The Visual Counter Turing Test (VCT^2): A Benchmark for Evaluating AI-Generated Image Detection and the Visual AI Index (V$_{\text{AI}}$)

Nasrin Imanpour[1,*], Abhilekh Borah[2,*], Shashwat Bajpai[3,*], Subhankar Ghosh[4,*], Sainath Reddy Sankepally[5,*], Hasnat Md Abdullah[6], Nishoak Kosaraju[7], Shreyas Dixit[8], Ashhar Aziz[9], Shwetangshu Biswas[10], Vinija Jain[11], Aman Chadha[12,13†], Song Wang[14], Amit Sheth[1], Amitava Das[15]

[1]University of South Carolina, USA, [2]Manipal University Jaipur, India, [3]BITS Pilani, Hyderabad, India, [4]Xpectrum AI, USA, [5]International Institute of Information Technology, India, [6]Texas A&M University, USA, [7]Carnegie Mellon University, USA, [8]Vishwakarma Institute of Information Technology, India, [9]IIIT Delhi, India, [10]National Institute of Technology, Silchar, India, [11]Amazon AI, USA, [12]Stanford University, USA, [13]Amazon GenAI, USA, [14]Shenzhen University of Advanced Technology, China, [15]BITS Pilani, Goa, India

Abstract

The rapid progress and widespread availability of text-to-image (T2I) generative models have heightened concerns about the misuse of AI-generated visuals, particularly in the context of misinformation campaigns. Existing AI-generated image detection (AGID) methods often overfit to known generators and falter on outputs from newer or unseen models. We introduce the **Visual Counter Turing Test (VCT^2)**, a comprehensive benchmark of 166,000 images, comprising both real and synthetic prompt-image pairs produced by six state-of-the-art T2I systems: Stable Diffusion 2.1, SDXL, SD3 Medium, SD3.5 Large, DALL·E 3, and Midjourney 6. We curate two distinct subsets: *COCO$_{AI}$*, featuring structured captions from MS COCO, and *Twitter$_{AI}$*, containing narrative-style tweets from The New York Times. Under a unified zero-shot evaluation, we benchmark 17 leading AGID models and observe alarmingly low detection accuracy, 58% on COCO$_{\text{AI}}$ and 58.34% on Twitter$_{\text{AI}}$. To transcend binary classification, we propose the **Visual AI Index (V$_{\text{AI}}$)**, an interpretable, prompt-agnostic realism metric based on twelve low-level visual features, enabling us to quantify and rank the perceptual quality of generated outputs with greater nuance. Correlation analysis reveals a moderate inverse relationship between V$_{\text{AI}}$ and detection accuracy: Pearson ρ of -0.532 on COCO$_{\text{AI}}$ and ρ of -0.503 on Twitter$_{\text{AI}}$, suggesting that more visually realistic images tend to be harder to detect, a trend observed consistently across generators. We release COCO$_{\text{AI}}$, Twitter$_{\text{AI}}$, and all codes to catalyze future advances in generalized AGID and perceptual realism assessment.

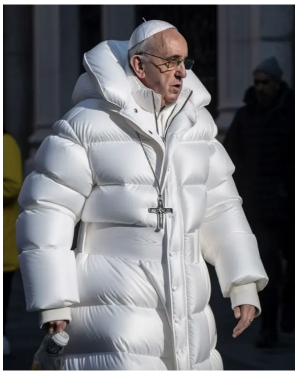


Figure 1: An AI-generated image of Pope Francis wearing a gigantic white puffer jacket went viral on social media platforms like Reddit and Twitter (X) in March 2023. This image sparked widespread media discussions on the potential misuse of generative AI technologies, becoming an iconic example of AI-generated misinformation. For more details, see the Forbes story.

1 Introduction

The rapid advancement of text-to-image (T2I) generative models, such as Stable Diffusion (Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024), DALL·E (Ramesh et al., 2021, 2022; Betker et al., 2023), Midjourney (Midjourney, 2024), and Imagen (Saharia et al., 2022), has revolutionized visual content creation. These models unlock powerful creative workflows and democratize image synthesis at scale. However, their widespread accessibility also raises critical concerns about visual misinformation and content authenticity. As illustrated in Figure 1, synthetic images can convincingly mimic journalistic or photographic style, blurring the boundary between real and generated content. This growing threat has prompted global attention. In March 2023, an open letter (Marcus, 2023) warned that generative AI could destabilize the global information ecosystem. The European

[*]These authors contributed equally to this work.
[†]Work does not relate to position at Amazon.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1847–1862
December 20-24, 2025 ©2025 Association for Computational Linguistics

Commission reported a significant decline in online content moderation accuracy, from 90.4% in 2020 to just 64.4% in 2022 (Commission, 2022). Meanwhile, social platforms process over 3.2 billion images and 720,000 hours of video daily (T.J. Thomson, 2020), with synthetic media projected to account for 90% of online content by 2026 (Europol, 2024).

Despite increasing demand for reliable detection tools, existing AI-generated image detection (AGID) methods often fail to generalize to images from unseen generators or real-world contexts. Watermark-based approaches remain fragile, easily circumvented via cropping, filtering, or adversarial manipulation (Zhao et al., 2025a). Meanwhile, prior AGID benchmarks (Zhu et al., 2023; Sha et al., 2023) suffer from limited real-image diversity, narrow prompt coverage, outdated model inclusion, and closed access, impeding rigorous evaluation and progress.

To address these limitations, we introduce the **Visual Counter Turing Test (VCT^2)**, a large-scale benchmark dataset for zero-shot AGID evaluation. VCT^2 contains approximately 166,000 images, including 26,000 real prompt-image pairs and 140,000 synthetic images produced by six state of the art T2I models: Stable Diffusion 2.1, SDXL, SD3 Medium, SD3.5 Large, DALL·E 3, and Midjourney 6, spanning both open-source and proprietary systems. The prompts in VCT^2 are drawn from two semantically distinct sources to capture both structured and open-ended language. The COCO$_{\mathrm{AI}}$ subset uses object-centric captions from MS COCO (Lin et al., 2014), a staple in vision-language research. The Twitter$_{\mathrm{AI}}$ subset comprises narrative-style tweets authored by The New York Times (`@nytimes`), providing real-world, journalistic prompts rich in nuance and context. This diversity allows us to evaluate AGID methods across a wide range of generation styles and domains.

To enable more nuanced evaluation beyond binary classification, we introduce the **Visual AI Index (V$_{\mathrm{AI}}$)**. This model-agnostic, interpretable metric quantifies the perceptual realism of an image based solely on its visual content. V$_{\mathrm{AI}}$ produces a scalar score derived from twelve handcrafted, low-level image features, including texture complexity, frequency-domain statistics, Haralick features, and image sharpness. These features have been selected based on their empirically observed alignment with human judgments of realism (Wang et al., 2004; Haralick et al., 2007; Canny, 2009; Corvi et al., 2023a). Correlation analysis further supports the utility of V$_{\mathrm{AI}}$ as a proxy for detection difficulty: we observe a moderate inverse relationship between V$_{\mathrm{AI}}$ scores and AGID detection accuracy across generative models (Pearson $\rho = -0.503$ on Twitter$_{\mathrm{AI}}$ and $\rho = -0.532$ on COCO$_{\mathrm{AI}}$), indicating that more visually realistic images tend to be harder to detect. Our realism scores offer a prompt and model-agnostic lens into the perceptual quality of generated images.

We evaluate **17 AGID** methods under a standardized zero-shot setting, using publicly available implementations and default model checkpoints. Our goal is to assess how well these methods generalize across a variety of text-to-image models, including open-source systems like Stable Diffusion 2.1, SDXL, SD3 Medium, and SD3.5 Large, as well as proprietary models such as DALL·E 3 and Midjourney 6, and across two domains: structured and high quality MS COCO captions and images and narrative-style tweets and real world images from The New York Times. Experimental results (Section 4) reveal model generalization gaps by noticeable detection performance degradation, with average detection accuracy of 58% on COCO$_{\mathrm{AI}}$ and 58.34% on Twitter$_{\mathrm{AI}}$. We observe lower detection accuracy on COCO$_{\mathrm{AI}}$ compared to Twitter$_{\mathrm{AI}}$. This is likely because COCO prompts produce images that are more photo-realistic and visually similar to real photos. In contrast, Twitter$_{\mathrm{AI}}$ generations often include creative or unusual visual patterns, leading to more detectable differences. Notably, DALL·E 3 and SD3.5 consistently yield the lowest detection accuracy across both domains. To summarize, our main contributions are:

(i) We introduce the Visual Counter Turing Test (VCT^2) benchmark to evaluate the generalization capabilities of AI-generated image detection methods across diverse prompt styles and real image sources, including MS COCO and Twitter, as well as six state-of-the-art synthetic image generators.

(ii) We define the VisualAI Index (V$_{\mathrm{AI}}$), a scalar metric to quantify perceptual realism based on twelve interpretable low-level visual features.

2 Recent Advances in AI-Generated Image Detection Techniques

AI-generated image detection (AGID) is becoming increasingly vital as synthetic content continues to grow in both photorealism and scale. Detection methods vary widely in their design assumptions, feature representations, and robustness to distribution shifts, such as changes in generative models, prompt styles, or real image characteristics that diverge from photographic norms.

To facilitate systematic evaluation, we categorize AGID approaches into three broad groups:

(i) Generation Artifact-Based Methods: These methods target low-level signals introduced during the image synthesis process, such as upsampling artifacts, denoising residuals, or color inconsistencies. While often computationally efficient, they tend to be fragile under post-processing or model variation.

(ii) Feature Representation-Based Methods: These approaches rely on high-level semantic or perceptual features extracted using convolutional neural networks (CNNs), Vision Transformers (ViTs), or CLIP-style encoders, computational models that have demonstrated strong performance across a wide range of vision tasks (Imanpour et al., 2021; Bagheri Rajeoni et al., 2023; Rajeoni et al., 2024; Cai et al., 2024; Zhao et al., 2025b). Such methods generally exhibit robust cross-domain generalization capabilities; however, they may struggle to capture subtle, fine-grained artifacts often present in generative content.

(iii) Hybrid Methods: These approaches combine both low-level artifact features (e.g., frequency, texture, or pixel-level traces) and high-level semantic representations (e.g., CNN, ViT, or CLIP embeddings). They often employ contrastive learning, multi-modal embeddings, or text–image alignment to integrate these complementary cues and enhance robustness under distributional shifts between real and synthetic data

Figure 2 illustrates this taxonomy. We evaluate 17 publicly available AGID models spanning all three categories, selected based on their methodological diversity, recent relevance, and open-source availability. Each method is tested in a standardized zero-shot setting using its default checkpoint, without any fine-tuning on the VCT^2 benchmark. This taxonomy provides a reference framework for interpreting detection trends discussed in Section 4, with further implementation details outlined in Appendix A.

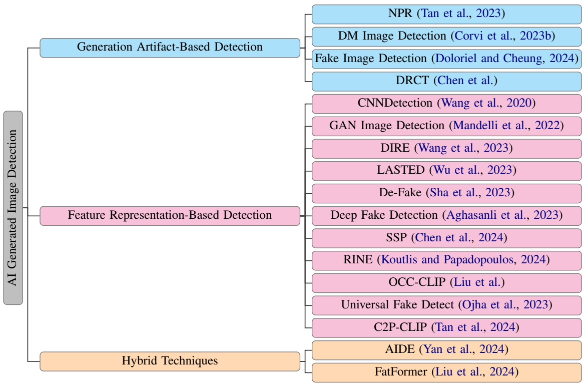

AI Generated Image Detection
Generation Artifact-Based Detection
Feature Representation-Based Detection
Hybrid Techniques
NPR (Tan et al., 2023)
DM Image Detection (Corvi et al., 2023b)
Fake Image Detection (Doloriel and Cheung, 2024)
DRCT (Chen et al.)
CNNDetection (Wang et al., 2020)
GAN Image Detection (Mandelli et al., 2022)
DIRE (Wang et al., 2023)
LASTED (Wu et al., 2023)
De-Fake (Sha et al., 2023)
Deep Fake Detection (Aghasanli et al., 2023)
SSP (Chen et al., 2024)
RINE (Koutlis and Papadopoulos, 2024)
OCC-CLIP (Liu et al.)
Universal Fake Detect (Ojha et al., 2023)
C2P-CLIP (Tan et al., 2024)
AIDE (Yan et al., 2024)
FatFormer (Liu et al., 2024)

Figure 2: The taxonomy of AI-generated image detection techniques, categorized into three main groups: Generation Artifact-Based Detection, Feature Representation-Based Detection, and Hybrid Techniques.

3 The Visual Counter Turing Test (VCT^2) Benchmark Dataset

We introduce the Visual Counter Turing Test (VCT^2), a large-scale benchmark designed to evaluate AI-generated image detection (AGID) techniques. VCT^2 includes the following:

- Approximately 26,000 real image–prompt pairs, combining our curated Twitter dataset and the benchmark MS COCO dataset;
- Approximately 140,000 synthetic images, generated using six state-of-the-art text-to-image models; open-source models: Stable Diffusion 2.1, SDXL, SD3 Medium, SD3.5 Large; and two proprietary models: DALL·E 3, Midjourney 6.
- In total, around 166,000 images derived from 26,000 unique prompts.

This scale provides a balance between structured, caption-based content and naturalistic, real-world prompts, positioning VCT^2 among the most comprehensive AGID datasets to date. All real images and prompts are organized into two structured subsets: Twitter$_{\text{AI}}$[*] and COCO$_{\text{AI}}$[†], which are publicly released.

3.1 Prompt Sources and Coverage

To ensure diversity in both semantic content and visual generation styles, we curated prompts from two distinct and complementary sources:

[*] https://huggingface.co/datasets/NasrinImp/Twitter_AI

[†] https://huggingface.co/datasets/NasrinImp/COCO_AI

Table 1: Topic Clusters in the NYT Twitter Subset.

Topic Cluster	Tweet Count	Top Keywords
Daily Briefings and News Summaries	1129	*know, need, day, morning, briefing, evening*
New York City and Culture	1961	*new, york, city, times, books, critics*
Art, Movies, and Obituaries	631	*photo, review, obituary, art, movie, critic*
Health, COVID-19, and Breaking News	1436	*coronavirus, health, opinion, news, breaking, people*
Opinion Pieces and Societal Reflections	914	*nytopinion, life, young, america, death, ebola*
Travel and International Destinations	605	*hours, italy, florida, china, japan, park*
World Events and Sports	903	*world, cup, photos, team, war, country*
Lifestyle and City Aesthetics	1164	*like, looks, look, city, love, idea*
Time, Life Stories, and Incarceration	1200	*years, life, ago, prison, time, close*
Home, Food, and Leisure	965	*make, recipes, summer, home, simple, best*
miscellaneous	5001	–

- $\sim$10,000 benchmark prompts from the MS COCO dataset (Lin et al., 2014), focused on object-centric and everyday scenes;

- $\sim$16,000 real-world prompts from the `@nytimes` Twitter account 2011–2023. To assess topical diversity, we identified ten topics and associated keywords, and then assigned each tweet to its most probable topic. Table 1 shows the ten dominant clusters. These clusters reflect both editorial depth and real-world content breadth. Their presence enhances the semantic realism of our benchmark and supports rigorous AGID evaluation across multiple domains.

3.2 Real Twitter Prompt-Image Dataset Collection and Processing

To construct a diverse and reliable dataset of real Twitter images, we employed an automated data collection pipeline using Python and Selenium. We focused on tweets from `@nytimes` (The New York Times) due to its editorial credibility, rigorous fact-checking, and diverse topical coverage.

Data Collection. Our pipeline sampled tweets spanning a 12-year period (2011-2023), retaining only those with attached media. The goal was to align real images with captions that could feasibly be used to generate synthetic counterparts.

Definition of Real Images. We define "real" images as those not generated by AI. This includes natural photographs as well as editorial media such as UI screenshots, infographics, and photojournalistic illustrations, provided they are not produced using generative models. This definition reflects the ambiguity present in real-world detection scenarios, where non-photographic content may still be authentic.

Data Filtration and Preprocessing. To ensure quality and consistency, we applied several filtering steps: (i) removal of duplicate tweets and media; (ii) exclusion of irrelevant content such as word games or puzzles; and (iii) filtering of non-English tweets. Additionally, preprocessing involved removal of hashtags and URLs, and retention of only alphanumeric characters to facilitate downstream analysis and clustering.

3.3 Benchmark Contributions

VCT^2 offers several key advantages over existing benchmarks:

- **Scale**. VCT^2 contains 166,000 images generated from 26,000 unique prompts.

- **Model diversity.** VCT^2 includes six cutting-edge generative models supporting broader evaluation across current-generation image generators.

- **Prompt realism.** VCT^2 uniquely combines benchmark-style prompts from MS COCO with naturalistic, real-world prompts curated from a 12-year archive of `@nytimes` tweets, capturing diverse linguistic styles and topics.

- **Mixed-media realism.** The real image subset includes ambiguous formats such as infographics, UI screenshots, and editorial photos, reflecting the heterogeneous content encountered in real-world detection scenarios.

- **Public accessibility.** All prompts, real and synthetic images, and evaluation scripts for 17 AGID baselines are publicly released to facilitate reproducibility and comparative benchmarking.

To our knowledge, VCT^2 is the first large-scale AGID benchmark to pair real-world journalistic prompts with diverse state-of-the-art text-to-image models, providing a robust and publicly available *testbed for evaluating both detection performance and perceptual realism across different prompt domains and generative model types.*

4 Evaluation and Results

We evaluate the VCT^2 benchmark under zero-shot settings using 17 state-of-the-art AI-generated image detection (AGID) methods. These span artifact-, feature-, and hybrid-based approaches. In the following, we present detection performance, examine cross-domain and cross-model generalization, and analyze detector sensitivity to detector type.

4.1 Evaluation Protocol

To simulate real-world deployment, we assess all detectors without fine-tuning. Public checkpoints and default hyperparameters are used. Performance is measured separately on COCO$_{\text{AI}}$ and Twitter$_{\text{AI}}$ subsets, reporting accuracy, precision, and recall. Results are summarized in Tables 2 and 3.

4.2 Cross-Domain and -Model Trends

Figure 3 presents the average detection accuracies per generator across the two domains. Overall, detection performance is low, with accuracy dropping further on COCO$_{\text{AI}}$ compared to Twitter$_{\text{AI}}$.

Detection performance also varies across generators. Images from earlier models like SD2.1 and SDXL remain relatively detectable. In contrast, newer or proprietary models such as SD3.5 Large and DALL·E 3 yield significantly lower detection results, suggesting that existing detectors may be overfitted to older, synthetic image distributions.

4.3 Comparative Detector Performance

We analyze per detector performance in Appendix B. The results indicate that there is no one-size-fits-all solution for detecting AI-generated images. Different generative models pose unique challenges, and the performance of each detection method varies based on its ability to identify specific artifacts. De-Fake and DRCT were the most consistent performers, highlighting their robustness across models. The latter collapse on proprietary models due to reliance on low-level artifacts often absent in advanced generators. Conversely, feature-based and contrastive methods benefit from semantic representations, allowing stronger generalization to unseen prompt styles and model outputs. Future research should aim to improve detection for models like SD 3.5 Large, Midjourney 6 and DALL.E 3, where many techniques struggled.

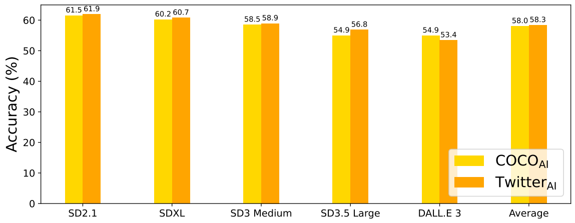


Figure 3: Average detection accuracy across the COCO$_{\text{AI}}$ and Twitter$_{\text{AI}}$ subsets for each generator.

5 The Visual AI Index (V$_{\text{AI}}$)

We introduce the *Visual AI Index* (V$_{\text{AI}}$), an interpretable, prompt-agnostic metric that scores the perceptual realism of images based on low-level visual features. V$_{\text{AI}}$ provides a continuous score that reflects where an image lies along a spectrum of visual realism. Many real images in web-scale datasets (e.g., news media, social platforms) are not pristine photographs; they may include screenshots, digital graphics, or compressed visuals. These images often lack sharpness, contrast, or structure. V$_{\text{AI}}$ quantifies perceptual quality by learning to

Table 2: Overall accuracy (Acc), recall (R), and precision (P) across COCO$_{\text{AI}}$ synthetic datasets generated from MS COCO prompts. All values are in %. Color-coded: Green ($\geq$ 90%), Yellow-Green (80–89%), Yellow (70–79%), Orange (60–69%), red (<60%).

Method	SD2.1			SDXL			SD3 Medium			SD3.5 Large			DALL.E 3			Midjourney 6		
	Acc	R	P	Acc	R	P	Acc	R	P	Acc	R	P	Acc	R	P	Acc	R	P
CNNDetection (Wang et al., 2020)	49.94	0.03	65.11	49.96	0.07	77.52	49.93	0.01	81.16	49.99	0.14	33.04	49.93	0.00	35.13	49.95	0.05	63.15
NPR (Tan et al., 2023)	26.76	1.89	34.26	26.68	1.73	33.15	27.96	4.29	34.41	70.32	48.37	79.44	25.81	0.00	41.13	25.81	0.00	48.13
DM Image Detection (Corvi et al., 2023b)	83.92	67.92	99.40	69.96	40.00	98.91	63.58	27.23	98.04	38.58	32.06	0.07	49.96	0.00	40.00	51.73	3.52	87.04
Fake Image Detection (Doloriel and Cheung, 2024)	49.84	0.49	63.58	49.83	0.48	66.68	50.02	0.86	66.91	48.24	0.11	62.57	49.59	0.00	34.90	49.79	0.40	62.89
DIRE (Wang et al., 2023)	47.08	93.40	37.66	49.67	98.57	47.07	48.59	96.40	38.88	50.63	99.23	58.68	48.89	97.01	43.25	50.04	99.31	52.74
LASTED (Wu et al., 2023)	54.00	8.67	56.62	61.13	9.86	61.20	51.87	9.61	57.67	55.21	10.11	57.35	66.18	44.85	76.21	68.21	14.37	63.14
GAN Image Detection (Mandelli et al., 2022)	51.87	82.93	51.16	56.35	91.75	53.72	58.26	95.35	54.74	45.61	79.08	47.37	48.10	74.93	48.77	57.15	93.42	54.14
AIDE (Yan et al., 2024)	60.30	20.98	93.77	64.34	28.91	96.75	57.11	14.45	94.28	50.83	5.01	52.31	50.00	0.02	61.23	76.01	52.25	96.92
SSP (Chen et al., 2024)	50.15	99.63	50.07	49.95	99.63	49.97	50.34	99.63	50.17	50.30	99.48	50.29	49.91	99.63	49.95	49.95	99.63	49.97
FatFormer (Liu et al., 2024)	50.00	0.00	0.00	50.00	0.00	0.00	50.00	0.01	100	50.28	0.00	0.00	48.01	0.00	0.00	48.01	0.00	0.00
DRCT (ConvB) (Chen et al.)	98.76	99.61	97.94	96.83	95.75	97.86	80.72	63.54	96.81	78.58	59.05	96.51	49.99	2.08	49.76	67.48	37.06	94.65
DRCT (UnivFD) (Chen et al.)	88.57	96.98	83.02	89.45	98.73	83.27	84.90	89.64	81.88	83.09	84.09	82.29	79.98	79.80	80.09	89.64	99.12	83.32
RINE (Koutlis and Papadopoulos, 2024)	74.43	49.63	98.49	56.47	13.71	94.76	61.99	24.75	97.03	55.34	27.72	95.34	50.05	0.87	53.37	63.13	27.02	97.27
OCC-CLIP (Liu et al.)	51.49	92.28	50.82	47.11	14.95	41.91	50.60	66.03	50.46	49.08	50.67	67.63	78.82	50.28	55.04	75.04	53.60	50.03
De-Fake (Sha et al., 2023)	92.37	97.90	88.15	91.23	95.62	87.90	91.30	95.76	87.92	52.57	86.05	5.11	90.58	94.31	87.76	86.22	85.59	86.68
Deep Fake Detection (Aghasanli et al., 2023)	49.49	49.49	49.03	51.43	51.43	49.65	49.85	49.85	49.97	50.66	50.66	52.19	52.73	52.73	53.02	52.87	52.87	54.09
Universal Fake Detect (Ojha et al., 2023)	74.42	77.15	73.15	69.18	65.84	70.56	70.11	68.79	70.66	57.46	60.10	57.09	50.00	99.99	50.00	53.23	76.28	52.21
C2P-CLIP (Tan et al., 2024)	53.38	7.53	90.73	53.69	8.15	91.30	55.50	11.77	93.87	52.13	4.40	97.01	49.76	0.29	27.36	50.31	1.40	64.52

Table 3: Overall accuracy (Acc), recall (R), and precision (P) across Twitter$_{\text{AI}}$ synthetic datasets generated from Twitter prompts. Midjourney 6 is not included as it blocks image generation for most Twitter prompts. All values are in %. Color-coded: Green ($\geq$ 90%), Yellow-Green (80–89%), Yellow (70–79%), Orange (60–69%), red (<60%).

Method	SD2.1			SDXL			SD3 Medium			SD3.5 Large			DALL.E 3		
	Acc	R	P	Acc	R	P	Acc	R	P	Acc	R	P	Acc	R	P
CNNDetection (Wang et al., 2020)	50.00	0.06	52.21	49.98	0.03	59.98	50.19	0.44	74.35	50.34	0.76	76.04	49.97	0.01	34.59
NPR (Tan et al., 2023)	50.23	2.22	50.89	50.46	2.68	60.58	51.45	4.66	68.26	52.12	7.81	67.44	49.12	0.00	42.20
DM Image Detection (Corvi et al., 2023b)	88.31	77.57	97.82	73.82	48.58	93.74	65.15	31.24	90.34	63.56	28.19	89.66	49.53	0.00	33.34
Fake Image Detection (Doloriel and Cheung, 2024)	49.86	0.53	56.33	49.88	0.58	60.83	50.35	1.51	66.15	48.57	1.12	62.13	49.59	0.01	33.11
DIRE (Wang et al., 2023)	43.90	86.95	36.20	48.57	96.29	46.29	48.49	96.13	38.48	49.41	98.01	45.11	46.33	91.81	36.19
LASTED (Wu et al., 2023)	77.60	1.93	59.60	83.60	2.75	66.04	83.24	2.75	61.52	82.71	2.59	61.90	78.77	25.57	76.81
GAN Image Detection (Mandelli et al., 2022)	53.26	77.37	52.25	55.84	82.36	53.86	60.01	91.04	56.21	54.78	80.60	53.19	53.99	79.44	52.68
AIDE (Yan et al., 2024)	55.69	11.81	81.98	60.43	21.29	89.61	56.49	13.41	87.40	56.57	6.16	58.42	49.93	0.25	43.61
SSP (Chen et al., 2024)	49.91	99.66	49.95	50.20	99.66	50.10	50.20	99.66	50.10	54.94	99.33	55.04	50.18	99.66	50.10
FatFormer (Liu et al., 2024)	50.04	0.08	100	50.04	0.08	100	50.00	0.00	0.00	55.10	0.02	100	50.02	0.00	0.00
DRCT (ConvB) (Chen et al.)	96.81	99.77	94.20	93.96	94.05	93.87	71.79	49.73	89.01	77.87	59.54	90.35	47.31	0.76	11.02
DRCT (UnivFD) (Chen et al.)	67.47	96.73	61.02	68.32	98.43	61.44	64.81	91.40	59.67	62.94	85.60	55.68	53.76	69.30	52.87
RINE (Koutlis and Papadopoulos, 2024)	77.07	55.40	97.79	57.86	16.97	93.13	62.13	25.50	95.32	66.36	44.37	94.36	49.61	0.48	27.64
OCC-CLIP (Liu et al.)	46.88	74.11	47.98	45.67	51.17	46.10	48.84	67.54	49.16	47.82	66.16	48.03	47.75	45.63	49.72
De-Fake (Sha et al., 2023)	81.13	91.51	75.78	78.16	85.57	74.53	79.39	88.03	72.06	40.80	0.00	0.00	79.95	89.14	75.29
Deep Fake Detection (Aghasanli et al., 2023)	50.80	50.80	51.84	53.64	53.64	56.59	51.44	51.44	51.51	49.19	49.19	56.38	55.30	55.30	60.34
Universal Fake Detect (Ojha et al., 2023)	72.88	74.17	72.31	69.47	73.91	67.89	68.41	72.58	67.00	55.38	45.60	56.69	50.00	99.99	50.00
C2P-CLIP (Tan et al., 2024)	52.21	6.74	88.76	53.28	7.98	91.25	47.92	0.97	26.17	54.16	8.47	98.39	49.73	0.21	53.45

score realism using a combination of handcrafted visual features, independent of prompts or model-specific information.

5.1 Feature Design

V$_{\text{AI}}$ uses twelve visual features grouped into three categories:

(i) Texture and Frequency: Texture Complexity, Haralick Contrast, Haralick Correlation, Haralick Energy, Frequency Mean, Frequency Standard Deviation.

(ii) Sharpness and Structure: Image Sharpness, Image Smoothness, Image Contrast.

(iii) Color and Semantics: Color Distribution Consistency, Object Coherence, Contextual Relevance.

Texture Complexity quantifies the variety and unpredictability of an image's texture. It is determined by computing the entropy of the normalized Local Binary Pattern (LBP) histogram of the grayscale image using the formula $-\sum_{k=0}^{P-1} \tilde{H}_{LBP}(k) \log_2(\tilde{H}_{LBP}(k) + \epsilon)$. Here, $\tilde{H}_{LBP}(k)$ represents the normalized histogram value for LBP bin k, and P is the total number of bins in the LBP histogram. The small constant ϵ (in our case, 1×10^{-6}) is used to avoid taking the logarithm of zero.

Haralick features are texture descriptors computed from the gray-level co-occurrence matrix (GLCM), which encodes the frequency $G(i, j)$ of pixel intensity pairs (i, j) occurring at a fixed spatial offset. We use three common features:

Haralick Contrast is defined as $\sum_{i,j}(i - j)^2 G(i, j)$, capturing local intensity variation.

Haralick Correlation is computed as $\sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)G(i,j)}{\sigma_i \sigma_j}$, where μ_i, μ_j and σ_i, σ_j are the means and standard deviations of the marginal GLCM distributions. It measures linear dependency between pixel pairs.

Haralick Energy (Angular Second Moment) is given by $\sum_{i,j} G(i, j)^2$, reflecting texture uniformity—higher values imply more homogeneous regions.

These values are averaged across multiple angles (e.g., 0°, 45°, 90°, 135°) to ensure rotation-invariant descriptors.

We extract frequency-domain features using the 2D Fast Fourier Transform (FFT) of the grayscale image I. Let $\hat{I}(u, v) = \text{FFT2}(I)$ denote the Fourier-transformed image, and let $M(u, v) = |\hat{I}(u, v)|$ be the magnitude spectrum.

Frequency Mean is defined as $\text{FreqMean} = \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} M(u, v)$, where $H \times W$ is the image resolution.

Frequency Standard Deviation is given by $\text{FreqStd} = \sqrt{\frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} (M(u, v) - \text{FreqMean})^2}$.

These two features capture the spectral energy and its variation. Higher values indicate detailed or noisy content, while lower values reflect smoother textures.

Image Sharpness is quantified as $\max(|I - I_{\text{blurred}}|)$. I and I_{blurred} denote the grayscale and blurred image with Gaussian kernel, respectively.

Image Smoothness evaluates how consistent the image's texture is. It is quantified as $\frac{1}{1+\text{var}(\Delta I)}$, where ΔI denotes the Laplacian of the grayscale image I.

Image Contrast measures the degree of variation in intensity across an image. It is quantified by calculating the standard deviation of the pixel values in the grayscale image, expressed as $\text{std}(I)$.

Color Distribution Consistency evaluates the variability in an image's color distribution by analyzing the standard deviation of the normalized color histogram in the HSV color space. It is calculated as std($\tilde{H}_{HSV}(h, s, v)$), where std($\cdot$) denotes the standard deviation of the normalized histogram $\tilde{H}_{HSV}(h, s, v)$ for hue h, saturation s, and value v.

Object Coherence evaluates the extent and clarity of edge detection in an image, providing insight into the consistency of object boundaries. It is determined using $\frac{\sum_{i,j} E(i,j)}{\sum_{i,j} 1}$, where $E(i, j)$ represents the value of the Canny edge image at pixel (i, j), and the $\sum_{i,j} 1$ represents the total number of pixels in the image.

Contextual Relevance evaluates the distribution of edge strengths across the image. It is given by var($\sqrt{{G_x}^2 + {G_y}^2}$), where var($\cdot$) denotes the variance, and G_x and G_y are the gradients computed using the Sobel filter in the horizontal and vertical directions, respectively.

Each low-level feature is first standardized using Z-score normalization as $f_i(x) = \frac{v_i(x) - \mu_i}{\sigma_i}$, where $v_i(x)$ is the raw value of feature i for image x, and μ_i, σ_i are the mean and standard deviation of that feature across the dataset.

To compute the Visual AI Index, we learn a set of weights that quantify how strongly each normalized feature contributes to the perceived realism of an image. We employ a logistic regression model to distinguish between real (label $y = 1$) and synthetic (label $y = 0$) images. Given a 12-dimensional feature vector $x = [f_1, f_2, \ldots, f_{12}]$, the model estimates the probability that an image is real as $p(x) = \frac{1}{1+e^{-w^\top x}}$, where w is the weight vector. The weights are learned by minimizing the binary cross-entropy loss $\mathcal{L}(w) = \frac{1}{N} \sum_{i=1}^{N} [-y_i \log p(x_i) - (1 - y_i) \log(1 - p(x_i))]$. After optimization, the final Visual AI Index is defined as $\mathrm{V}_{\mathrm{AI}}(x) = p(x; w^*) = \frac{1}{1+e^{-w^{*\top} x}}$, where w^* denotes the optimized weights obtained after training. We train two models separately, one for $\mathrm{COCO}_{\mathrm{AI}}$ and one for $\mathrm{Twitter}_{\mathrm{AI}}$, each tailored to the distribution of real images in its respective domain. Table 4 reports the final learned weights.

Table 4: Learned V_{AI} feature weights for $\mathrm{COCO}_{\mathrm{AI}}$ and $\mathrm{Twitter}_{\mathrm{AI}}$ domains.

Feature	**$\mathrm{COCO}_{\mathrm{AI}}$**	**$\mathrm{Twitter}_{\mathrm{AI}}$**
Texture Complexity	4.13	1.15
Color Dist. Consistency	−0.15	−0.05
Object Coherence	−0.87	1.56
Contextual Relevance	−1.33	2.42
Haralick Contrast	1.02	−4.52
Haralick Correlation	−0.42	−0.07
Haralick Energy	−1.46	−1.59
Freq. Mean	−0.87	−1.36
Freq. Std	−1.30	−0.20
Image Smoothness	0.37	−0.06
Image Sharpness	2.08	2.26
Image Contrast	0.50	−0.18

5.2 V_{AI} Analysis

We report the average V_{AI} scores for real and generated images across the $\mathrm{COCO}_{\mathrm{AI}}$ and $\mathrm{Twitter}_{\mathrm{AI}}$ subsets in Figures 4 and 5. As expected, real images achieve the highest V_{AI} scores in both domains, reflecting the benchmark's ability to assign higher realism to naturally occurring images. Among generative models, DALL·E 3 obtains the highest V_{AI} in both subsets (0.626 for $\mathrm{COCO}_{\mathrm{AI}}$, 0.593 for $\mathrm{Twitter}_{\mathrm{AI}}$), indicating its outputs most closely align with real images in terms of low-level features such as texture complexity, edge coherence, and color consistency. A cluster of diffusion-based models, SD2.1, SD3 Medium, and SD3.5 Large, follow DALL·E 3 with relatively similar V_{AI} scores, suggesting comparable levels of photorealism. SDXL ranks lower across both domains, i.e. 0.496 $\mathrm{COCO}_{\mathrm{AI}}$ and 0.573 $\mathrm{Twitter}_{\mathrm{AI}}$, which may be attributed to its tendency toward stylistic exaggeration or generation artifacts that deviate from natural image statistics. These artifacts can influence frequency-domain, edge-based, or texture descriptors negatively, despite the model's high perceptual fidelity. Midjourney yields the lowest V_{AI} in the $\mathrm{COCO}_{\mathrm{AI}}$ subset (0.432) and is excluded from the $\mathrm{Twitter}_{\mathrm{AI}}$ analysis due to the unavailability of corresponding generated images.

The accuracy heat maps in Figures 4 and 5 highlight differences in AGID methods across models. The results indicates that the texture and artifact characteristics differ significantly across models, affecting detection reliability. While some detection methods, like De-Fake and DRCT, performed consistently well, the V_{AI} scores reveal that realism score of generated image plays a significant role in detection difficulty.

Factor-level analysis is provided in the Appendix C, where we highlight specific contributing factors through LBP (Local Binary Pattern) analysis, and pairwise factor plots.

5.3 Correlation with Detection Accuracy

To evaluate whether the Visual AI Index aligns with the difficulty of detecting AI-generated im-

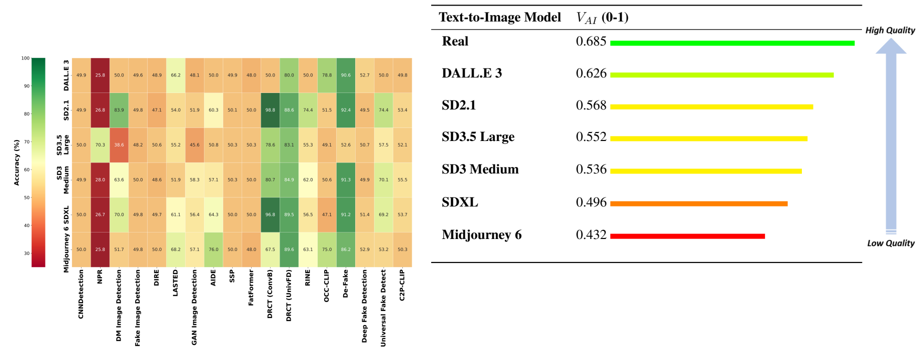


Figure 4: Right: V_{AI} scores of COCO$_{\text{AI}}$ dataset. Left: Accuracy heat maps showing the average accuracy of each AGID method.

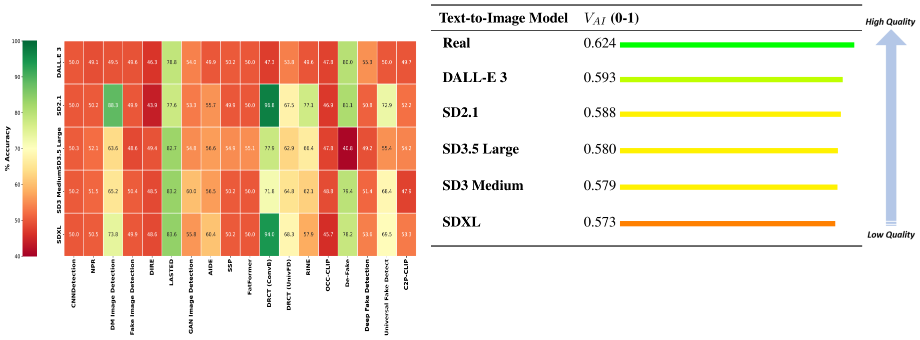


Figure 5: Right: V_{AI} scores of Twitter$_{\text{AI}}$ dataset. Left: Accuracy heat maps showing the average accuracy of each AGID method.

ages, we compute the Pearson correlation coefficient between average V$_{\text{AI}}$ scores and AGID detection accuracy across five generative models. The Pearson correlation coefficient ρ is defined as:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where x_i and y_i represent the V$_{\text{AI}}$ score and AGID accuracy for model i, and $\bar{x}$ and $\bar{y}$ are their respective means. The coefficient ρ ranges from -1 to 1: a value near 1 implies a strong positive correlation, near -1 implies a strong negative correlation, and a value near 0 suggests no linear relationship. We compute ρ separately for the Twitter$_{\text{AI}}$ and COCO$_{\text{AI}}$ datasets. As shown in Table 5, our results indicate a moderate inverse relationship: models with higher visual realism tend to be harder to detect. However, the correlations are not statistically significant, likely due to the small number of generative models, i.e. $n = 5$, and should be interpreted cautiously.

Table 5: Pearson correlation between V$_{\text{AI}}$ and AGID detection accuracy.

Dataset	Pearson ρ	p-value
Twitter$_{\text{AI}}$	-0.503	0.388
COCO$_{\text{AI}}$	-0.532	0.356

6 Conclusion

In this paper, we introduced (VCT^2), a comprehensive benchmark for evaluating AI-generated image detection (AGID) across diverse generative models, including cutting-edge proprietary systems like DALL·E 3 and Midjourney 6. By incorporating both real-world prompts and standardized captions, VCT² offers a challenging, realistic dataset for assessing generalization. The VCT^2 benchmark provides a critical resource for evaluating AGID techniques under challenging and varied conditions, highlighting performance gaps and guiding the development of more robust detection methods.

To assess the realism of images, we present the Visual AI Index (V_{AI}) that evaluates characteristics like texture complexity, Haralick correlation, frequency mean, and image sharpness. Our findings reveal that real images generally achieve higher V_{AI} scores than AI-generated images.

Limitations and Future Work

While our work provides a strong foundation for evaluating AGID methods and realism metrics, future directions include expanding to diverse domains (e.g., social platforms, synthetic video), integrating temporal and multimodal features into V_{AI}, and adapting it for localization or attribution. We also plan to explore human alignment and psychometric grounding of these continuous realism scores. As generative models evolve, updating the benchmark and exploring hybrid detection techniques will be key to ensuring resilience against increasingly sophisticated AI imagery.

References

Agil Aghasanli, Dmitry Kangin, and Plamen Angelov. 2023. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 467–474.

Alireza Bagheri Rajeoni, Breanna Pederson, Daniel G Clair, Susan M Lessner, and Homayoun Valafar. 2023. Automated measurement of vascular calcification in femoral endarterectomy patients using deep learning. *Diagnostics*, 13(21):3363.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Pingping Cai, Canyu Zhang, Lingjia Shi, Lili Wang, Nasrin Imanpour, and Song Wang. 2024. Einet: Point cloud completion via extrapolation and interpolation. In *European Conference on Computer Vision*, pages 377–393. Springer.

John Canny. 2009. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*.

Jiaxuan Chen, Jieteng Yao, and Li Niu. 2024. A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*.

European Commission. 2022. Eu code of conduct against online hate speech: latest evaluation shows slowdown in progress.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023a. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 973–982.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023b. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chandler Timm Doloriel and Ngai-Man Cheung. 2024. Frequency masking for universal deepfake detection. *arXiv preprint arXiv:2401.06506*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Europol. 2024. Facing reality: Law enforcement and the challenge of deepfakes. Accessed: 2024-08-30.

Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. 2007. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.

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

Nasrin Imanpour, Ahmad R Naghsh-Nilchi, Amirhassan Monadjemi, Hossein Karshenas, Kamal Nasrollahi, and Thomas B Moeslund. 2021. Memory-and time-efficient dense network for single-image super-resolution. *IET Signal Processing*, 15(2):141–152.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

Christos Koutlis and Symeon Papadopoulos. 2024. Leveraging representations from intermediate encoder-blocks for synthetic image detection. *arXiv preprint arXiv:2402.19091*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, and Jindong Gu. Which model generated this image? a model-agnostic approach for origin attribution.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. 2024. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. 2022. Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3091–3095. IEEE.

Gary Marcus. 2023. Pause giant ai experiments: An open letter.

Midjourney. 2024. `https://www.midjourney.com/home`.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *CVPR*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Alireza Bagheri Rajeoni, Breanna Pederson, Ali Firooz, Hamed Abdollahi, Andrew K Smith, Daniel G Clair, Susan M Lessner, and Homayoun Valafar. 2024. Vascular system segmentation using deep learning. *Artificial Intelligence: Machine Learning, Convolutional Neural Networks and Large Language Models*, 1:85.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 2023. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. 2024. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. *arXiv preprint arXiv:2408.09647*.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2023. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. *arXiv preprint arXiv:2312.10461*.

Paula Dootson T.J. Thomson, Daniel Angus. 2020. 3.2 billion images and 720,000 hours of video are shared online daily. can you sort real from fake?

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*, pages 8695–8704.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

H. Wu, J. Zhou, and S. Zhang. 2023. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint:2305.13800*.

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 preprint arXiv:2406.19435*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. 2025a. Invisible image watermarks are provably removable using generative ai. *Advances in Neural Information Processing Systems*, 37:8643–8672.

Ziyu Zhao, Xiaoguang Li, Lingjia Shi, Nasrin Imanpour, and Song Wang. 2025b. Dpseg: Dual-prompt cost volume learning for open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25346–25356.

Mingjian Zhu, Hanting Chen, Mouxiao Huang, Wei Li, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. Gendet: Towards good generalizations for ai-generated image detection. *arXiv preprint arXiv:2312.08880*.

Appendix

A Detection Techniques

This appendix provides detailed descriptions of the **17** AI-generated image detection (AGID) techniques evaluated on our benchmark. These methods span three major detection paradigms: generation artifact-based, feature representation-based, and hybrid approaches. This taxonomy is designed to reflect the breadth of design assumptions across the literature and serves as the foundation for our performance analysis in Section 4.

Artifact-based methods exploit low-level visual artifacts, such as frequency distortions, edge inconsistencies, or upsampling traces, introduced during the image generation process. Feature-based methods, in contrast, analyze semantic-level inconsistencies by leveraging deep neural representations from CNNs, vision transformers, or CLIP encoders. Hybrid methods combine both low-level and high-level signals, often incorporating alignment objectives or learned fusion strategies to improve robustness.

The detectors described here were selected based on recency, diversity, and public availability, and represent both classical and state-of-the-art AGID strategies. Each technique is evaluated under zero-shot settings using default public checkpoints, and grouped by detection paradigm below.

A.1 Generation Artifact-Based Detection

Generation artifact-based detection techniques focus on identifying visual artifacts produced during the generation process, analyzing both spatial and frequency domains.

(Tan et al., 2023) found that the up-sampling operator introduces artifacts not only in frequency patterns but also in pixel arrangements within images. The authors introduce the concept of Neighboring Pixel Relationships to capture and characterize these generalized structural artifacts caused by up-sampling operations.

(Corvi et al., 2023b) observed that synthetic images, especially those generated by diffusion models like GLIDE and Stable Diffusion, exhibit distinctive differences in mid-to-high frequency signals compared to real images. However, this distinction is less pronounced in images produced by newer models, such as DALL-E and ADM. Although their method accurately distinguishes synthetic and real images in controlled settings, it struggles in real-world scenarios.

(Doloriel and Cheung, 2024) explored masked image modeling for universal fake image detection. Their approach involves both spatial and frequency domain masking, leading to a deepfake detector based on frequency masking.

(Chen et al.) enhance detector generalization diffusion generated images by generating hard samples through high-quality diffusion reconstruction. These reconstructed images, which closely resemble real ones but retain subtle artifacts, train detectors to differentiate between real and generated images, including those from unseen diffusion models.

A.2 Feature Representation-Based Detection

Feature representation-based detection methods distinguish real images from synthesized images by leveraging deep learning models to extract and analyze complex visual features.

(Wang et al., 2020) proposed a universal detector using a ResNet-50 classifier (He et al., 2016) with random blur and JPEG compression data augmentation. When trained on images generated by a single CNN generator (ProGAN), their model demonstrated strong generalization across unseen architectures, including StyleGAN2 (Karras et al., 2020) and StyleGAN3 (Karras et al., 2021).

(Mandelli et al., 2022) developed a GAN-generated image detector based on an ensemble of CNNs. Their method emphasizes generalization by ensuring orthogonal results from CNNs and prioritizing original images during testing.

(Wang et al., 2023) introduced a technique that measures the error between an input image and its reconstructed counterpart generated by a pre-trained diffusion model. They observed that diffusion-generated images are more accurately reconstructed than real images, highlighting a key discrepancy for detection.

(Wu et al., 2023) employed language-guided contrastive learning to capture inherent differences in the distributions of real and synthetic images. Their method augments training images with designed textual labels, enabling joint image-text contrastive learning for forensic feature extraction.

(Sha et al., 2023) addressed the challenges of fake image detection and attribution. Their approach involves: (i) building a machine learning classifier to detect fake images generated by various text-to-image models, including DALL-E 2, Stable Diffusion, GLIDE, and Latent Diffusion, and benchmark prompt-image datasets such as MS

COCO and Flickr30k (Young et al., 2014); (ii) attributing fake images to their respective generative models to enhance accountability; and (iii) examining how prompts influence detection and attribution performance.

(Aghasanli et al., 2023) introduced a deepfake detection method that combines fine-tuned Vision Transformers (ViTs) with Support Vector Machines (SVMs). Their method provides interpretability by analyzing the SVMs' support vectors to distinguish between real and fake images generated by various diffusion models.

(Chen et al., 2024) proposed a straightforward method that extracts the simplest patch from an image and sends its noise pattern to a binary classifier, demonstrating effectiveness with minimal complexity.

(Koutlis and Papadopoulos, 2024) utilized intermediate outputs from CLIP's image encoder for enhanced AI-generated image detection. They introduced a Trainable Importance Estimator to dynamically assess the contributions of each Transformer block, boosting generalization across generative models.

(Liu et al.) presented OCC-CLIP, a CLIP-based framework for few-shot one-class classification. This method is particularly effective when only a few images generated by a model are available, and access to the model's parameters is restricted. OCC-CLIP combines high-level and adversarial data augmentation techniques to attribute images to specific generative models accurately.

To enhance generalization to unseen generative models, (Ojha et al., 2023) propose an approach that avoids explicitly training a classifier to distinguish real from fake images. Instead, their method leverages the feature space of large pre-trained vision-language models and employs techniques such as nearest neighbor classification.

(Tan et al., 2024) enhance the image encoder's ability to detect deepfakes by integrating category-related prompts into the text encoder of CLIP.

A.3 Hybrid Techniques

Hybrid techniques combine low-level artifact analysis with high-level semantic feature extraction to effectively distinguish AI-generated images from real ones.

(Yan et al., 2024) propose a hybrid-feature model that integrates high-level semantic information (using CLIP) with low-level artifact analysis to improve detection robustness.

(Liu et al., 2024) incorporate a forgery-aware adapter that integrates local forgery traces from both image and frequency domains. Their method employs language-guided alignment, using contrastive objectives between image features and text prompts to enhance generalization.

To guide our benchmark evaluation, we selected 17 state-of-the-art AGID methods spanning all three categories. This categorization enables us to evaluate model robustness from complementary perspectives: from low-level artifact exploitation to high-level semantic inconsistency analysis.

B Detection Performance Overview

Tables 2 and 3 provide the performance of detection techniques across COCO_{AI} and $\text{Twitter}_{\text{AI}}$ synthetic datasets, respectively. The metrics measured are Accuracy (Acc), Recall (R), and Precision (P), providing insights into each model's ability to differentiate real from AI-generated images. Below, we analyze performance across different detection techniques.

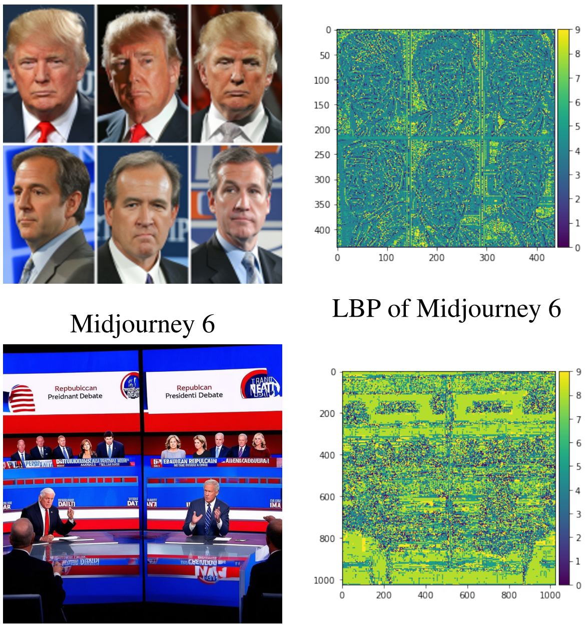
- **CNNDetection, NPR and Fake Image Detection**: These methods showed variable results, characterized by low recall but higher precision across several models. This indicates a tendency to correctly identify generated images when detected, but with many instances being missed (false negatives).

- **DM Image Detection and De-Fake**: DM Image Detection demonstrated high precision across all models, particularly excelling with Stable Diffusion versions and Midjourney 6, effectively capturing generated images. De-Fake consistently maintains strong metrics across SD (2.1, XL and 3), DALL.E 3, and Midjourney 6 but struggles with SD3.5 Large images, exhibiting lower accuracy, precision, and recall. This drop in performance likely results from SD3.5's refined generation and post-processing that minimize the artifacts and noise patterns AGID techniques depend on.

- **GAN Image Detection, SSP and DIRE**: These methods had mixed performance, particularly excelling in recall.

- **DRCT (ConvB and UnivB)**: Both versions of DRCT showed strong accuracy, recall, and precision across most models but experienced a slight performance drop with DALL.E 3, indicating challenges with proprietary models.

- **OCC-CLIP and Deep Fake Detection**: OCC-CLIP had lower recall with SDXL but balanced performance for DALL.E 3 and Midjourney 6; while Deep Fake Detection demonstrated steady, consistent performance, with all of its metrics remaining within a similar range.
- **Universal Fake Detect**: Universal Fake Detect performed better on SD (2.1, XL, and 3) models but its performance dropped when applied to SD3.5, DALL.E 3, and Midjourney 6. Notably, we observed a significant increase in recall for DALL.E 3-generated images across both datasets.
- **C2P CLIP**: C2P CLIP consistently performs poorly with low accuracy and recall, clearly showing that it often misses AI-generated images. Although its precision remains high across both datasets overall, it declines significantly for DALL.E 3 images in both datasets and for SD3 images in the Twitter dataset.



Midjourney 6

LBP of Midjourney 6

Stable Diffusion 3

LBP of Stable Diffusion 3

Figure 6: Comparative analysis of texture patterns in images generated by different T2I models using Local Binary Pattern (LBP) representation.

C Factor-level analysis

C.1 LBP Texture Analysis

Local Binary Pattern (LBP) is commonly used for texture analysis. AI-generated images vary in fine-grained texture generation. In Figure 6 we can see that an image generated by Midjourney 6 has specific facial textures and subtle expression lines whereas image generated by SD3 has inconsistencies and lack of texture in certain areas. facial features, facial structures, hair lines, edges in clothing, and wrinkles are preserved in each segment for the Midjourney image but SD3 image completely lost it.

C.2 Pairwise Scatter Plot Analysis

The pairwise scatter plots in Figures 7 through 11 visualize relationships among different features such as texture complexity, object coherence, and contextual relevance across five text-to-image models. DALL·E 3 shows a dense, compact cluster, indicating strong internal correlations between low- and high-level features and a high degree of structural stability. Midjourney 6 produces a broader, moderately dispersed cluster, reflecting greater stylistic variability and weaker statistical coupling among features, consistent with its lower Visual AI Index. Stable Diffusion 3, 3.5, and XL exhibit intermediate distributions, more organized than SD 2.1 but still fragmented, signifying partial improvements in texture-structure alignment. Stable Diffusion 2.1 displays the widest and most irregular scatter, denoting high variance and minimal feature coherence. Overall, the progression from SD 2.1 to DALL·E 3 illustrates a clear tightening of feature relationships and increasing internal consistency, highlighting how newer models achieve smoother integration of texture, structure, and semantic cues while reducing detectable artifacts.

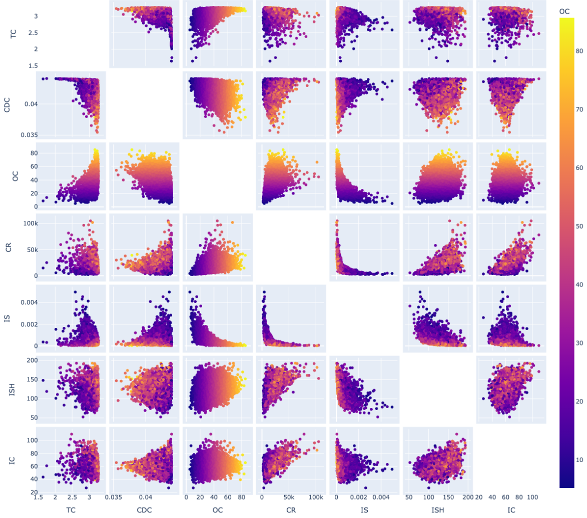


Figure 7: DALL·E 3

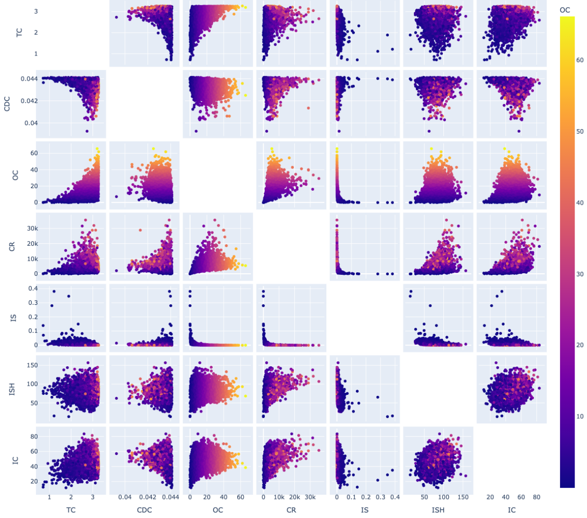


Figure 8: Midjourney 6

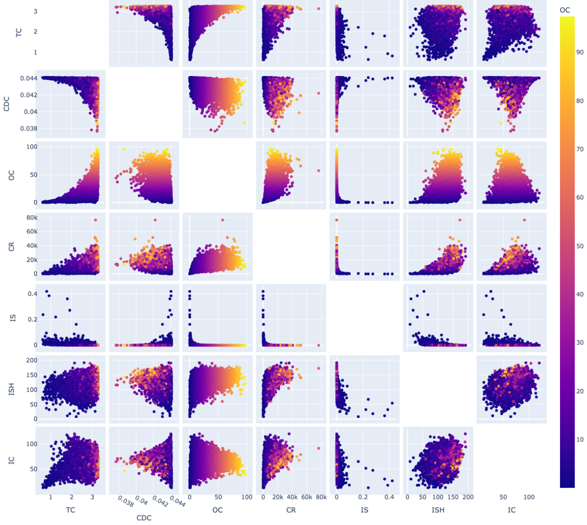


Figure 9: Stable Diffusion 3

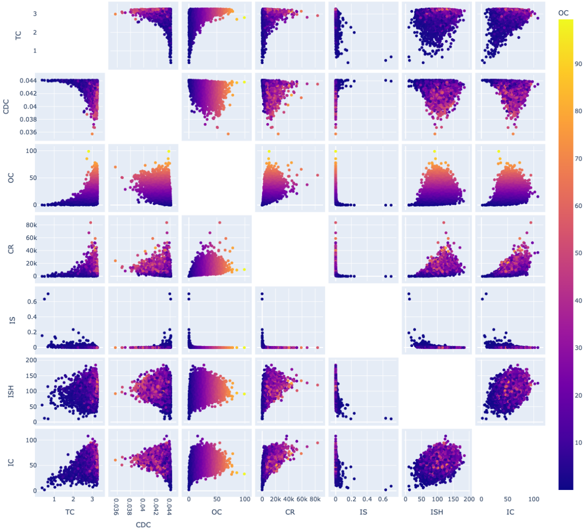


Figure 10: Stable Diffusion 2.1

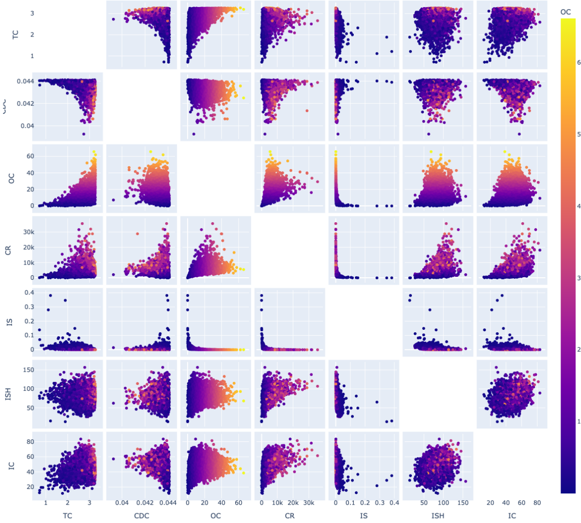
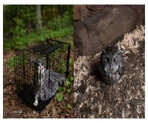
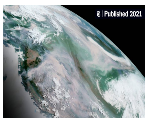
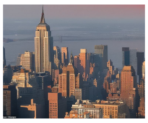
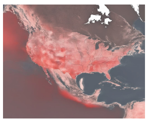
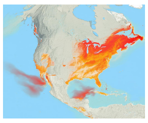
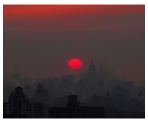
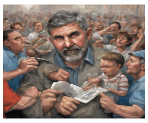
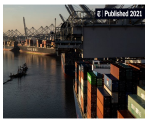
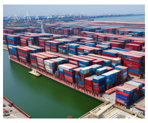
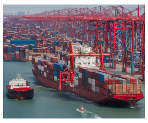
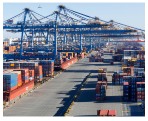
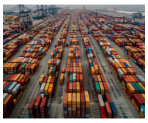
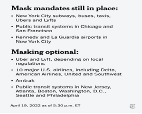
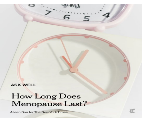


Figure 11: Stable Diffusion XL

Table 6: Real images and synthetic images generated by different models.

Real Image	SD2.1	SDXL	SD3 Medium	SD3.5 Large	DALL.E 3	Midjourney 6
						
						
						
						
 **Mask mandates still in place:** • New York City subways, buses, taxis, Ubers and Lyfts • Public transit systems in Chicago and San Francisco • Kennedy and La Guardia airports in New York City **Masking optional:** • Uber and Lyft, depending on local regulations • 10 major U.S. airlines, including Delta, American Airlines, United and Southwest • Amtrak • Public transit systems in New Jersey, Atlanta, Boston, Washington, D.C., Seattle and Philadelphia April 19, 2022 as of 5:30 p.m. ET						
 ASK WELL How Long Does Menopause Last? Aileen Son for The New York Times