The Unwinnable Arms Race of AI Image Detection

Till Aczel   Lorenzo Vettor   Andreas Plesner   Roger Wattenhofer
ETH Zürich, Switzerland
{taczel, lvettor, aplesner, wattenhofer}@ethz.ch

Abstract

The rapid progress of image generative AI has blurred the boundary between synthetic and real images, fueling an arms race between generators and discriminators. This paper investigates the conditions under which discriminators are most disadvantaged in this competition. We analyze two key factors: data dimensionality and data complexity. While increased dimensionality often strengthens the discriminator's ability to detect subtle inconsistencies, complexity introduces a more nuanced effect. Using Kolmogorov complexity as a measure of intrinsic dataset structure, we show that both very simple and highly complex datasets reduce the detectability of synthetic images; generators can learn simple datasets almost perfectly, whereas extreme diversity masks imperfections. In contrast, intermediate-complexity datasets create the most favorable conditions for detection, as generators fail to fully capture the distribution and their errors remain visible.

1 Introduction

Over the past decade, generative AI has enabled highly realistic synthetic media, including deepfakes [1]. These technologies blur the line between reality and fabrication, creating significant societal challenges [2]. While these advances have opened new possibilities in art and design [3], they have also introduced risks in disinformation, fraud, and media authenticity verification [4, 5]. Reports show thousands of deepfake attacks annually, causing hundreds of millions in financial losses and eroding public trust in digital media [6–10]. Despite growing awareness, unaided human observers perform only slightly better than chance at distinguishing AI-generated images from real photographs [11, 12]. Traditional verification systems struggle to detect AI-generated content highlighting the urgent need for robust detection methods [13]. The ability to distinguish synthetic images from real ones has therefore become increasingly important, both for security and for maintaining trust in digital media.

This dynamic has evolved into an arms race between generators, which strive to produce indistinguishable samples, and discriminators, which attempt to detect fakes. Over time, both AI-generated content and detection methods will improve, but the battle remains inherently asymmetric: if a generator perfectly captures the data distribution, no discriminator can ever win. Thus, the generator can always improve and approach a point where detection becomes impossible. Understanding the limits of detection is crucial for developing reliable tools to safeguard digital content.

Existing benchmarks focus on selecting the best discriminator. Little is known about the conditions under which discriminators are most disadvantaged, particularly when considering the full spectrum from simple, structured datasets to highly diverse, complex ones. For example, simple datasets include MNIST, which consists of centered grayscale digits with minimal variation, whereas complex datasets include CIFAR-10, which contains small color images across ten classes with significant variability in objects, backgrounds, and lighting. We quantify complexity in terms of Kolmogorov complexity, which measures the inherent compressibility or structure of a dataset. This metric is particularly relevant in the context of generative modeling and detection, as datasets that are highly

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: What Makes a Good Video: Next Practices in Video Generation and Evaluation.

compressible are easier for generators to reproduce and harder for discriminators to exploit, whereas less compressible datasets introduce variability that challenges both sides. As dataset complexity increases, the task becomes more challenging for both the generator and the discriminator. When the task is very simple, the generator can achieve near-perfect modeling, leaving the discriminator at a disadvantage. Conversely, if the dataset is extremely complex, neither the generator nor the discriminator can fully capture the distribution, and the discriminator again struggles to reliably detect fakes. In this sense, intermediate complexity presents a unique regime where the generator is imperfect but the data is structured enough for the discriminator to identify inconsistencies.

By examining this spectrum from simple to complex datasets and low to high dimensionality, we aim to map the conditions under which synthetic data are easiest and hardest to detect. Our contributions are as follows:

1. **Formal proof of the inherent challenge of detection:** We show that distinguishing generated image content from real images is an unwinnable battle, establishing theoretical limits for discriminators.
2. **Impact of dataset complexity:** We systematically analyze how the complexity of datasets, measured through Kolmogorov complexity, affects the detectability of synthetic images.
3. **Role of input resolution:** We quantify how changes in image resolution influences the ability of discriminators to detect synthetic images.

Our work serves both as a theoretical study, establishing formal limits on detectability, and as a conceptual framework, mapping how dataset complexity and resolution shape discriminator performance. Together, these perspectives reveal both the long term impossibility of perfect detection and the practical regimes where it remains feasible.

2 Related Works

Over the past decade, image generation has advanced rapidly, evolving from Variational Autoencoders (VAEs) [14] to Generative Adversarial Networks (GANs) [15], and more recently to diffusion models [16]. These generative models progressively improve the realism of synthetic images, effectively blurring the boundary between real and artificial content. This progress creates new challenges for detection, as even humans struggle to distinguish AI-generated images from authentic ones, achieving only around 62% accuracy [11, 17]. Such limitations motivate the development of automated methods capable of reliably identifying synthetic media.

Early detection approaches focused on identifying artifacts inherent to generative models. These artifacts include inconsistencies in pixel patterns, unnatural textures, irregular noise distributions, and subtle distortions in geometry or lighting [18, 19]. As generative models have become more sophisticated, deep learning classifiers have been increasingly applied to detect AI-generated images [20, 21]. Hybrid forensic systems, combining deep learning with traditional forensic techniques, have further improved detection effectiveness [22, 23]

Several novel methods emerge to address specific challenges in detection. DIRE utilizes reconstruction errors derived from diffusion model inversion to detect AI-generated images [24]. Similarly, SSP shows that even a single, carefully selected image patch can suffice for accurate detection, highlighting the presence of localized artifacts [25], and GANs have been found to generally have identifying artifacts/fingerprints [26–28]. Recent research also explores leveraging multimodal large language models, which provide visually grounded explanations for detection decisions and enhance interpretability [29]. These advances collectively illustrate the rapid evolution of detection strategies, reflecting the ongoing arms race between generative models and discriminators.

Benchmarking plays a crucial role in evaluating detector performance. Large-scale datasets such as GenImage [30] and Chameleon [22] provide diverse evaluation scenarios across a wide range of generative models, including Stable Diffusion [31], Midjourney [32], and BigGAN [33]. These benchmarks assess not only detection accuracy but also robustness under real-world conditions, such as low-resolution images, compression artifacts, and blurring. Analyses from GenImage indicate that higher-resolution images reveal finer, more detectable artifacts, improving detection performance, whereas low-resolution or compressed images present greater challenges [30].

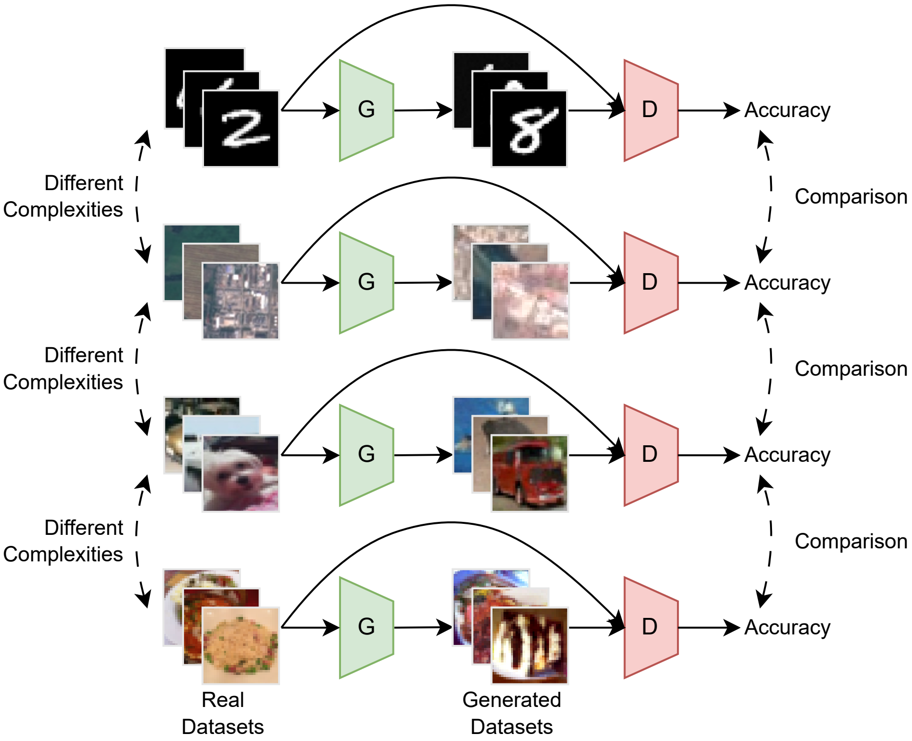


Figure 1: overview of the experimentation setup used in this project. We take multiple datasets with different (Kolmogorov) complexities or resolutions, and independently train copies of the same diffusion-based image generator and convolutional discriminator for the datasets. We then compare the accuracies of the discriminators against differences in the dataset complexities.

Some studies focus on leveraging semantic information, such as the number of fingers on a hand [34, 35]. At the same time, other work investigated how detectors designed and trained to detect GAN-generated images fail to generalize to diffusion-generated images and how detectors fail when images have been compressed [28, 36]. However, no work has investigated how dataset distribution and the resulting complexity influence detector performance.

Kolmogorov complexity provides a framework to quantify intrinsic dataset complexity by measuring the length of the shortest program capable of reproducing it [37]. For image datasets, low Kolmogorov complexity corresponds to highly structured or repetitive content, which generators can learn easily, producing nearly indistinguishable synthetic images [38].

While previous studies have focused on developing AI-generated image detectors and evaluating them on large-scale benchmarks, none have explicitly analyzed the dynamics of the ongoing arms race between generators and discriminators over time. In this work, we are the first to systematically investigate how this long-term battle unfolds, identifying conditions under which detectors hold the greatest advantage.

3 Methodology

We begin by formalizing the asymmetry of the detection problem. Let $p(x)$ denote the true data distribution and $q(x)$ the distribution induced by a generator. As long as $p \neq q$, there exists a discriminator D with non-trivial accuracy in distinguishing real from synthetic samples. In the limiting case where $q(x) = p(x)$, however, the detection task becomes ill-posed: no discriminator can do better than always guessing the more probable class.

Proposition 1. *Let a dataset consist of real and generated samples with priors π_r and $\pi_f = 1 - \pi_r$. If the generator distribution $q(x)$ equals the data distribution $p(x)$ for all x, then every discriminator has accuracy equal to* $\max\{\pi_r, \pi_f\}$.

Proof. Let $p(x)$ and $q(x)$ denote the densities of real and generated samples w.r.t. a common dominating measure. The mixture (marginal) density is

$$m(x) = \pi_r p(x) + \pi_f q(x). \tag{1}$$

Conditioned on $X = x$, the posterior probabilities of the two classes are

$$\Pr(\text{real} \mid x) = \frac{\pi_r p(x)}{m(x)}, \qquad \Pr(\text{fake} \mid x) = \frac{\pi_f q(x)}{m(x)}. \tag{2}$$

The Bayes-optimal classifier chooses the class with larger posterior probability. Hence, the pointwise probability of a correct decision given x is

$$\max\left\{\frac{\pi_r p(x)}{m(x)}, \frac{\pi_f q(x)}{m(x)}\right\}. \tag{3}$$

The overall Bayes-optimal accuracy is

$$\mathrm{Acc}^\star = \int_{\mathcal{X}} \max\left\{\frac{\pi_r p(x)}{m(x)}, \frac{\pi_f q(x)}{m(x)}\right\} m(x)\, dx = \int_{\mathcal{X}} \max\{\pi_r p(x), \pi_f q(x)\}\, dx. \tag{4}$$

If $q(x) = p(x)$ everywhere, then $\max\{\pi_r p(x), \pi_f q(x)\} = \max\{\pi_r, \pi_f\}\, p(x)$, so

$$\mathrm{Acc}^\star = \max\{\pi_r, \pi_f\} \int_{\mathcal{X}} p(x)\, dx = \max\{\pi_r, \pi_f\}. \tag{5}$$

Thus, when $q = p$, the Bayes-optimal accuracy equals the prior of the more probable class, and no discriminator can outperform this baseline. □

This establishes the theoretical limit of detection and motivates our investigation of the practical regimes where discriminators retain predictive power. In particular, we study how two dataset-dependent factors shape discriminator performance: (i) dataset complexity and (ii) input dimensionality. We use Kolmogorov complexity $K(\mathcal{D})$ as a conceptual measure of dataset complexity, and approximate it empirically using lossless compression. Dimensionality, in contrast, refers to the raw input dimension d (e.g. the number of pixels per image). While larger d expands the feature space in which distributions p and q can be separated, it also increases the sample complexity required for reliable discrimination. Our experiments thus explore the trade-off between these two factors across a wide range of datasets.

3.1 Diffusion Model and Discriminator Training

For each dataset, a diffusion model is trained to generate synthetic samples. Synthetic subsets are denoted as $D^{\text{fake}}_{\text{train}}$, $D^{\text{fake}}_{\text{val}}$, and $D^{\text{fake}}_{\text{test}}$, with sizes matched to the corresponding real subsets. The discriminator is trained on $D^{\text{real}}_{\text{train}} \cup D^{\text{fake}}_{\text{train}}$ and validated on $D^{\text{real}}_{\text{val}} \cup D^{\text{fake}}_{\text{val}}$, while evaluation is performed on $D^{\text{real}}_{\text{test}} \cup D^{\text{fake}}_{\text{test}}$. The setup is shown in Figure 1 and enables controlled comparisons of detection performance across datasets that differ in both complexity and resolution.

We evaluate six discriminator configurations that differ in architecture and input representation. The `Base` discriminator is a compact convolutional neural network with approximately $40{,}000$ parameters. The `Big` discriminator is a deeper variant with about $520{,}000$ parameters, providing increased capacity while maintaining architectural similarity to the base model.

Both the `Base` and `Big` discriminators are trained under two input modalities. In the *Pixel* setting, and in the *Fourier* setting. In the latter, a two-dimensional Fast Fourier Transform (FFT) is applied, and the log-magnitude spectrum of each channel is used as input. This results in four models: `Pixel-Base`, `Pixel-Big`, `Fourier-Base`, and `Fourier-Big`.

In addition, we evaluate a pretrained discriminator based on ResNet-18 [39] pretrained on ImageNet, comprising approximately 11M parameters. Two training regimes are considered. In the `Linear-ResNet` setting, only the final classification layer is optimized while the ResNet backbone remains frozen. In the `Finetuned-Resnet` setting, the entire network is updated end-to-end.

3.2 Approximation of Dataset Complexity

Kolmogorov complexity is not computable due to the undecidability of the halting problem[40], but compression-based methods provide a tractable and meaningful approximation [37]. Modern lossless compressors exploit redundancies and regularities in the data, yielding an effective upper bound on true Kolmogorov complexity [41]. Datasets that compress strongly exhibit high internal structure, whereas datasets that compress poorly contain greater variability.

We quantify the *Complexity* of a dataset D using its *Compression Ratio*:

$$C(D) = \frac{S_{\text{comp}}(D)}{S_{\text{orig}}(D)}, \tag{6}$$

where $S_{\text{orig}}(D)$ is the size of the dataset in bytes (raw NumPy representation) and $S_{\text{comp}}(D)$ is the size after compression. All images are concatenated into a single PNG file prior to compression to maximize exploitation of spatial redundancies [42]. This compressed size then serves as a practical proxy for Kolmogorov complexity [41].

Choice of compressor. While several compression algorithms could be employed, we adopt PNG [42] as our primary measure due to its widespread use and high optimization. For robustness, we also computed dataset complexity using `zip`, `bzip2`, `zstd`, and NumPy's `npz` (Table 3). The relative ranking of datasets remained highly consistent across methods (Spearman $\rho \approx 0.80$–0.95), confirming that PNG compression serves as a reliable proxy for dataset complexity.

4 Experiments

We evaluate the detectability of AI-generated images along two primary axes: dataset complexity and image resolution. For the complexity experiments, all datasets are zero-padded or resized to a common resolution of 32×32 pixels. For the resolution experiments, we vary the input resolution using the OrganAMNIST dataset [43, 44] as a case study. To capture a broad spectrum of dataset complexity, we consider 19 datasets drawn from diverse distributions. The complete list of datasets is provided in Table 2.

To ensure comparability across datasets, a consistent preprocessing and compression pipeline is applied to all datasets used for complexity evaluation. Diffusion models are trained for each dataset using a standardized configuration appropriate for 32×32 resolution. Apart from controlled experimental variations such as image size or dataset, architecture depth and channel width are kept constant across training runs. We employ a conditional U-Net backbone trained with the standard DDPM (Denoising Diffusion Probabilistic Model) formulation [16]. The U-Net consists of 3 levels of encoder–decoder with symmetric skip connections and self-attention blocks at multiple resolutions to capture both local and global dependencies. The base channel width is 128 and doubles after each level.

Each dataset is split into training ($D^{\text{real}}_{\text{train}}$), validation ($D^{\text{real}}_{\text{val}}$), and test ($D^{\text{real}}_{\text{test}}$) subsets. When predefined validation and test splits with reasonable sizes are available, they are preserved. Otherwise, all available images are pooled and randomly partitioned. Validation and test sets each contain either $10{,}000$ images or one eighth of the total dataset size, whichever is smaller, with the remainder used for training. For datasets with limited sample counts, standard data augmentations such as horizontal and vertical flips are applied to increase effective sample size.

For each dataset, a diffusion model is trained to produce synthetic samples, resulting in $D^{\text{fake}}_{\text{train}}$, $D^{\text{fake}}_{\text{val}}$, and $D^{\text{fake}}_{\text{test}}$ splits that mirror the sizes of their real counterparts. Discriminators are trained using $D^{\text{real}}_{\text{train}} \cup D^{\text{fake}}_{\text{train}}$, validated on $D^{\text{real}}_{\text{val}} \cup D^{\text{fake}}_{\text{val}}$, and evaluated on $D^{\text{real}}_{\text{test}} \cup D^{\text{fake}}_{\text{test}}$. This setup enables a controlled comparison of detection performance across datasets of varying complexity and dimensionality.

Each of the six discriminator variants (see Section 3) is trained and evaluated independently. For each variant, experiments are repeated five times with different random seeds, and the mean detection performance is reported.

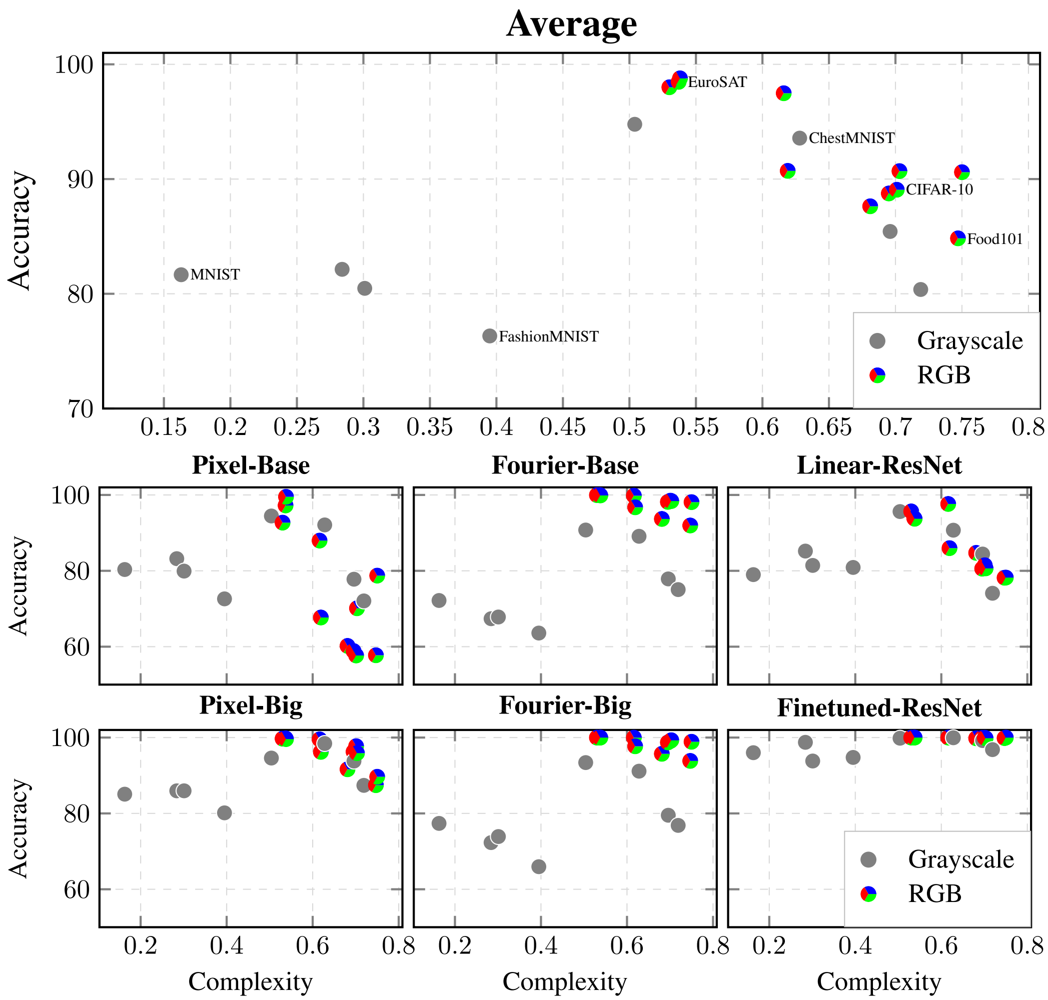


Figure 2: Discriminator accuracy across dataset complexity. **Top:** Overview across all datasets and models, gray points indicate grayscale images and tri-color points indicate RGB images. Medium-complexity datasets are easiest to detect, while simple datasets are nearly perfectly reproduced, and complex datasets mask generator errors. **Bottom:** Model-specific performance breakdown. Overall, increasing model capacity improves performance, particularly on high-complexity datasets. Fourier preprocessing boosts detection for RGB datasets, while fine-tuning ResNets achieves near-perfect accuracy across most datasets. Diminishing returns are observed when combining large models with Fourier preprocessing, and low-resolution grayscale datasets benefit less from these enhancements.

5 Results

5.1 Results Across Dataset Complexity

Figure 2 (top) summarizes the relationship between dataset complexity and discriminator accuracy. Detailed results for each dataset are reported in Table 1. The figure shows average performance across all discriminator architectures.

At low complexity, the generator can capture the distribution almost perfectly, making very few mistakes, and the discriminator has a hard task. At high complexity, the data distribution is wide, and the discriminator cannot reliably distinguish mistakes from genuine variability. At medium complexity, the discriminator excels: the generator struggles to learn the distribution, producing systematic errors, while the dataset's diversity is limited enough that these errors are clearly detectable.

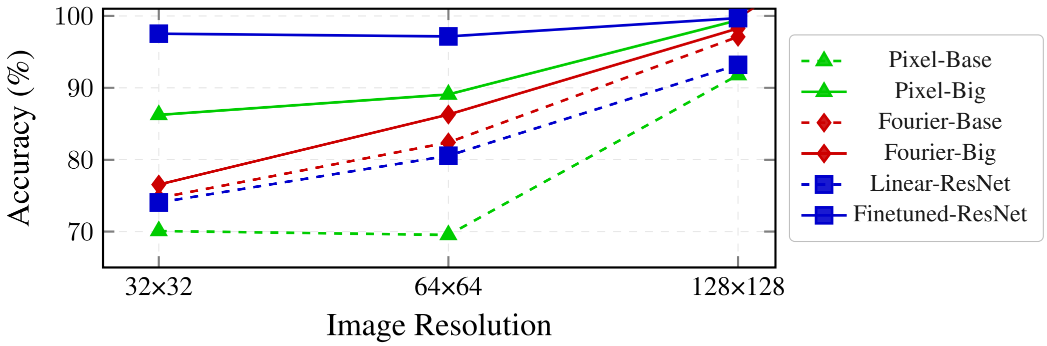


Figure 3: Classification accuracy comparison across different model architectures on the OrganAMNIST dataset. Results show performance scaling with image resolution from 32×32 to 128×128 pixels. Fine-tuned ResNet consistently achieves the highest accuracy across all resolutions.

5.2 Discriminator Capacity

Figure 2 (bottom) summarizes discriminator performance across datasets of varying complexity. Per-dataset results are given in Table 1. All models struggle on low-complexity datasets, where generators closely match the real distribution. Accuracy improves on intermediate-complexity datasets, where imperfections are more visible. For high-complexity datasets, smaller models decline, while larger ones retain accuracy by capturing subtler inconsistencies.

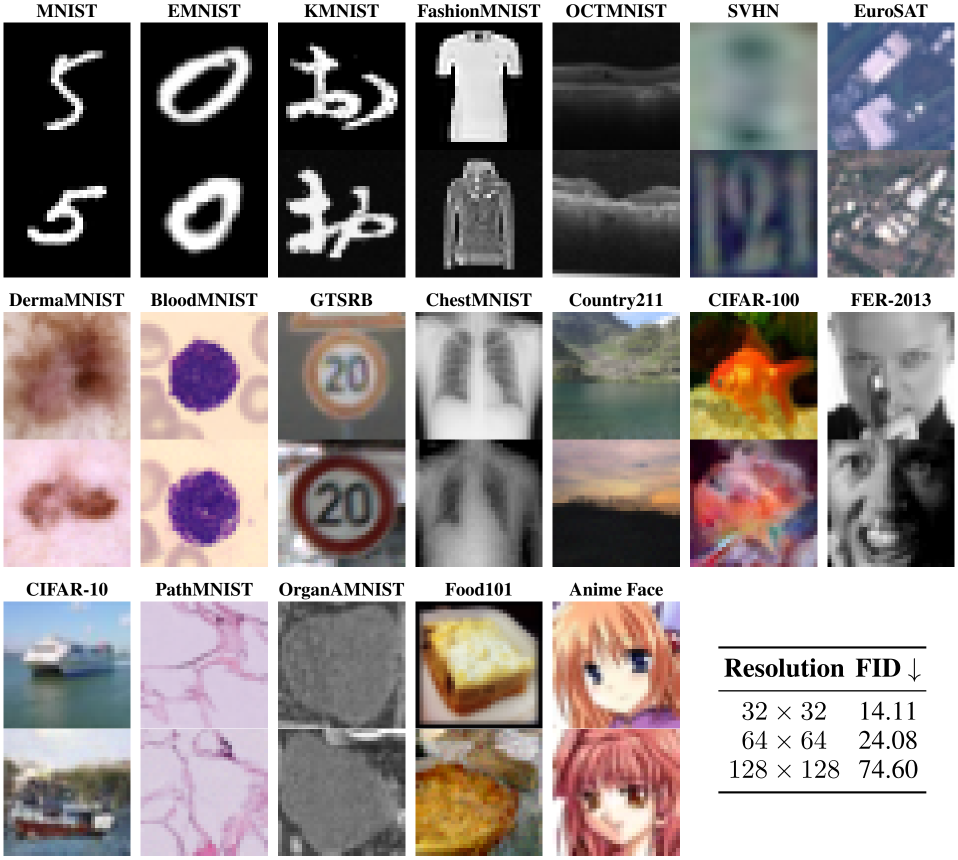
Increasing model capacity consistently improves performance. Transitioning from `Base` to `Big` CNNs boosts accuracy, particularly in high-complexity regimes. Fourier preprocessing stabilizes training and enhances detection for RGB datasets, as artifacts are spread in the frequency domain. However, combining Fourier transforms with larger CNNs yields only marginal additional improvement, suggesting diminishing returns when both capacity and preprocessing are maximized. For low-resolution grayscale datasets, Fourier preprocessing offers minimal benefit.

`ResNet` discriminators follow similar trends. Pretrained `ResNets` with linear projection perform well on intermediate datasets but struggle on the simplest and most complex ones. Fine-tuning boosts accuracy across most datasets, with more errors on low-complexity and fewer on complex datasets, showing generators still make detectable mistakes. These results show that model capacity is key for complex datasets. As complexity rises, discriminators need more expressive architectures, but practical constraints in large-scale applications emphasize the need for efficient, high-capacity models. In practical applications, such as content moderation at scale, computational constraints limit the extent to which discriminators can be enlarged, highlighting the need for efficient yet high-capacity models.

5.3 Results Across Dataset Resolution

For the multi-resolution experiment, we used the OrganAMNIST dataset and evaluated diffusion-generated images at three resolutions: 32×32, 64×64 and 128×128 pixels. At 32×32, discriminator accuracy is comparatively low across most architectures, suggesting that low-resolution generations obscure many artifacts and are therefore harder to classify as fake. The fine-tuned ResNet, however, already achieves relatively strong performance at this resolution and continues to improve as resolution increases. At 64×64, performance improves slightly across models, indicating that increased resolution exposes additional cues for detection. At 128×128, all models achieve very high accuracy, showing that higher resolutions amplify detectable differences from real data, likely due to high-frequency artifacts introduced by the generator. These results are consistent with previous studies showing that higher resolution makes it easier to detect synthetic images [30].

This highlights a key dynamic in the generator-discriminator arms race: the generator's struggle to maintain high fidelity at larger scales, evidenced by the rising FID score (Figure 4), provides a clearer signal for detectors. The high-frequency artifacts that degrade the generator's performance appear to be the very same cues that enable discriminators to achieve near-perfect accuracy at high resolution.



Resolution	FID ↓
32×32	14.11
64×64	24.08
128×128	74.60

Figure 4: Real and generated images across datasets of varying complexity. Each column shows a dataset, with real images on top and generated images on the bottom. The bottom-right table reports FID scores across resolutions. Generated images closely follow the real distributions, but FID increases with resolution, indicating that the diffusion model struggles to capture fine details.

5.4 Evaluation of the Diffusion Model

Diffusion models are a leading approach for image generation due to their ability to capture complex data distributions and produce high-quality samples. We evaluate the model qualitatively and quantitatively across datasets of varying complexity and resolution.

Image quality across dataset complexity. Figure 4 shows real images on top and generated images on the bottom for each dataset. For simpler datasets such as FashionMNIST or KMNIST, generated samples are nearly indistinguishable from real images, reflecting low Real-AI FID values (see Table 4). In contrast, for more complex datasets like CIFAR-100, SVHN, or EuroSAT, subtle imperfections remain visible, corresponding to higher Real-AI FID scores. These trends indicate that image quality decreases with increasing dataset complexity and help explain why discriminators perform better on more complex datasets, as subtle artifacts are easier to detect.

Image quality across resolution. For OrganAMNIST, the model achieves strong fidelity at lower resolutions (see Figure 4, with performance gradually decreasing as resolution increases. Overall, the diffusion model generates realistic and diverse samples for simpler datasets. These results align with the discriminator performance shown in Section 5.3 and Figure 3; as resolution increases, discriminators become more effective at detecting fake samples.

6 Limitations and Future Work

Despite our study's insights, several limitations suggest directions for future work. The analysis is limited by the choice of generative model, including architecture and hyperparameters. Diffusion model performance can vary with factors such as latent diffusion variants, number of diffusion steps, noise schedules, and other training settings, affecting image quality and artifact types. Results may differ for other generative architectures or parameter configurations. Extending the analysis to text-to-image models is also promising, as prompts introduce additional complexity that could influence discriminator performance. Finally, complexity was measured at a fixed resolution of 32×32 pixels, yet both dataset complexity and generative performance can scale with resolution, warranting further exploration.

Our study captures only a snapshot and does not consider the historical evolution of the generator-discriminator arms race. Studying past improvements could provide context for current detection challenges and reveal trends in model development. Also, human perception was not incorporated in the evaluation. Humans often serve as effective detectors of AI-generated content, so benchmarking against human performance could offer complementary insights.

The measure of dataset complexity relies on standard PNG compression, which may not fully capture the intrinsic diversity of the data. For example, a dataset of random noise could appear highly complex under this metric, even though a learned compression model could efficiently encode it. Employing learned compression schemes tailored to each dataset could provide a more accurate assessment of structural complexity.

Together, these limitations highlight both methodological constraints and opportunities for future research, including exploring higher-resolution images, alternative generative models, temporal dynamics, human-centered evaluations, and improved complexity metrics. Addressing these aspects would deepen our understanding of the conditions under which AI-generated content is most and least detectable.

7 Conclusions

Our study highlights the interplay between AI-generated image detectability, dataset complexity, data resolution, and discriminator capacity. Diffusion models are highly effective at learning simple datasets, producing images that closely match the real distribution. As dataset complexity increases, these models begin to make systematic errors, which discriminators can exploit to distinguish real from generated content. However, when datasets are extremely complex, even discriminators struggle to reliably detect fakes, as the diversity and variability in the data mask generator imperfections.

Increasing data dimensionality, such as higher-resolution images, provides the discriminator with more features and subtle cues, improving detection accuracy. Larger discriminators further enhance performance, particularly in high-complexity regimes. A key aspect we have only begun to explore is the synergistic effect of both complexity and resolution on detectability. As generators become more capable of producing high-resolution, complex images, the nature of the detectable artifacts may shift from global inconsistencies to subtle, high-frequency errors. This suggests that the "sweet spot" of intermediate complexity for detection may itself be resolution-dependent, a fascinating phenomenon that presents a rich and promising direction for further investigation to truly understand the boundaries of AI-generated content detection.

Looking forward, the rapid evolution of AI-generated content, including high-resolution images, text-to-image models, and multimodal media, presents both opportunities and challenges. Generators will continue to produce increasingly realistic content, while discriminators must adapt to maintain reliable detection. However, given the asymmetric nature of this arms race, it is likely that this battle will eventually be lost: as generative models approach perfect emulation of real data distributions, discriminators will be fundamentally limited in their ability to detect fakes. Understanding the limits of detection and the factors that influence it remains essential for building robust systems to safeguard digital media, mitigate misinformation, and preserve trust in online content. Our work provides a foundation for future research in this evolving landscape, guiding the development of both generative and discriminative AI in a responsible and informed manner.

References

[1] M. Westerlund, “The emergence of deepfake technology: A review,” *Technology innovation management review*, vol. 9, no. 11, 2019.

[2] A. De Ruiter, “The distinct wrong of deepfakes,” *Philosophy & Technology*, vol. 34, no. 4, pp. 1311–1332, 2021.

[3] E. Zhou and D. Lee, “Generative artificial intelligence, human creativity, and art,” *PNAS Nexus*, vol. 3, no. 3, p. pgae052, 03 2024. [Online]. Available: https://doi.org/10.1093/pnasnexus/pgae052

[4] P. Fraga-Lamas and T. M. Fernández-Caramés, “Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality,” *IT Professional*, vol. 22, no. 2, pp. 53–59, 2020.

[5] A. Verma, “Deepfakes and the crisis of digital authenticity: ethical challenges in the age of synthetic media,” *Journal of Information, Communication and Ethics in Society*, 08 2025. [Online]. Available: https://doi.org/10.1108/JICES-04-2025-0083

[6] VentureBeat, “Deepfakes will cost $40 billion by 2027 as adversarial ai gains momentum,” 2025. [Online]. Available: https://venturebeat.com/security/deepfakes-will-cost-40-billion-by-2027-as-adversarial-ai-gains-momentum/

[7] D. R. J. (DRJ), “Financial losses from deepfake-related fraud have reached almost $900 million,” 2025. [Online]. Available: https://drj.com/industry_news/financial-losses-from-deepfake-related-fraud-have-reached-almost-900-million/

[8] Surfshark, “Deepfake fraud losses report: 2025 mid-year update,” https://surfshark.com/research/chart/deepfake-fraud-losses, 2025.

[9] Wall Street Journal, “Ai drives rise in ceo impersonator scams,” Aug. 2025. [Online]. Available: https://www.wsj.com/articles/ai-drives-rise-in-ceo-impersonator-scams-2bd675c4

[10] T. Sippy, F. Enock, J. Bright, and H. Z. Margetts, “Behind the deepfake: 8% create; 90% concerned. surveying public exposure to and perceptions of deepfakes in the uk,” *arXiv preprint arXiv:2407.05529*, 2024. [Online]. Available: https://arxiv.org/abs/2407.05529

[11] T. Roca, A. C. Roman, J. T. Vega, M. Duarte, P. Wang, K. White, A. Misra, and J. L. Ferres, “How good are humans at detecting ai-generated images? learnings from an experiment,” *arXiv preprint arXiv:2507.18640*, 2025.

[12] M. Groh, Z. Epstein, C. Firestone, and R. Picard, “Deepfake detection by human crowds, machines, and machine-informed crowds,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, Dec. 2021. [Online]. Available: http://dx.doi.org/10.1073/pnas.2110013119

[13] A. Mahara and N. Rishe, “Methods and trends in detecting generated images: A comprehensive review,” *arXiv preprint arXiv:2502.15176*, 2025.

[14] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.

[16] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[17] Z. Lu, D. Huang, L. Bai, J. Qu, C. Wu, X. Liu, and W. Ouyang, “Seeing is not always believing: Benchmarking human and model perception of ai-generated images,” *Advances in neural information processing systems*, vol. 36, pp. 25 435–25 447, 2023.

[18] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International conference on machine learning*. PMLR, 2020, pp. 3247–3258.

[19] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8695–8704.

[20] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, and B. Manjunath, "Detecting gan generated fake images using co-occurrence matrices," *arXiv preprint arXiv:1903.06836*, 2019.

[21] J. J. Bird and A. Lotfi, "Cifake: Image classification and explainable identification of ai-generated synthetic images," *IEEE Access*, vol. 12, pp. 15 642–15 650, 2024.

[22] S. Yan, O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, and W. Xie, "A sanity check for ai-generated image detection," 2025. [Online]. Available: https://arxiv.org/abs/2406.19435

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[24] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 445–22 455.

[25] J. Chen, J. Yao, and L. Niu, "A single simple patch is all you need for ai-generated image detection," *arXiv preprint arXiv:2402.01123*, 2024.

[26] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to gans: Learning and analyzing gan fingerprints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[27] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[29] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 480–24 489.

[30] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang, "Genimage: A million-scale benchmark for detecting ai-generated image," 2023. [Online]. Available: https://arxiv.org/abs/2306.08571

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

[32] Midjourney, "Midjourney," https://www.midjourney.com/home, n.d., accessed: 2025-08-18.

[33] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[34] S. Cheng, L. Lyu, Z. Wang, X. Zhang, and V. Sehwag, "Co-spy: Combining semantic and pixel features to detect synthetic images by ai," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 13 455–13 465.

[35] Y. Zhang, Z. Qin, Y. Liu, and D. Campbell, "Detecting and restoring non-standard hands in stable diffusion generated images," *arXiv preprint arXiv:2312.04236*, 2023.

[36] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, “ Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art ,” in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2021, pp. 1–6. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICME51207.2021.9428429

[37] M. Li, P. Vitányi *et al.*, *An introduction to Kolmogorov complexity and its applications*. Springer, 2008, vol. 3.

[38] H. Zenil, F. Soler-Toscano, J.-P. Delahaye, and N. Gauvrit, “Two-dimensional kolmogorov complexity and an empirical validation of the coding theorem method by compressibility,” *PeerJ Computer Science*, vol. 1, p. e23, 2015.

[39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[40] G. J. Chaitin, A. Arslanov, and C. Calude, “Program-size complexity computes the halting problem,” Department of Computer Science, The University of Auckland, New Zealand, Tech. Rep., 1995.

[41] P. Grunwald and P. Vitányi, “Shannon information and kolmogorov complexity,” *arXiv preprint cs/0410002*, 2004.

[42] World Wide Web Consortium (W3C), “Png specification: Recommendations for encoders,” https://www.w3.org/TR/PNG-Encoders.html, 1996, accessed: 2025-08-22.

[43] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, “Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification,” *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[44] J. Yang, R. Shi, and B. Ni, “Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis,” in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 191–195.

[45] L. Deng, “The mnist database of handwritten digit images for machine learning research [best of the web],” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[46] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, “Emnist: an extension of mnist to handwritten letters,” 2017. [Online]. Available: https://arxiv.org/abs/1702.05373

[47] P. Team, “Torchvision datasets documentation,” https://pytorch.org/vision/main/datasets.html, 2025, accessed: 2025-07-22.

[48] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” 2017. [Online]. Available: https://arxiv.org/abs/1708.07747

[49] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng, “Reading digits in natural images with unsupervised feature learning,” *NIPS*, 01 2011.

[50] P. Helber, B. Bischke, A. Dengel, and D. Borth, “Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” 2019. [Online]. Available: https://arxiv.org/abs/1709.00029

[51] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, “The German Traffic Sign Recognition Benchmark: A multi-class classification competition,” in *IEEE International Joint Conference on Neural Networks*, 2011, pp. 1453–1460.

[52] A. Krizhevsky, V. Nair, and G. Hinton, “Cifar-10 and cifar-100 datasets,” https://www.cs.toronto.edu/~kriz/cifar.html, 2009.

[53] msambare, “Fer-2013: Facial expression recognition dataset,” Kaggle, urlhttps://www.kaggle.com/datasets/msambare/fer2013, 2020.

[54] L. Bossard, M. Guillaumin, and L. Van Gool, “Food-101 – mining discriminative components with random forests,” in *European Conference on Computer Vision*, 2014.

[55] S. Churchill and B. Chao, “Anime face dataset,” 2019. [Online]. Available: https://www.kaggle.com/ds/379764

[56] S. Sadat, O. Hilliges, and R. M. Weber, “Eliminating oversaturation and artifacts of high guidance scales in diffusion models,” 2025. [Online]. Available: https://arxiv.org/abs/2410.02416

[57] T. Karras, M. Aittala, T. Kynkäänniemi, J. Lehtinen, T. Aila, and S. Laine, “Guiding a diffusion model with a bad version of itself,” 2024. [Online]. Available: https://arxiv.org/abs/2406.02507

8 Appendix

8.1 Extended Results

Table 1 provides a comprehensive overview of classification accuracy across all datasets and discriminator architectures. The datasets span a wide range of intrinsic complexity, from simple digit datasets such as MNIST and KMNIST, to moderately complex datasets such as OCTMNIST and SVHN, to highly diverse datasets including CIFAR-10, PathMNIST, and Food101. The complexity column quantitatively reflects the structural richness and variability of each dataset, providing context for the observed performance trends.

Table 1: Classification accuracy comparison across different models and datasets. The complexity column provides a quantitative measure of each dataset's structural richness and variability. Average accuracy across all models is also included for comparison. Fine-tuned ResNet consistently achieves the highest accuracy across datasets, demonstrating the importance of model capacity for handling complex and diverse image data.

Dataset	Complexity	Average	Pixel-Base	Pixel-Big	Fourier-Base	Fourier-Big	Linear-ResNet	Finetuned-ResNet
MNIST[45]	0.163	81.7	80.3 ± 2.7	85.1 ± 4.9	72.2 ± 0.6	77.4 ± 0.6	79.0 ± 0.1	96.0 ± 3.9
EMNIST[46]	0.284	82.1	83.2 ± 1.0	86.0 ± 11.0	67.4 ± 0.8	72.3 ± 1.6	85.2 ± 0.1	98.7 ± 1.3
KMNIST[47]	0.301	80.5	79.9 ± 0.4	86.0 ± 1.1	67.8 ± 0.7	73.9 ± 1.4	81.4 ± 0.2	93.8 ± 3.8
FashionMNIST[48]	0.395	76.3	72.6 ± 1.2	80.1 ± 3.1	63.6 ± 0.3	66.0 ± 0.5	80.9 ± 0.1	94.8 ± 4.4
OCTMNIST[43, 44]	0.504	94.8	94.5 ± 2.9	94.6 ± 10.4	90.8 ± 0.4	93.4 ± 0.4	95.6 ± 0.1	99.8 ± 0.0
SVHN[49]	0.530	98.0	92.7 ± 7.1	99.7 ± 0.1	100.0 ± 0.0	100.0 ± 0.0	95.7 ± 0.1	100.0 ± 0.0
Eurosat-AUG[50]	0.537	98.5	97.2 ± 4.5	99.8 ± 0.2	100.0 ± 0.0	100.0 ± 0.0	93.9 ± 0.1	100.0 ± 0.0
DermaMNIST-AUG[43, 44]	0.538	98.8	99.5 ± 0.3	99.6 ± 0.3	99.9 ± 0.0	100.0 ± 0.0	93.7 ± 0.2	100.0 ± 0.0
BloodMNIST-AUG[43, 44]	0.616	97.5	88.0 ± 5.7	99.5 ± 0.7	99.8 ± 0.0	100.0 ± 0.0	97.6 ± 0.0	100.0 ± 0.0
GTSRB[51]	0.619	90.7	67.7 ± 0.6	96.2 ± 2.5	96.7 ± 0.2	97.7 ± 0.1	86.0 ± 0.2	100.0 ± 0.0
ChestMNIST[43, 44]	0.628	93.6	92.1 ± 5.6	98.4 ± 0.6	89.1 ± 0.6	91.1 ± 0.3	90.7 ± 0.1	99.9 ± 0.0
Country211[47]	0.681	87.6	60.2 ± 3.0	91.6 ± 3.2	93.7 ± 0.3	95.7 ± 0.1	84.7 ± 0.2	99.8 ± 0.0
CIFAR-100[52]	0.695	88.7	58.9 ± 1.1	96.3 ± 1.3	98.2 ± 0.1	98.7 ± 0.0	80.6 ± 0.3	99.9 ± 0.0
Fer-2013-AUG[53]	0.696	85.4	77.8 ± 5.2	93.8 ± 1.3	77.9 ± 0.2	79.5 ± 0.5	84.4 ± 0.3	99.2 ± 0.1
CIFAR-10[52]	0.701	89.1	57.7 ± 1.5	97.7 ± 0.6	98.5 ± 0.2	99.1 ± 0.1	81.5 ± 0.3	99.9 ± 0.0
PathMNIST[43, 44]	0.703	90.7	70.1 ± 0.6	95.9 ± 2.3	98.4 ± 0.1	99.3 ± 0.1	80.6 ± 0.1	99.8 ± 0.1
OrganAMNIST[43, 44]	0.719	80.4	72.1 ± 1.5	87.4 ± 4.2	75.0 ± 0.4	76.8 ± 0.3	74.1 ± 0.1	96.8 ± 1.1
Food101[54]	0.747	84.8	57.7 ± 3.0	87.4 ± 4.6	92.0 ± 0.8	93.8 ± 0.4	78.2 ± 0.1	99.9 ± 0.0
Anime Face Dataset[55]	0.750	90.6	78.7 ± 2.6	89.7 ± 6.9	98.1 ± 0.2	98.9 ± 0.1	78.2 ± 0.1	100.0 ± 0.0

8.2 Dataset Collection and Preparation

The datasets (see Table 2) used in this study are primarily obtained from *Torchvision Datasets* [47] and the *MedMNIST+* collections [43, 44]. Table 2 summarizes all datasets, including the number of training, validation, and test samples, as well as total sizes.

To mitigate the limited number of samples in some datasets, data augmentation is applied. Augmented datasets are indicated by the suffix *AUG*. EuroSAT and FER-2013 are augmented via horizontal flips. BloodMNIST and DermaMNIST are augmented using horizontal flips, vertical flips, and combined horizontal and vertical flips.

To ensure consistent image dimensions, MNIST, EMNIST, KMNIST, and FashionMNIST (28×28) are padded with black pixels to reach 32×32. All other datasets are either already at the target resolution or resized directly to 32×32.

To study the effect of input dimensionality, we select OrganAMNIST because it is available at a high resolution of 128×128, which can be downscaled to 64×64 and 32×32 as needed.

8.2.1 Dataset Complexity

To quantify the intrinsic complexity of the datasets, we approximate Kolmogorov complexity using compression-based measures. Each dataset is concatenated into a single PNG file, and the resulting compression ratio serves as a practical proxy for complexity. Datasets that compress efficiently exhibit more regularity and lower complexity, whereas datasets that compress poorly contain higher variability and are considered more complex.

Table 3 reports the complexity values for all datasets, obtained both from PNG concatenation and alternative compression methods (`Zip`, `bzip2`, `Zstd`, `NumPy NPZ`). Lower values correspond to simpler datasets such as MNIST, while higher values correspond to more complex datasets such as Food-101 and Anime.

Table 2: Summary of datasets used in our study. Train, validation and test split are listed. Overview on the augmentation of smaller datasets is also provided.

Dataset	**Train**	**Val**	**Test**	**Total**
MNIST[45]	50,000	10,000	10,000	70,000
EMNIST[46]	102,800	10,000	10,000	122,800
KMNIST[47]	50,000	10,000	10,000	70,000
FashionMNIST[48]	50,000	10,000	10,000	70,000
OCTMNIST[43, 44]	89,309	10,000	10,000	109,309
SVHN[49]	79,289	10,000	10,000	99,289
EuroSAT[50]	20,250	3,375	3,375	27,000
EuroSAT-AUG	40,500	6,750	6,750	54,000
DermaMNIST[43, 44]	7,513	1,251	1,251	10,015
DermaMNIST-AUG	30,052	5,004	5,004	40,060
BloodMNIST[43, 44]	12,820	2,136	2,136	17,092
BloodMNIST-AUG	51,280	8,544	8,544	68,368
GTSRB[51]	38,881	6,479	6,479	51,839
ChestMNIST[43, 44]	92,120	10,000	10,000	112,120
Country211[47]	47,476	7,912	7,912	63,300
CIFAR-100[52]	40,000	10,000	10,000	60,000
FER-2013[53]	25,887	5,000	5,000	35,887
FER-2013-AUG	51,774	10,000	10,000	71,774
CIFAR-10[52]	40,000	10,000	10,000	60,000
PathMNIST[43, 44]	87,180	10,000	10,000	107,180
OrganAMNIST[43, 44]	44,124	7,353	7,353	58,830
Food-101[54]	81,000	10,000	10,000	101,000
Anime Face Dataset[55]	47,675	7,945	7,945	63,565

Random removal of samples from a dataset changes its size and compressed size but does not affect the underlying data distribution and maintains the same Compression Ratio. Formally, if $D = \{x_1, \ldots, x_N\}$ and $D' \subset D$ is obtained by removing k samples, the expected empirical distribution of D' satisfies

$$\mathbb{E}[\hat{p}_{D'}(x)] = \hat{p}_D(x),$$

demonstrating that the underlying distribution remains preserved. For example, removing $10{,}000$ samples from MNIST ($N = 50{,}000$) reduces the compressed size by $1/5$, while the data distribution remains approximately the same.

8.3 Diffusion Model Training Procedure

The training of our diffusion model follows the framework described in the theoretical background and incorporates several practical considerations to ensure consistency across experiments.

Controlling Experimental Variables: To isolate the effect of dataset complexity and image resolution, we control all other training variables to prevent confounding factors.

Number of Iterations: All models are trained for a fixed total of 5 million iterations. This number was chosen empirically to ensure smooth convergence across all datasets.

Architecture: We use a conditional U-Net with three levels of encoder–decoder blocks, symmetric skip connections, and self-attention layers at multiple resolutions to capture both local and global dependencies. The base channel width is 64, doubling at each successive level. Architecture depth and channel width are held constant across experiments to isolate dataset and resolution effects.

Optimization Strategy: Training uses the AdamW optimizer with a one-cycle learning rate scheduler and weight decay to facilitate stable convergence. A linear noise schedule is applied, and an exponential moving average (EMA) of the model weights is maintained, as EMA weights generally yield higher sample quality at inference.

Table 3: Complexity values of datasets obtained using PNG concatenation as well as alternative compression methods (PNG folder compression, Zip, bzip2, Zstd, and NumPy NPZ). The values serve as proxies for the intrinsic complexity of each dataset, with lower values indicating simpler, more regular datasets and higher values indicating more complex, diverse datasets.

Dataset	**PNG concat.**	**PNG Folder**	**Zip of PNG**	**bzip2**	**Zstd**	**NPZ**
MNIST[45]	0.16	0.27	0.37	0.13	0.16	0.16
EMNIST[46]	0.28	0.41	0.50	0.18	0.24	0.25
KMNIST[47]	0.30	0.43	0.53	0.27	0.31	0.31
FashionMNIST[48]	0.40	0.51	0.61	0.39	0.43	0.44
OCTMNIST[43, 44]	0.50	0.62	0.72	0.50	0.68	0.66
SVHN[49]	0.53	0.56	0.59	0.79	0.95	0.88
EuroSAT[50]	0.54	0.56	0.60	0.59	0.76	0.73
DermaMNIST[43, 44]	0.54	0.57	0.60	0.77	0.94	0.87
BloodMNIST[43, 44]	0.62	0.66	0.70	0.58	0.80	0.79
GTSRB[51]	0.62	0.65	0.68	0.71	0.86	0.83
ChestMNIST[43, 44]	0.63	0.74	0.84	0.75	0.98	0.94
Country211[47]	0.68	0.71	0.74	0.85	0.97	0.90
CIFAR-100[52]	0.70	0.73	0.76	0.86	0.96	0.91
FER-2013[53]	0.70	0.82	0.92	0.81	0.97	0.97
CIFAR-10[52]	0.70	0.74	0.77	0.86	0.97	0.92
PathMNIST[43, 44]	0.70	0.73	0.76	0.63	0.84	0.81
OrganAMNIST[43, 44]	0.72	0.84	0.94	0.76	0.88	0.89
Food-101[54]	0.75	0.78	0.82	0.91	1.00	0.97
Anime Face Dataset[55]	0.75	0.82	0.85	0.89	0.97	0.95

Model Selection: After each epoch, the model is validated on $D_{\text{val}}^{\text{real}}$ using the same MSE objective employed during training. This validation allows for consistent monitoring of convergence and ensures comparability across runs.

By carefully controlling these factors, any observed differences in generative performance can be confidently attributed to variations in dataset complexity or image resolution, rather than inconsistencies in architecture, optimization, or training procedure.

8.4 AI Image Generation

During image generation, noise is iteratively transformed according to a conditional label. Generated datasets are sampled to match the size of the corresponding real subsets:

$$|D_{\text{train}}^{\text{real}}| = |D_{\text{train}}^{\text{gen}}|, \quad |D_{\text{val}}^{\text{real}}| = |D_{\text{val}}^{\text{gen}}|, \quad |D_{\text{test}}^{\text{real}}| = |D_{\text{test}}^{\text{gen}}|.$$

Figures 5 and 6 illustrate the effect of varying the classifier-free guidance (CFG) parameter on image saturation for FashionMNIST and CIFAR-10. Increasing CFG values (0, 2, 10) produces visibly more saturated images, demonstrating how generation parameters can influence dataset characteristics.

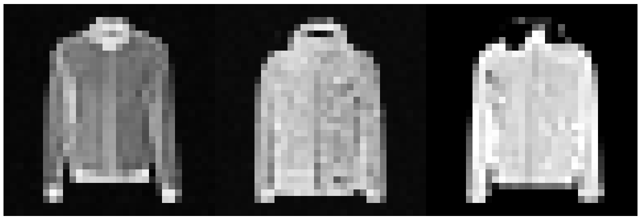


Figure 5: FashionMNIST images with increasing CFG values (0, 2, 10). Saturation increases with higher CFG.

We do not apply techniques such as Classifier-Free Guidance (CFG) to improve sample quality. CFG introduces a weight parameter w that requires careful tuning, which varies across datasets. Optimizing w would introduce an uncontrolled variable that could influence discriminator performance.

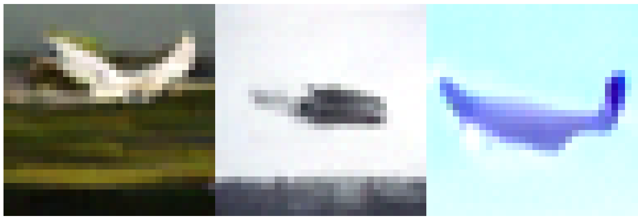


Figure 6: CIFAR-10 images with increasing CFG values (0, 2, 10). Saturation increases with higher CFG.

Prior work shows that CFG affects image saturation [56], potentially providing a trivial signal for discriminators.

Other methods, such as Autoguidance [57] or APG [56], mitigate this issue or improve FID. However, these methods also require dataset-specific optimization, which would similarly introduce non-controlled variables. By avoiding these techniques, we ensure that discriminator evaluation reflects intrinsic dataset and model characteristics rather than artifacts of sampling parameter tuning.

Table 4 reports the Fréchet Inception Distance (FID) between real and generated datasets, alongside the FID between training and validation subsets of the real data. The Real-AI FID provides a quantitative measure of how closely the diffusion model replicates the distribution of real images, with lower values indicating higher fidelity. The Train-Val FID serves as a baseline and a lower bound on achievable FID, capturing the natural variability within the real dataset itself. Across datasets, FID values vary significantly, reflecting differences in dataset complexity, image diversity, and the inherent difficulty of generation. For simpler datasets like FashionMNIST or KMNIST, Real-AI FID is low and close to the Train-Val baseline, while complex datasets such as ChestMNIST or PathMNIST show substantially higher FID, indicating that the model struggles more to capture intricate visual patterns. Interestingly, datasets with intermediate complexity such as SVHN and EuroSAT are the ones with highest FID scores. These results highlight both the strengths and limitations of the diffusion model in reproducing diverse datasets and provide context for subsequent discriminator evaluations.

Table 4: Fréchet Inception Distance (FID) comparisons for real versus generated datasets (Real-AI FID) and between training and validation splits of real data (Train-Val FID) across multiple datasets. Lower FID indicates higher fidelity to the real distribution.

Dataset	**Real-AI FID**	**Train-Val FID**
MNIST[45]	14.97	0.74
KMNIST[47]	8.19	0.95
FashionMNIST[48]	6.27	1.47
EMNIST[46]	11.01	0.69
SVHN[49]	68.71	1.90
EuroSAT[50]	44.20	5.23
BloodMNIST[43, 44]	9.35	1.99
GTSRB[51]	15.64	3.04
Country211[47]	23.27	4.39
CIFAR-100[52]	21.91	3.69
CIFAR-10[52]	17.45	3.23
FER-2013[53]	15.30	3.32
OrganAMNIST[43, 44]	14.11	2.82
Food101[54]	19.41	2.82
DermaMNIST[43, 44]	19.99	6.22
OCTMNIST[43, 44]	10.26	1.04
ChestMNIST[43, 44]	15.45	1.16
PathMNIST[43, 44]	14.77	1.51
Anime Face Dataset [55]	13.64	2.36

Table 5: Overview of discriminator variants, including input modality and the number of tunable parameters.

Model	Input Modality	# Tunable Parameters
Base	Normalized	$\approx 40{,}000$
Base	Fourier	$\approx 40{,}000$
Big	Normalized	$\approx 520{,}000$
Big	Fourier	$\approx 520{,}000$
ResNet (Frozen)	Normalized	≈ 500
ResNet (Fine-tuned)	Normalized	$\approx 11{,}000{,}000$

8.5 Discriminator Model Training

This section describes the training procedure for discriminators tasked with distinguishing real images from AI-generated ones. Table 5 gives an overview of the models used.

The input data consists of two sources: real images and generated images. To ensure a balanced dataset, the number of real samples is matched to the number of AI-generated samples.

The split of D^{real} corresponds to that used during diffusion model training, which only used $D^{\text{real}}_{\text{train}}$ and $D^{\text{real}}_{\text{val}}$.

For discriminator training, the number of AI-generated images in $D^{\text{gen}}_{\text{train}}$ is capped at 50,000, yielding a maximum total of 100,000 images. The training, validation, and test sets for the discriminator contain real and generated images in equal proportion, forming $D^{\text{discr}} = D^{\text{real}} \cup D^{\text{gen}}$.

Each image is assigned a binary label:

$$y = \begin{cases} 1 & \text{if the image is real,} \\ 0 & \text{if the image is AI-generated.} \end{cases}$$

We evaluate six discriminator variants, differing in architecture and input representation, while keeping hyperparameters consistent.

All models are trained using the AdamW optimizer with

$$\alpha = 2 \times 10^{-4}, \quad \beta_1 = 0.5, \quad \beta_2 = 0.999.$$

The loss function is Binary Cross-Entropy with logits. Training is performed for 1 million iterations.

After each epoch, the discriminator is evaluated on the validation set $D^{\text{discr}}_{\text{val}}$. The model achieving the lowest validation loss is selected for final evaluation.

9 Computational Resources

All experiments were conducted on an internal compute cluster equipped with RTX 3090 GPUs. In total, we logged 2,261 GPU-hours for both exploratory experiments and the reported results.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach is only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See for instance Section 8.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is provided in a zip file

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they are chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they are calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See overall numbers in Section 9.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There are no direct impacts from this work. The model sizes are too small for general usage of deep fakes.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent is obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks are disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) are obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.