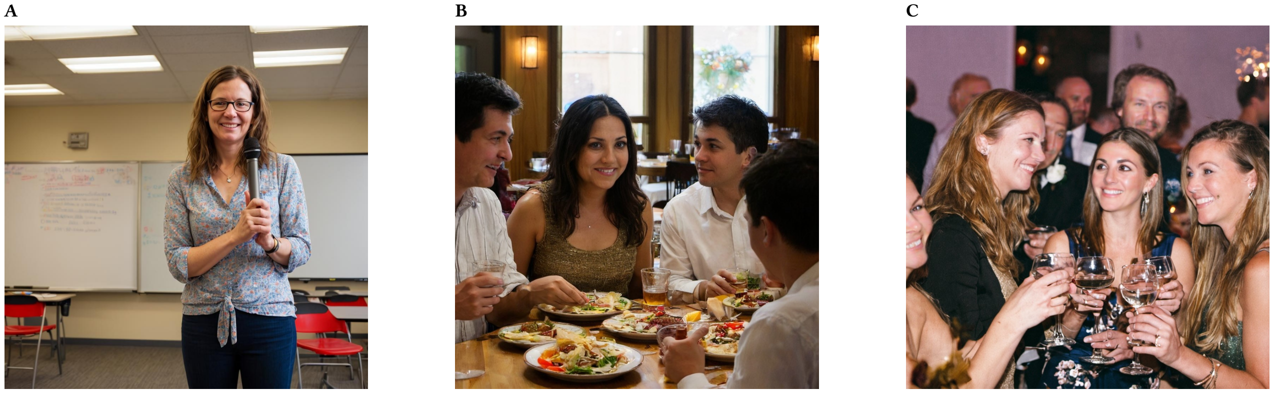


Figure 13: Exemplar AI-generated images for which a closer look improves accuracy. A. Accuracy: 38% at 1 second display time to 65% at 20 second display time. **B.** Accuracy: 44% at 1 second display time to 82% at 20 second display time. **C.** Accuracy: 27% at 1 second display time to 70% at 20 second display time.

examples of participant comments for four images and how they were categorized into themes. Figure 14 displays the distribution of themes across the comments and the related concept from our taxonomy in parentheses.

Based on GPT-3.5 Turbo, we find that 61% of participants' comments mentioned relying on anatomical implausibilities. The next most common concept referred to is stylistic artifacts, which is mentioned in 30% of comments. Participants mentioned functional implausibilities in 21% of comments, violations of physics in 15% of comments, and sociocultural implausibilities in only 4% of comments.

Based on the authors' annotations of artifacts, we find functional implausibilities to be the most prevalent, appearing in 58.7% of images, followed by anatomical implausibilities in 51.4% and stylistic artifacts in 39.0% of images. We identify violations of physics and sociocultural implausibilities in only 9.17% and 5.50% of images, respectively.

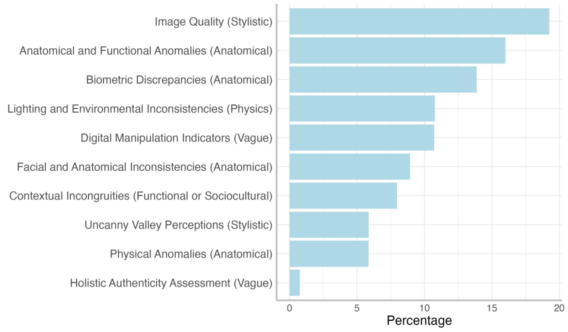


Figure 14: Distribution of themes identified in participant comments.

While functional artifacts were the most prevalent in human researcher annotated images, they were less frequently mentioned in participant comments annotated by GPT–3.5. Conversely, anatomical artifacts were emphasized more in participant comments than in their prevalence in annotated images.

5.7 Accuracy by Models

In the process of generating the images for this experiment's stimuli set, we noticed that Midjourney, Firefly, and Stable Diffusion have different capabilities and limitations. For example, we noticed that Midjourney often produced images with persistent stylistic artifacts that were challenging to eliminate. Firefly, on the other hand, frequently exhibited a tendency toward synthetic emotional expressions, with subjects often appearing unnaturally and overly cheerful, necessitating multiple iterations to produce more realistic results. Stable Diffusion struggled significantly with generating group images, often introducing artifacts such as anatomical inconsistencies. In light of the limitations to generate non-portrait images with Stable Diffusion, 75% of the Stable Diffusion-generated stimuli in this experiment were portraits. On the other hand, 30% of Midjourney and Firefly-generated images in this experiment depict portraits. In order to compare the three models fairly, we focus our comparison on portrait images. Figure 16 presents accuracy shown on portraits by each of the three models and reveals that participants' mean accuracy on Midjourney, Stable Diffusion, and Firefly were 76% (95% CI: [75.2, 75.8]), 74% (95% CI: [73.9, 74.8]), and 73% (95% CI: [72.7, 73.3]), respectively.

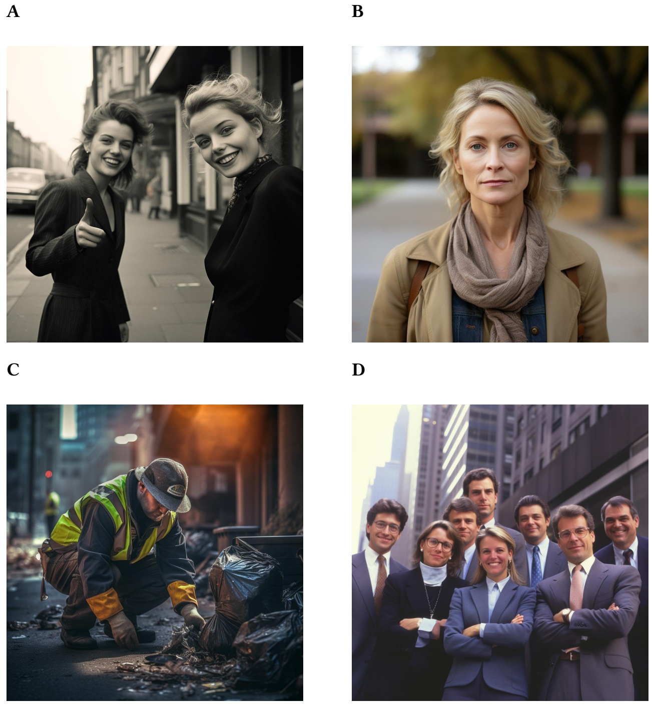


Figure 15: Examples of participant comments mapped to themes. A. "Cosmetic style out of character with vintage setting": Contextual Incongruities. **B.** "Skin too smooth, depth of field shallow.": Image Quality, Lighting Inconsistencies. **C.** "If this is not AI then it is a staged photograph like a movie set because of the lighting and he is an actor.": Lighting inconsistencies, Contextual Incongruities. **D.** "Group looks pasted onto background.": Digital Manipulation Indicators.

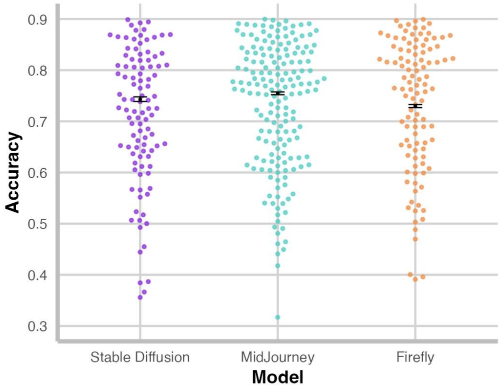


Figure 16: Accuracy across generative AI models Each point represents an image. The black dots and error bars show the mean accuracy and 95% bootstrapped confidence intervals for each model

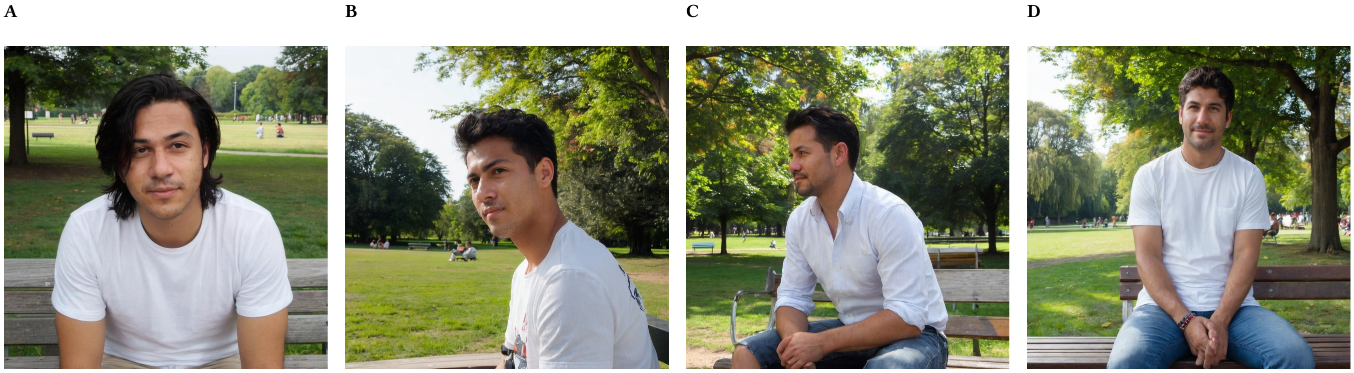


Figure 17: Re-generated images from the same prompt. A. Stage 1 image generated by Stable Diffusion and curated by our team (37% accuracy) **B.** Most photorealistic of 12 prompt-matched image generations by Stable Diffusion (42% accuracy) **C.** Median photorealistic of 12 prompt-matched image generations by Stable Diffusion (59% accuracy) **D.** Least photorealistic of 12 prompt-matched image generations by Stable Diffusion(83% accuracy)

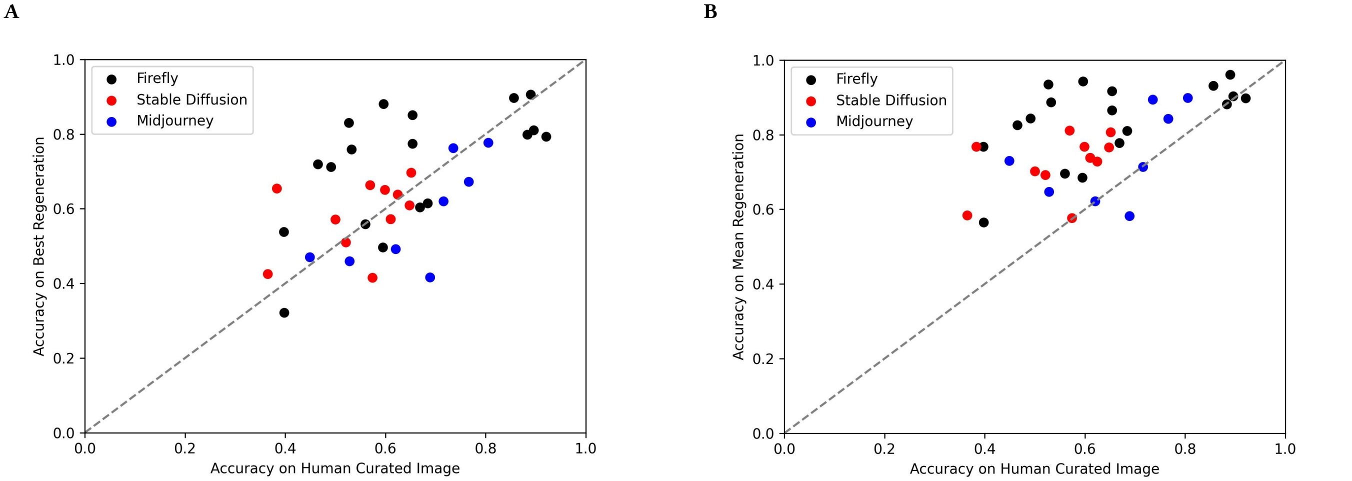


Figure 18: Comparing accuracy scores on curated images and uncurated prompt-matched images. A. Scatterplot showing human detection accuracy of the original curated image compared to human detection accuracy of its most photorealistic regeneration out of 11 to 24 prompt-matched images labeled as re-generations. **B.** Scatterplot showing human detection accuracy of the original curated image compared to human detection accuracy of its mean photorealistic regeneration out of 11 to 24 re-generations.

5.8 Accuracy on Human Curated Images vs. Uncurated Images

Generating photorealistic AI-generated images involves three key ingredients: the diffusion model, the prompt, and human curation. In this section, we examine how human curation of diffusion model-generated images influences the aggregate accuracy scores of human participants. In order to show this influence, we compare diffusion model images from the main experiment, which were curated by our research team, with multiple diffusion model images generated from the same prompt as the curated images. This comparison reveals the increase in photorealism (as measured by the decrease in participants' accuracy) on the curated images relative to the prompt-matched images.

In this second phase of the experiment, we randomly sampled 39 AI-generated images from the main stimuli set, where the sample was stratified on 10 percentage point wide bins on human detection accuracy. For each of these 39 images, we generated at least 11 prompt-matched images using Midjourney, Firefly, and the same pipeline in Stable Diffusion. Figure 17 displays a Stable Diffusion-generated image from our original stimuli set and three of the twelve generations using the same prompt. We generated 482 total additional images, with at least 11 per prompt. These 482 images were included alongside the 149 real images on the experiment website.

In Figure 18, we present scatterplots comparing human detection accuracy on the initial curated images and the best prompt-matched images in panel A, and mean prompt-matched images in panel B. We find the human-curated images have lower human detection accuracy than the best regenerated image in 18 of 39 instances and the mean re-generated image in 35 of 39 instances. In total, the human-curated images were perceived to be more photorealistic than 408 of the 482 (84%) uncurated prompt-matched images. Specifically, we find the marginal value added by human curation for images that were initially detected in the range of 30% to 50% is 31 percentage points, 50 to 60% is 23 percentage points, 60-70% is 11 percentage points, 70-80% is 8 percentage points, and 80+% is 4 percentage points. Across the stimuli selected from Midjourney, Firefly, and Stable Diffusion, the marginal value of human curation is 7.8, 19.0, and 16.9 percentage points, respectively.

The two panels in Figure 18 illustrate the positive correlation between accuracy on the human-curated image and accuracy on the regeneration. This reveals how the prompt influences photorealism. The Pearson Correlation Coefficient between accuracy on curated images and their best, mean, and worst re-generations are .58, .53, and .32, respectively. This positive correlation suggests the choice of a prompt plays a significant role in the photorealism of an image. Figure S5 displays two original curated images where A is generated by a prompt in which re-generations achieved low human detection accuracy (a 'good' prompt), and B is generated by a prompt in which re-generations achieved a high human detection accuracy (a 'bad' prompt). Prompts that consistently generate easily detectable images often have elements that are difficult to generate and result in artifacts. The prompt "Persian woman astronaut in astronaut clothes, family photo with husband and two toddlers, high resolution, realistic" for Figure S5B generates a posed group image that tends to be easy to detect. On the other hand, the prompt "American woman faculty portrait, not a close-up, blond" for Figure S5A generates a portrait image that tends to be perceived as more photorealistic.

6 Discussion

While diffusion models can generate highly realistic images, most of the images they produce still contain visible artifacts. In particular, we find that only 17% of diffusion model-generated images are misclassified as real at rates consistent with random guessing. Notably, this misclassification rate increases to 43% when the viewing duration is restricted to 1 second. By curating a dataset of 599 images and conducting a large scale digital experiment, we can begin to answer fundamental questions about what drives the appearance of photorealism in diffusion model-generated images.

First, we find that images with greater scene complexity tend to introduce more opportunities for artifacts to appear, making it easier for participants to detect AI-generated images. Our results reveal that participants were less accurate at identifying AI-generated portraits compared to more complex scenes, such as those involving multiple people in candid settings. Based on qualitative analysis of the images, we identify three main reasons for this difference. First, portraits often feature a single person against a blurred background, which can obscure details and provide fewer cues compared to full-body or group images. Second, portraits typically involve fewer and simpler poses, focusing only on the face and torso, leaving fewer opportunities for errors or inconsistencies to be apparent. Third, the prevalence of edited and retouched portraits in real-world photography complicates the distinction between real and AI-generated portraits, addressing the question of how subject type and context (e.g., unknown people vs. public figures) influence the perceived authenticity of an image. In contrast, more complex images, like full-body or group shots, involve a greater number of elements, increasing the likelihood of noticeable errors or inconsistencies. Similar to our results on AI-generated images, we find that real images with lower scene complexity are also harder to identify as real.

Second, we identify five high-level categories of artifacts and implausibilities and find that the easiest images to identify as diffusion model generated are the ones with anatomical implausibilities, such as unrealistic body proportions and stylistic artifacts like overly glossy or waxy features.

Third, by randomizing display time, we identify the relationship between how long an individual looks at an image and their accuracy at distinguishing between real and AI-generated images. Specifically, we find that participants' accuracy at identifying an AI-generated image upon a quick glance of 1 second is 72% and increases by 5 percentage points with just an additional 4 seconds of viewing time and 10 percentage points when unconstrained by time. Given the nature of rapid scrolling on social media and how much time people have to see advertisements as they pass by billboards on a highway, these results reveal the importance of attentive viewing of images before making judgments about an image's veracity.

Fourth, we find that human curation had a notable negative impact on participants' accuracy compared to uncurated images generated by the same prompts as the human-curated AI-generated images. In particular, the images curated by our research team were harder to identify as AI-generated than 84% of the uncurated images generated using the same prompts as the curated images. This finding reveals the limitation of state-of-the-art diffusion models in producing images of consistent quality. It also suggests that human curation is a bottleneck to generating fake images at scale. The process of generating high-quality AI images is inherently iterative—users refine prompts and select outputs until they achieve their desired result. This fundamental aspect of AI image generation is evident across all applications, from advertising and marketing to education and beyond. While concerns exist about fake images being used to mislead or impersonate, many use cases exist for business and educational applications [27, 31, 78]. The critical role of human curation in this iterative process further emphasizes how the photorealism of images produced by diffusion models depends not only on the capabilities of the diffusion model but also on the quality of human curation, choice of prompts, and context of the scene. Given the importance of these factors beyond the generative AI model, these results reveal the importance of considering these factors in research examining human perception of AI-generated images. Without considering these elements, it is possible to produce biased findings showing AI-generated images are more or less realistic than they really appear in real-world settings.

The taxonomy offers a practical framework on which to build AI literacy tools for the general public. We synthesized information from diverse sources such as social media posts, scientific literature, and our online behavioral study with 50,444 participants to systematically categorize artifacts in AI-generated images. Through this process, we identify five key categories: anatomical implausibilities, which involve unlikely artifacts in individual body parts or inconsistent proportions, particularly in images with multiple people; stylistic artifacts, referring to overly glossy, waxy, or picturesque qualities of specific elements of an image; functional implausibilities, arising from a lack of understanding of real-world mechanics and leading to objects or details that appear impossible or nonsensical; violations of physics, which include inconsistencies in shadows, reflections, and perspective that defy physical logic; and sociocultural implausibilities, focusing on scenarios that violate social norms, cultural context, or historical accuracy. Our taxonomy builds upon the Borji 2023 taxonomy [7] and focuses on images that appear more realistic at first glance, which is useful for comparing and contrasting real photographs with diffusion model generated images for revealing the nuances of the artifacts and implausibilities [40]. Moreover, this taxonomy offers a shared language by which practitioners and researchers can communicate about artifacts commonly seen in AI-generated images and exposes the persistent challenges that can help people identify AI-generated images.

6.1 Future Work and Limitations

In addition to aiding in identifying AI-generated content, the taxonomy offers insights into the open problems for producing realistic AI-generated images. Future work may explore integrating such taxonomies into model evaluation frameworks to provide iterative feedback during the development of generative models. As models advance to address the weaknesses presented in this taxonomy, new and more subtle artifacts may emerge, requiring future updates to this taxonomy. This dynamic interplay between detection and generation capabilities demonstrates why we need to maintain robust human detection abilities even as models evolve. We acknowledge the potential dual use of these insights to create more deceptive synthetic media, and we believe that transparent documentation of artifacts does more good than harm by offering detection strategies and an opportunity to develop general awareness in the public.

Large-scale digital experiments with participants who participate based on their own interests come with certain limitations. First, we did not collect demographic data from participants. Participants were not recruited for this experiment; instead, participants found the experiment organically and participated. Given the organic nature of the participation, we prioritized maximizing engagement, which involves questions unrelated to distinguishing AI-generated and real images like demographic questions. While this approach enabled substantial data collection, it limits analysis by excluding factors like age, gender, and cultural background that may influence detection.

Second, we provided feedback on the correct answer after each participant made an observation, which has the potential to introduce learning effects. Future research could address these open questions by collecting demographic data to design more inclusive AI literacy tools and evaluating how performance changes with and without feedback.

This research focused on images generated by state-of-the-art generative models available in 2024, and the findings are inherently tied to the state of diffusion models and generative AI technologies as of 2024. In the future, models are likely to change, and the somewhat visible errors that emerge will also likely change. Past state-of-the-art GAN models such as StyleGAN2 [43] and BigGAN [8], often produced more noticeable artifacts in facial features, color balance, and overall photorealism, making their outputs more easily distinguishable. Nonetheless, the current taxonomy on diffusion models points out elements like anatomical implausibilities and stylistic artifacts that can be mapped to the facial feature and color balance cues. These recurring issues offer evidence of the taxonomy's robustness to differences across model generations, but future studies should explore how the taxonomy may need to adapt to these changes, which may involve adding or removing categories or may involve further identifying nuances within these categories. As an example of how this taxonomy may be applied to AI-generated video, Figure S6 presents an example of an anatomical implausibility that we never saw in diffusion model-generated images because it involves a temporal inconsistency. Future research on the realism of AI-generated audio and video may also consider following the three-step process involved in building this taxonomy for images generated by diffusion models. Based on first surveying AI literacy resources, academic literature, and social media, second generating media with state-of-the-art models, and third collecting empirical data on the human ability to distinguish AI-generated media from authentically recorded media, researchers can build empirical insights towards characterizing realism and categorizing the artifacts in AI-generated media.

The empirical insights on the photorealism of AI-generated images and the resulting taxonomy designed to help people better navigate real and synthetic images online lead to a practical research question: How can AI literacy interventions improve people's ability to distinguish real photographs and AI-generated images? Future research may address this question via randomized experiments comparing a control group with no intervention to a treatment group that receives training based on the taxonomy presented in this paper. Likewise, future research may explore this with just-in-time interventions to direct people's attention to the cues identified in the taxonomy.

7 Conclusion

Our work contributes empirical insights on the photorealism of AI-generated images and a taxonomy of artifacts commonly found in AI-generated images, organized into five categories: anatomical implausibilities, stylistic artifacts, functional implausibilities, violations of physics, and sociocultural implausibilities. We find that the photorealism of AI-generated images depends on the scene complexity of the image, the kind of artifacts and implausibilities, if any, detectable in an image, the duration of visual attention to an image, and the quality of human effort to select appropriate prompts and curate images. A question such as "How photorealistic are state-of-the-art diffusion models" may sound simple, but the answer is more complex and depends on many details, including what images are generated and selected, how photorealism is measured, what real images are included in the experiment, and how much time, skill, and effort a human participant has and willing to offer. This paper offers an initial exploration into how we can address this question and develops a practical taxonomy that offers scaffolding for building AI–literacy interventions to help people navigate the capabilities and limitations of diffusion models and whether an image is AI-generated or not.

Acknowledgments

This material is based upon work supported by Robert Pozen, and in part with funding from the Department of Defense (DoD). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DoD or any agency or entity of the United States Government. We thank Will Thompson from Kellogg Research Support for performing a replication check.

References

[1] Adobe. 2024. *Adobe Firefly.* Adobe Inc. https://www.adobe.com/products/firefly.html Accessed: 2024-08-19.

[2] Associated Press. 2024. Trump arrested? Putin jailed? Fake AI images spread online. https://apnews.com/article/ai-misinformation-trump-putin-new-york-42ac9c41c5504d05412b492e48bcaded Accessed: 2024-08-26.

[3] Associated Press. 2024. Trump, Harris, and the Detroit Crowd Size Photo Controversy. https://apnews.com/article/trump-harris-detroit-crowd-size-photo-ff54a66d8e3197c90068ba94847297cf Accessed: 2024-08-26.

[4] Quentin Bammey. 2024. Synthbuster: Towards Detection of Diffusion Model Generated Images. *IEEE Open Journal of Signal Processing* 5 (2024), 1–9. doi:10.1109/OJSP.2023.3337714

[5] Sarah Barrington and Hany Farid. 2024. People are poorly equipped to detect AI-powered voice clones. arXiv:2410.03791 [cs.HC] https://arxiv.org/abs/2410.03791

[6] Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman. 2023. Detecting Generated Images by Real Images Only. arXiv:2311.00962 [cs.CV] https://arxiv.org/abs/2311.00962

[7] Ali Borji. 2023. Qualitative failures of image generation models and their application in detecting deepfakes. *Image and Vision Computing* 137 (2023), 104771.

[8] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv:1809.11096 [cs.LG] https://arxiv.org/abs/1809.11096

[9] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376789

[10] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. arXiv:2304.06408 [cs.CV] https://arxiv.org/abs/2304.06408

[11] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. 2024. Raising the Bar of AI-generated Image Detection with CLIP. arXiv:2312.00195 [cs.CV] https://arxiv.org/abs/2312.00195

[12] Reality Defender. 2024. Deepfake Detection Guide. https://www.realitydefender.com/blog/deepfake-detection-guide Accessed: 2024-08-19.

[13] Chengdong Dong, Ajay Kumar, and Eryun Liu. 2022. Think Twice Before Detecting GAN-generated Fake Images from their Spectral Domain Imprints. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Piscataway, NJ 08854 USA, 7855–7864. doi:10.1109/CVPR52688.2022.00771

[14] R. Durall, M. Keuper, and J. Keuper. 2020. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Piscataway, NJ 08854 USA, 7887–7896. doi:10.1109/CVPR42600.2020.00791

[15] David C. Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. 2023. Online Detection of AI-Generated Images. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, Piscataway, NJ 08854 USA, 382–392. doi:10.1109/ICCVW60793.2023.00045

[16] Ziv Epstein, Mengying C Fang, Antonio A Arechar, and David G Rand. 2023. What label should be applied to content produced by generative AI? doi:10.31234/osf.io/v4mfz

[17] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.

[18] eWeek Staff. 2024. What is a Deepfake? Understanding the Technology and its Impact. https://www.eweek.com/artificial-intelligence/deepfake/ Accessed: 2024-08-19.

[19] Hany Farid. 2022. Lighting (In)consistency of Paint by Text. arXiv:2207.13744 [cs.CV] https://arxiv.org/abs/2207.13744

[20] Hany Farid. 2022. Perspective (In)consistency of Paint by Text. arXiv:2206.14617 [cs.GR] https://arxiv.org/abs/2206.14617

[21] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, Vienna, Austria, 3247–3258. https://proceedings.mlr.press/v119/frank20a.html

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc., Boston, 2672–2680. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[23] Matthew Groh. 2022. Identifying the Context Shift between Test Benchmarks and Production Data. arXiv:2207.01059 [cs.LG] https://arxiv.org/abs/2207.01059

[24] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022), e2110013119. doi:10.1073/pnas.2110013119 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2110013119

[25] Matthew Groh, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, and Rosalind Picard. 2024. Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications* 15, 1 (2024), 7629.

[26] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. 2022. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, Piscataway, NJ 08854 USA, 2904–2908.

[27] Andrés Gvirtz and Oguz A Acar. 2023. Why Text-to-Image AI Requires a New Branding Mindset. *MIT Sloan Management Review* 65, 1 (2023), 1–4.

[28] Anna Yoo Jeong Ha, Josephine Passananti, Ronik Bhaskar, Shawn Shan, Reid Southen, Haitao Zheng, and Ben Y. Zhao. 2024. Organic or Diffused: Can We Distinguish Human Art from AI-generated Images? arXiv:2402.03214 [cs.CV] https://arxiv.org/abs/2402.03214

[29] Michael Hameleers, Toni GLA van der Meer, and Tom Dobber. 2024. They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication* 39, 1 (2024), 56–70.

[30] Chaeeun Han, Prasenjit Mitra, and Syed Masum Billah. 2024. Uncovering Human Traits in Determining Real and Spoofed Audio: Insights from Blind and Sighted Individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 949, 14 pages. doi:10.1145/3613904.3642817

[31] Jochen Hartmann, Yannick Exner, and Samuel Domdey. 2023. The power of generative marketing: Can generative AI reach human-level visual marketing content? *Available at SSRN* (2023).

[32] Eliot Higgins. 2023. Tweet by Eliot Higgins, March 2023. https://x.com/EliotHiggins/status/1637927681734987777 Accessed: 2024-08-19.

[33] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[34] Shu Hu, Yuezun Li, and Siwei Lyu. 2021. Exposing GAN-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, Piscataway, NJ 08854 USA, 2500–2504.

[35] Nils Hulzebosch, Sarah Ibrahimi, and Marcel Worring. 2020. Detecting CNN-Generated Facial Images in Real-World Scenarios. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Piscataway, NJ 08854 USA, 2729–2738. doi:10.1109/CVPRW50498.2020.00329

[36] Transparent Statistics in Human–Computer Interaction Working Group. 2019. Transparent Statistics Guidelines. doi:10.5281/zenodo.1186169 (Available at https://transparentstats.github.io/guidelines).

[37] Benjamin N Jacobsen. 2024. Deepfakes and the promise of algorithmic detectability. *European Journal of Cultural Studies* 0, 0 (2024), 13675494241240028. doi:10.1177/13675494241240028

[38] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2022. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America* 120, 11 (2022), e2208839120. https://api.semanticscholar.org/CorpusID:249674779

[39] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. 2024. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, Piscataway, NJ 08854 USA, 4324–4333.

[40] Negar Kamali, Karyn Nakamura, Angelos Chatzimparmpas, Jessica Hullman, and Matthew Groh. 2024. How to Distinguish AI-Generated Images from Authentic Photographs. arXiv:2406.08651 [cs.HC] https://arxiv.org/abs/2406.08651

[41] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, Online, 26 pages. https://openreview.net/forum?id=Hk99zCeAb

[42] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Piscataway, NJ 08854 USA, 4401–4410.

[43] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Piscataway, NJ 08854 USA, 8107–8116. doi:10.1109/CVPR42600.2020.00813

[44] Federica Lago, Cecilia Pasquini, Rainer Bohme, Helene Dumont, Valerie Goffaux, and Giulia Boato. 2022. More Real Than Real: A Study on Human Visual Perception of Synthetic Faces [Applications Corner]. *IEEE Signal Processing Magazine* 39, 1 (Jan. 2022), 109–116. doi:10.1109/msp.2021.3120982

[45] N. S. Lakshmiprabha, J. Bhattacharya, and S. Majumder. 2011. Face recognition using multimodal biometric features. In *2011 International Conference on Image Information Processing*. IEEE Computer Society, Piscataway, NJ 08854 USA, 1–6. doi:10.1109/ICIIP.2011.6108945

[46] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Fei-Fei Li, Jiajun Wu, Stefano Ermon, and Percy S Liang. 2023. Holistic Evaluation of Text-to-Image Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 57 Morehouse Lane; Red Hook; NY; United States, 69981–70011. https://proceedings.neurips.cc/paper_files/paper/2023/file/dd83eada2c3c74db3c7fe1c087513756-Paper-Datasets_and_Benchmarks.pdf

[47] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ 08854 USA, 19401–19411.

[48] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. 2024. Detecting Multimedia Generated by Large AI Models: A Survey. arXiv:2402.00045 [cs.MM] https://arxiv.org/abs/2402.00045

[49] ltdrdata. 2024. ComfyUI-Impact-Pack. https://github.com/ltdrdata/ComfyUI-Impact-Pack. https://github.com/ltdrdata/ComfyUI-Impact-Pack Accessed: 2024-04-12.

[50] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

[51] Siwei Lyu and Hany Farid. 2005. How realistic is photorealistic? *IEEE Transactions on Signal Processing* 53, 2 (2005), 845–850.

[52] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. 2023. Exposing the Fake: Effective Diffusion-Generated Images Detection. arXiv:2307.06272 [cs.CV] https://arxiv.org/abs/2307.06272

[53] New York Magazine. 2024. Ukraine War Diary: Latest Updates. https://nymag.com/intelligencer/article/ukraine-war-diary-updates.html Accessed: 2024-08-19.

[54] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. 2019. Do GANs Leave Artificial Fingerprints?. In *2nd IEEE Conference on Multimedia Information Processing and Retrieval, MIPR 2019, San Jose, CA, USA, March 28-30, 2019*. IEEE Computer Society, Piscataway, NJ 08854 USA, 506–511. doi:10.1109/MIPR.2019.00103

[55] Elizabeth J Miller, Ben A Steward, Zak Witkower, Clare AM Sutherland, Eva G Krumhuber, and Amy Dawel. 2023. AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological science* 34, 12 (2023), 1390–1403.

[56] Jaron Mink, Miranda Wei, Collins W. Munyendo, Kurt Hugenberg, Tadayoshi Kohno, Elissa M. Redmiles, and Gang Wang. 2024. It's Trying Too Hard To Look Real: Deepfake Moderation Mistakes and Identity-Based Bias. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 778, 20 pages. doi:10.1145/3613904.3641999

[57] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (jan 2021), 41 pages. doi:10.1145/3425780

[58] Shivansh Mundra, Gonzalo J. Aniano Porcile, Smit Marvaniya, James R. Verbus, and Hany Farid. 2023. Exposing GAN-Generated Profile Photos From Compact Embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, Piscataway, NJ 08854 USA, 884–892.

[59] Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems* 22, 3 (2013), 336–359.

[60] Sophie J Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* 119, 8 (2022), e2120481119.

[61] Norton. 2024. What Are Deepfakes? https://us.norton.com/blog/emerging-threats/what-are-deepfakes Accessed: 2024-08-19.

[62] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2024. Towards Universal Fake Image Detectors that Generalize Across Generative Models. arXiv:2302.10174 [cs.CV] https://arxiv.org/abs/2302.10174

[63] Kyle Orland. 2024. Kamala Harris' Rally Crowds Aren't AI-Generated. Here's How You Can Tell. https://www.wired.com/story/kamala-harris-rally-crowds-ai-trump-conspiracy/

[64] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *CVPR*. IEEE, Los Alamitos, CA, USA, 14277–14286. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2023.html#OtaniTSINRH023

[65] Qiyao Peng, Yingdan Lu, Yilang Peng, Sijia Qian, Xinyi Liu, and Cuihua Shen. 2024. Crafting Synthetic Realities: Examining Visual Realism and Misinformation Potential of Photorealistic AI-Generated Images. arXiv:2409.17484 [cs.CY] https://arxiv.org/abs/2409.17484

[66] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, Online, 21 pages. https://openreview.net/forum?id=di52zR8xgf

[67] Gonzalo J. Aniano Porcile, Jack Gindi, Shivansh Mundra, James R. Verbus, and Hany Farid. 2024. Finding AI-Generated Faces in the Wild. arXiv:2311.08577 [cs.CV] https://arxiv.org/abs/2311.08577

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, Virtual Event, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[69] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. ICLR, Online, 16 pages. http://arxiv.org/abs/1511.06434

[70] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. 2024. Towards the Detection of Diffusion Model Deepfakes. arXiv:2210.14571 [cs.CV] https://arxiv.org/abs/2210.14571

[71] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 10674–10685. doi:10.1109/CVPR52688.2022.01042

[72] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. 2024. Shadows Don't Lie and Lines Can't Bend! Generative Models don't know Projective Geometry... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Piscataway, NJ 08854 USA, 28140–28149.

[73] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. 2006. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proc. IEEE* 94, 11 (2006), 1948–1962.

[74] TechTarget Staff. 2024. How to Detect Deepfakes Manually and Using AI. https://www.techtarget.com/searchsecurity/tip/How-to-detect-deepfakes-manually-and-using-AI Accessed: 2024-08-19.

[75] Moritz Stephan, Alexander Khazatsky, Eric Mitchell, Annie S Chen, Sheryl Hsu, Archit Sharma, and Chelsea Finn. 2024. RLVF: Learning from Verbal Feedback without Overgeneralization. arXiv:2402.10893 [cs.LG] https://arxiv.org/abs/2402.10893

[76] Stuart Thompson. 2024. AI is Getting Better Fast. Can You Tell What's Real Now? https://www.nytimes.com/interactive/2024/06/24/technology/ai-deepfake-facebook-midjourney-quiz.html Accessed: 2024-08-27.

[77] u/KudzuEye. 2023. Boring America Photorealism - Reddit Post. https://www.reddit.com/r/midjourney/comments/157hsdd/boring_america_photorealism/ Accessed: 2024-04-19.

[78] Henriikka Vartiainen and Matti Tedre. 2023. Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity* 34, 1 (2023), 1–21.

[79] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Piscataway, NJ 08854 USA, 8695–8704.

[80] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. 2023. Deepfake Detection: A Comprehensive Study from the Reliability Perspective. arXiv:2211.10881 [cs.CV] https://arxiv.org/abs/2211.10881

[81] Xuan Wang and Zhigang Zhu. 2023. Context understanding in computer vision: A survey. *Computer Vision and Image Understanding* 229 (2023), 103646.

[82] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. DIRE for Diffusion-Generated Image Detection. arXiv:2303.09295 [cs.CV] https://arxiv.org/abs/2303.09295

[83] Chloe Wittenberg, Ziv Epstein, Gabrielle Péloquin-Skulski, Adam J Berinsky, and David G Rand. 2024. Labeling AI-Generated Media Online.

[84] Leslie Wöhler, Martin Zembaty, Susana Castillo, and Marcus Magnor. 2021. Towards Understanding Perceptual Differences between Genuine and Face-Swapped Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 240, 13 pages. doi:10.1145/3411764.3445627

[85] Ziyi Xi, Wenmin Huang, Kangkang Wei, Weiqi Luo, and Peijia Zheng. 2023. AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Piscataway, NJ 08854 USA, 1463–1470. doi:10.1109/APSIPAASC58517.2023.10317126

[86] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2024. A Sanity Check for AI-generated Image Detection. arXiv:2406.19435 [cs.CV] https://arxiv.org/abs/2406.19435

[87] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surveys* 56, 4 (2023), 1–39.

[88] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Piscataway, NJ 08854 USA, 8261–8265. doi:10.1109/ICASSP.2019.8683164

[89] Xin Yang, Yuezun Li, Honggang Qi, and Siwei Lyu. 2019. Exposing GAN-synthesized Faces Using Landmark Locations. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security* (Paris, France) *(IH&MMSec'19)*. Association for Computing Machinery, New York, NY, USA, 113–118. doi:10.1145/3335203.3335724

[90] YOLOv8. 2024. YOLOv8: Real-Time Object Detection. https://yolov8.com Accessed: 2024-08-19.

[91] N. Yu, L. Davis, and M. Fritz. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 7555–7565. doi:10.1109/ICCV.2019.00765

[92] Haitao Zhang. 2022. A Survey of Anti-forensic for Face Image Forgery. *Journal of Information Hiding and Privacy Protection* 4 (01 2022), 41–51. doi:10.32604/jihpp.2022.031707

[93] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

[94] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. 2019. Detecting and Simulating Artifacts in GAN Fake Images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, Piscataway, NJ 08854 USA, 1–6. doi:10.1109/WIFS47025.2019.9035107

[95] Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. 2019. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems* 32 (2019), 3449–3461.

A Further Methodological Details

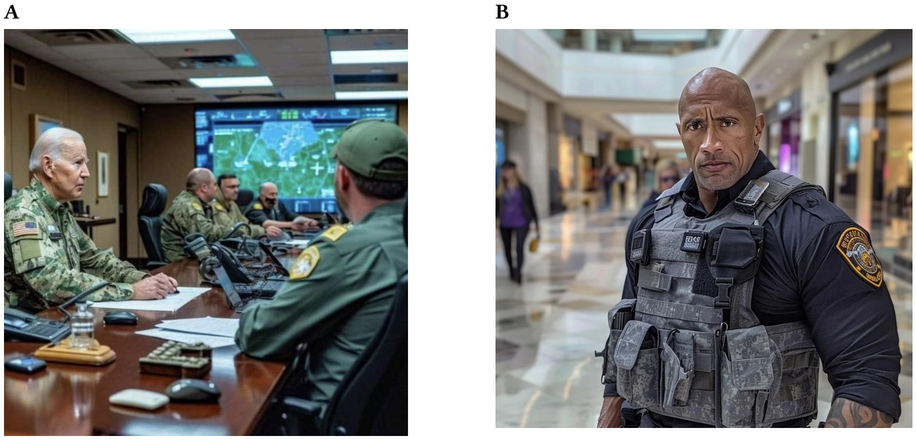


Figure S1: AI-Generated Images from New York Times Quiz A. NYT's explanation for evidence pointing to this image as AI-generated is: "Though the resemblance to President Biden is striking, he would not be wearing military fatigues as a civilian." [76] B. NYT's explanation for evidence pointing to this image as AI-generated is "One giveaway in this image is the badge, which includes garbled text." [76]

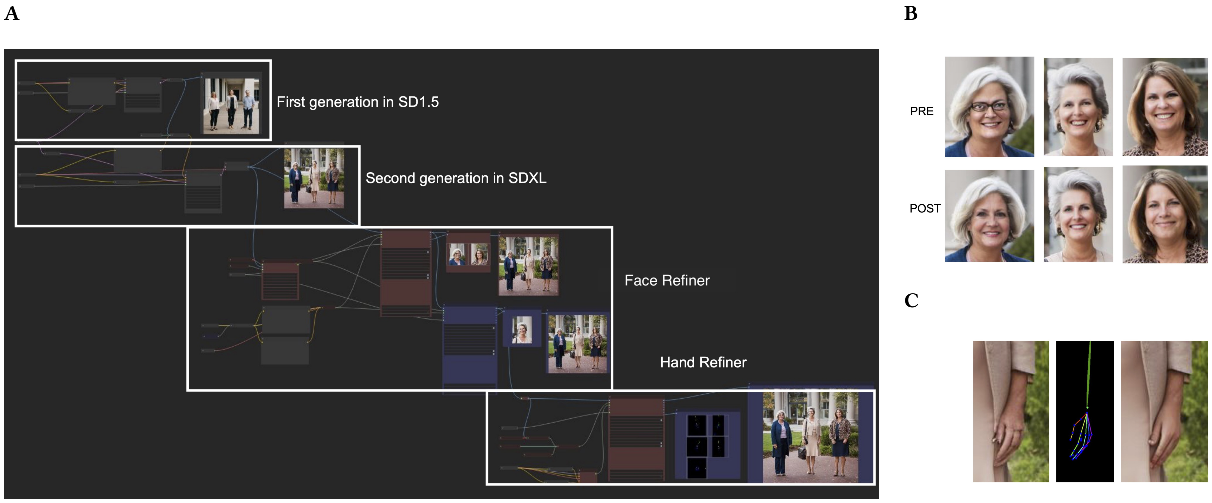
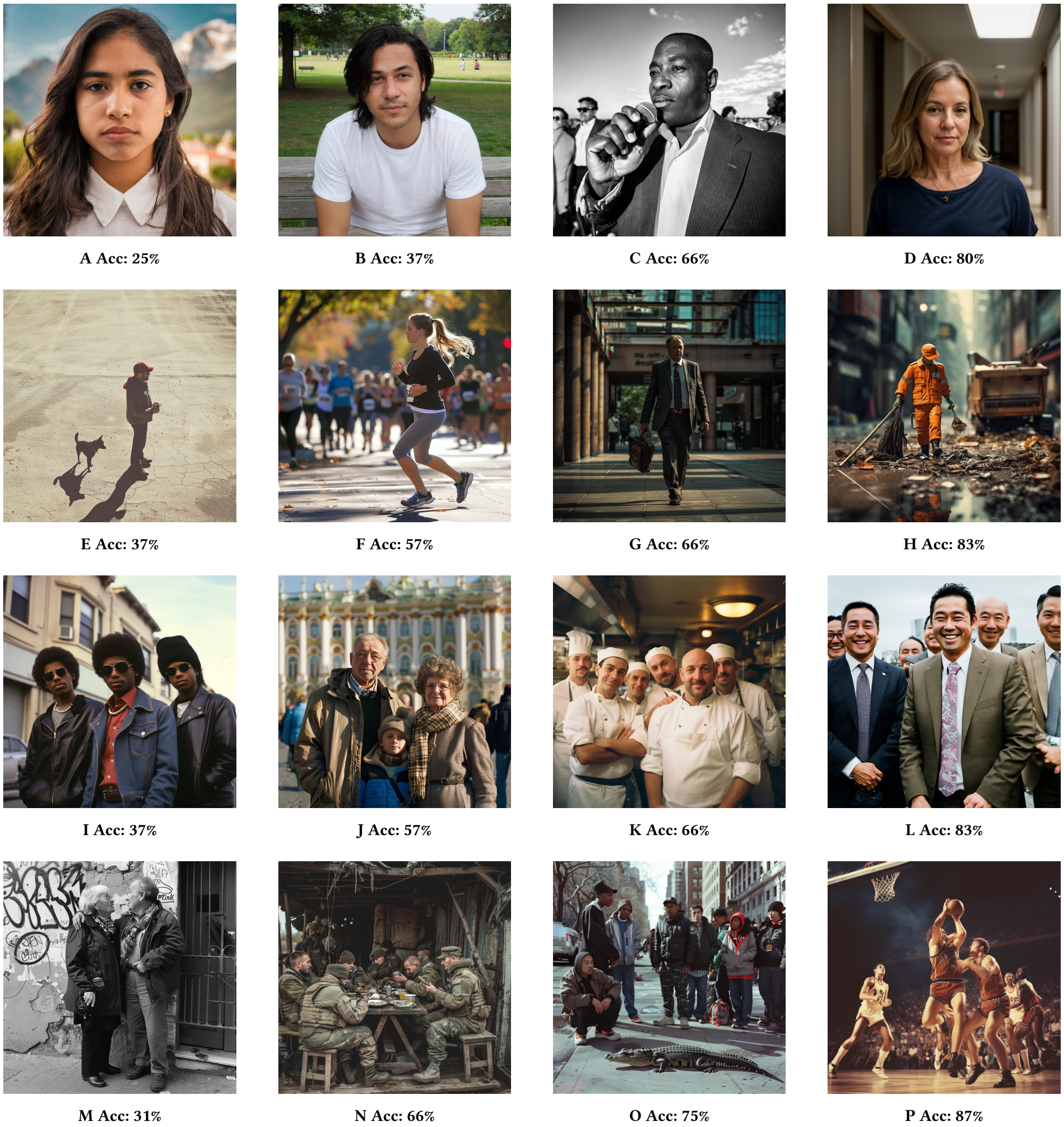


Figure S2: Image generation process in Stable Diffusion A. Four stage image generation pipeline where the image is first generated in SD1.5. The output image is then encoded as latent and upscaled to be re-generated in SDXL with ControlNets applied for pose consistency. This is passed to the face refiner [49] which detects dominant and background faces in the image via YOLOv8 [90] and re-generates them using an SDXL pipeline. Finally, the resulting image is passed to the hand refiner [49] which detects hands in the image via YOLOv8 and predicts the hand pose used to guide the re-generation of the hands. B. Faces in the image before and after the face refining process C. Hand refining process. The left image shows the initial generation of the hand. The center image shows a predicted skeleton for the hand that is used for a ControlNet that guides the re-generation of the hand shown in the image on the right.

According to a New York Times (NYT) quiz, qualities that typically signify AI generation include missing fingers, misaligned eyes, repeated elements, and garbled or nonsensical details [76]. Examples are shown in S1. The NYT quiz also discusses qualities that may cause a real image to look AI-generated, such as repeated cropping and compression that often happens over social media.

A screenshot of the pipeline, along with images before and after refinement, is shown in Figure S2.

Figure S3 displays more examples of the four pose complexities and their average accuracies.



A Acc: 25% **B Acc: 37%** **C Acc: 66%** **D Acc: 80%**

E Acc: 37% **F Acc: 57%** **G Acc: 66%** **H Acc: 83%**

I Acc: 37% **J Acc: 57%** **K Acc: 66%** **L Acc: 83%**

M Acc: 31% **N Acc: 66%** **O Acc: 75%** **P Acc: 87%**

Figure S3: More examples of the four pose complexities and their average accuracies. The first row shows Portraits, the second row Full Body images, the third row Posed Groups, and the last row Candid Groups.

A.1 Robustness Check: Dataset Comparison

To ensure the validity of our conclusions, we conducted a robustness check comparing the results from our full dataset against a subset excluding data collected before May 10th, 2024. This comparison addresses potential biases introduced by the initial experimental design, which did not implement stratified randomization as mentioned in Section 3.3.2.

Table S1 presents the accuracy metrics for both the full dataset and the dataset excluding pre-May 10th data. The table includes overall accuracy, as well as specific accuracy for AI-generated and real images, along with their respective 95% confidence intervals.

Table S1: Comparison of accuracy: Full Dataset vs. Dataset excluding data before May 10th

Dataset	Overall		AI-generated		Real	
	Accuracy	95% CI	Accuracy	95% CI	Accuracy	95% CI
Full Dataset	0.75	[0.74, 0.76]	0.76	[0.74, 0.77]	0.73	[0.71, 0.75]
Dataset excluding data before May 10th	0.75	[0.74, 0.76]	0.76	[0.75, 0.77]	0.7201	[0.70, 0.74]

Figure S4 visualizes the distribution of image accuracies for both datasets. This comparison allows for direct observation of any potential shifts in accuracy distributions between the full dataset and the subset, excluding early data. This robustness check supports the validity of using the full dataset in our main analysis.

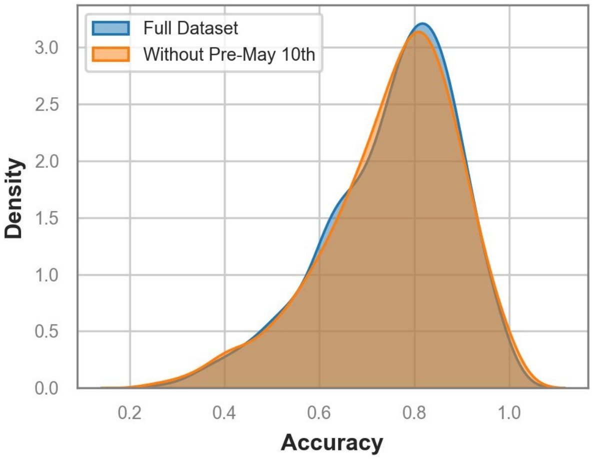


Figure S4: Comparison of accuracies between the full dataset and the dataset excluding data before 10th data.

B Curated and Uncurated AI-generated Images

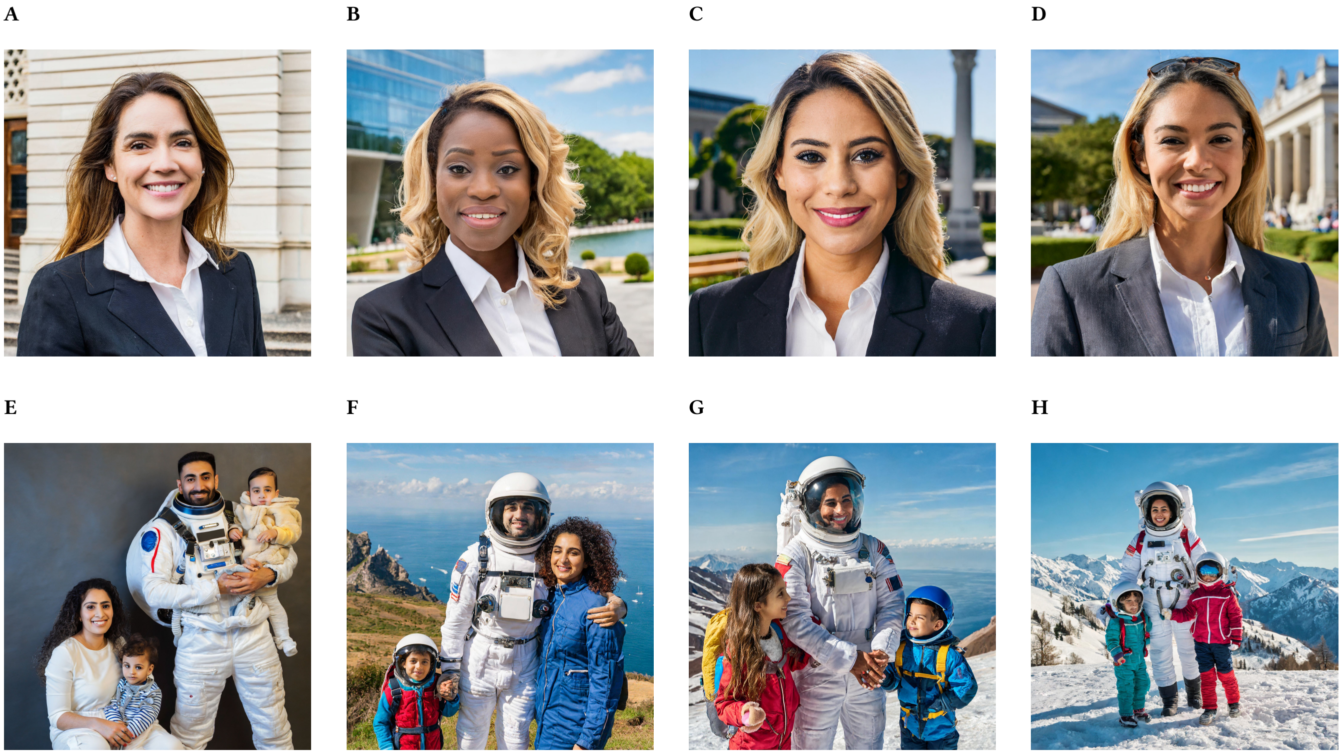


Figure S5: Example images generated by consistently photorealistic and consistently detectable prompts. A. Curated image generated with a consistently photorealistic prompt: "American woman faculty portrait, not a close-up, blond." **B-D** Reprompted images generated with the same consistently photorealistic prompts. **E.** Curated image generated with a consistently detectable prompt: "Persian woman astronaut in astronaut clothes, family photo with husband and two toddlers, high resolution, realistic." **F-H** Reprompted images of the same consistently detectable prompts.

C Future Work on Videos

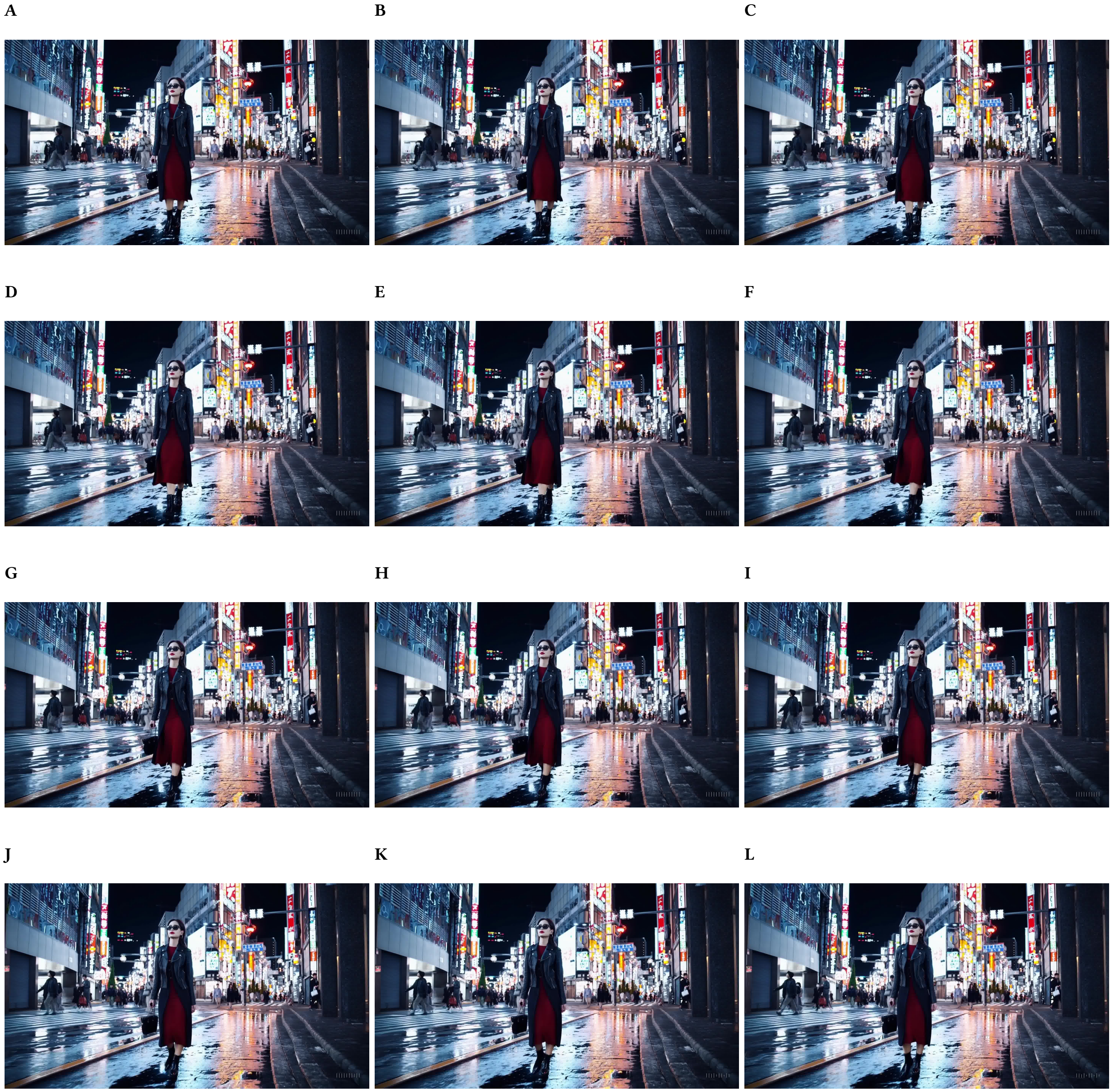


Figure S6: Example frames from an AI-generated video with a temporal anatomical implausibility. 9 frames from a video generated by OpenAI's Sora diffusion-transformer model where the subject's right leg morphs into the left leg somewhere between E and J. Each frame is separated by 1/10 of a second. This particular artifact fits into the anatomical implausibility category of the taxonomy, but it's different from any anatomical plausibility seen in diffusion model-generated images. In particular, this implausibility has a temporal element: the transition from A to L involves the subject's right leg becoming her left in a split second, which does not fit with what we know about human anatomy.