VIRTUAL CLASSIFIER: A REVERSED APPROACH FOR ROBUST IMAGE EVALUATION

Jizhe Zhang[1]* **Yifei Wang**[2]* **Yisen Wang**[3, 4]†
[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2] MIT CSAIL
[3] National Key Lab of General Artificial Intelligence,
School of Intelligence Science and Technology, Peking University
[4] Institute for Artificial Intelligence, Peking University

ABSTRACT

Reliable evaluation of visual generative models has been a long-lasting problem. Existing evaluation metrics like Inception score and FID all follow the same methodology, that is, to calculate feature statistics of generated images based on a backbone network pretrained from real-world images (e.g., ImageNet). However, recent papers find that these methods are often biased and inconsistent with humans. Besides, we find that these metrics are very sensitive to slight (even imperceptible) image perturbations. To develop a more robust metric aligned with humans, we explore a new *reversed* approach, that is to pretrain a model from generated training data and evaluate it on natural test data. It is based on the insight that a lower test error on natural data would, in turn, indicate that the training data are of higher quality. We show that this metric, we call Virtual Classifier Error (VCE), aligns better with human evaluation compared to FID, while being more robust against image noises. Conceptually, VCE suggests a new *pragmatic* perspective to measure data quality by their usefulness for model training instead of perceptual similarities.

1 INTRODUCTION

In the past few years, generative models have rapidly developed and the quality of generated images have become increasingly close to real images, and sometimes it even becomes hard for ordinary people to distinguish them (Croitoru et al., 2023). The evaluation of generative models provides a vital role in this development by providing objective metrics to guide the technical progress (Bińkowski et al., 2018; Rambhatla and Misra, 2023).

Table 1: Comparing FID and VCE under very small Gaussian noise $\mathcal{N}(0, 0.01)$ and round noises (0.1) with BigGAN on CIFAR-10 (examples in Appendix E).

Noise	None	Gaussian	Round
FID	3.87	13.49	39.01
VCE	14.79	15.14	16.1

A commonly used evaluation metrics currently is Fréchet Inception Distance (FID) (Heusel et al., 2017), which calculates the Wasserstein-2 distance between real and generated features extracted by Inception-V3 (Szegedy et al., 2016) pretrained on ImageNet. However, employing a pretrained backbone to evaluate generative models has major drawbacks. Firstly, as pointed out by Veeramacheneni et al. (2023), FID tends to perform better when the evaluation set closely resembles ImageNet classes or if the generators utilize ImageNet weights, which can bias the output distribution towards ImageNet. As a result, FID-like metrics favor certain generative models (like GANs) that tend to generate object-centric generated data, even if they do not look realistic to humans. Secondly, these metrics are highly sensitive to noise. As shown in Table 1, FID varies a lot under subtle image perturbations that are perceptually indifferent to humans. This is because the features obtained from pretrained backbones are not robust to input perturbations.

*Equal Contribution.

†Corresponding Author: Yisen Wang (yisen.wang@pku.edu.cn).

Existing solutions to these drawbacks mainly focus on adopting other pretrained backbones like DINO (Stein et al., 2023), which may still suffer from similar risks. To fundamentally alleviate these drawbacks, we explore a new methodology for image evaluation that reverses the evaluation pipeline: We first train a classifier (named a virtual classifier) using *generated training data*, and use the test error (named Virtual Classifier Error (VCE)) on *natural test data* as the image quality measurement (lower the better). Intuitively, generated data drawn from a poor image generator deviate far from the real data, so when used for training, it will have limited generalization to natural test data. On the contrary, high-quality generated data are close to the real data distribution and can attain lower test error. Using generated data for training instead of for testing has the following benefits: 1) it is more robust against perturbations in generated data, which can be largely ignored during training and does not have a significant impact on final test error; 2) by training on generated data, neural networks can elicit more features that naturally pretrained models may overlook — especially those *not* present in natural data, *e.g.,* unnatural fingers, racial and gender bias, and subtle spurious features — and testing their generalization on natural data gives a fair account for the influences of these features; 3) it is more pragmatism-centred, and has a better indication for data quality under scenarios when the generated data are to be used to enrich the training dataset. These advantages motivate us to investigate the use of virtual classifiers for image quality evaluation.

We conduct preliminary studies of VCE with some well-known state-of-the-art generative models, including both GANs and diffusion models. The experiments result suggest that the virtual classifier can attain better alignment with human scores of image quality, mitigate subtle noises and variations in the generated dataset during the training process, and alleviate ImageNet class bias.

Future Works. In this workshop version, we mainly focus on conditional generative models where generated data are labeled, so that we can train a synthetic classifier. We leave unconditional models to be explored in future works, where one may utilize self-supervised learning methods, or even generative models themselves, for evaluating the generalization. Besides, the main idea of VCE is generic and can be generalized to other domains as well (like text). Considering the advantages elaborated above, it has the potential to become a competitive evaluation criterion when robustness, generalization, and data utility are prioritised.

2 Virtual Classifier Error

In this section, we introduce a novel metric called VCE which utilizes a classifier trained on generated datasets to evaluate generative models. Our evaluation focuses on widely prevalent conditional generative models (Pérez et al., 2020; Li et al., 2023; Montserrat et al., 2019). In the future, we will extend this metric to evaluate unconditional generative models(Hong et al., 2023) by self-supervised learning (Jing and Tian) or unsupervised learning (Khanum et al., 2015) methods.

Denote the distribution of the real training $\mathcal{D}_{tr}$ and test dataset $\mathcal{D}_{te}$ as P_d, and that of the generated data $\mathcal{D}_g$ as P_g. A generative model is trained on $\mathcal{D}_{tr}$, aiming to learn a distribution P_g that closely approximates P_d. To evaluate generative models, we draw N samples from $\mathcal{D}_g$ as training dataset to train a virtual classifier $F_g : \mathbb{R}^d \to \mathcal{Y}$, where $\mathcal{Y} = \{1, \dots, c\}$ denote the label space. We adopt the classification error of F_g on natural test data as a measure of generate data quality, written as

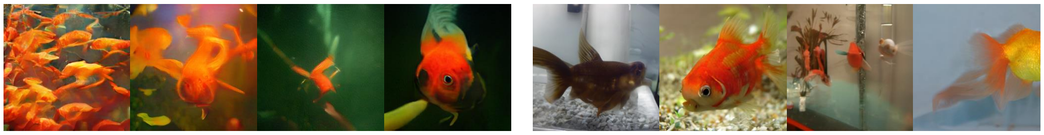
$$\mathrm{VCE}(\mathcal{D}_g) = \mathbb{E}_x \mathbb{1}(F_g(x) \neq y), \tag{1}$$

where $x \in \mathcal{D}_{te}$ is a test sample, and $y \in \mathcal{Y}$ is the true label of x. Intuitively, a smaller VCE indicates that the generalization ability of F_g on $\mathcal{D}_{te}$ is stronger, implying a smaller gap between P_g and P_d.

Theoretical Connection. Next, we provide formal proof of the effectiveness of VCE. The distribution gap between real and generated data can be characterized by the following Theorem 2.1 (see proof in Appendix B.1).

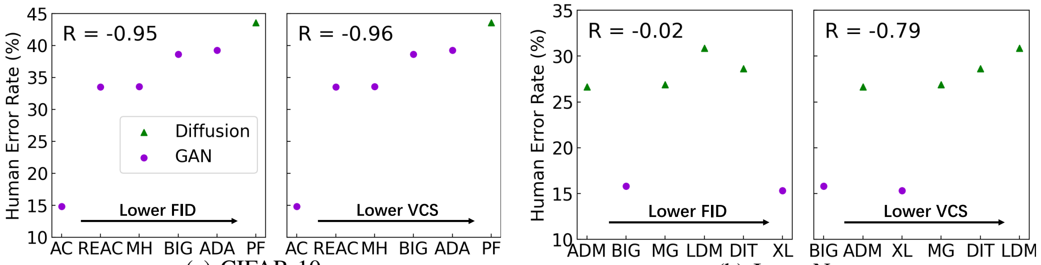
Theorem 2.1. $D_{\mathrm{KL}}(P_d(Y|X)\|P_g(Y|X)) \leq D_{\mathrm{KL}}(P_d(X,Y)\|P_g(X,Y))$*, where* $D_{KL}(p\|q)$ *is the KL-divergence between the distributions p and q,* $P(Y|X)$ *is the conditional class distribution, and* $P(X,Y)$ *is the joint probability distribution.*

Observe that $D_{\mathrm{KL}}(P_d(Y|X)\|P_g(Y|X)) = H(P_d(Y|X)) - \mathbb{E}_{P_d(X,Y)} P_g(Y|X)$, where the latter term corresponds to the CE classification loss. Therefore, the classification error on $\mathcal{D}_e$ of F_g reflects the distribution gap between P_d and P_g.



(a) BigGAN, FID=7.34, VCE=79.76 (b) ADM, FID=10.27, VCE=77.61

Figure 1: Generated goldfish from BigGAN and ADM (diffusion model). BigGAN excels at capturing the key feature of goldfish (the yellow color) but overlooks other details and backgrounds. In contrast, ADM focuses more on the entire image, including both the goldfish and the background, resulting in generated images that better represent the real distribution.



(a) CIFAR-10 (b) ImageNet

Figure 2: FID *v.s.* VCE on generative models of CIFAR-10 and ImageNet. R represents the correlation coefficient between FID,VCE and human error rate.

3 Experiments

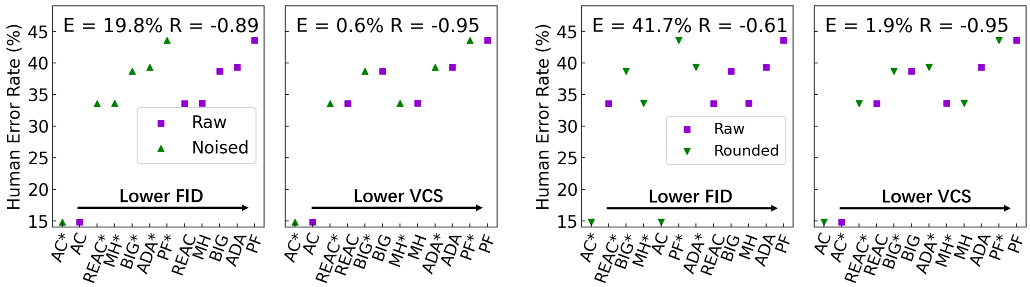
In this section, we evaluate VCE and FID on popular generative models. The results demonstrate that VCE aligns better with human perception, and exhibits greater robustness than FID.

Experiment Setup. We conduct experiments on CIFAR-10 (Krizhevsky et al., 2009) and ImageNet1K (Deng et al., 2009). The generated images are derived from two popular types of generative models: GANs (Szegedy et al., 2014) and diffusion models (Ho et al., 2020). The generative models of CIFAR-10 we used are PFGM++ (PF) (Xu et al., 2023)), StyleGAN2-ADA (ADA) (Karras et al., 2020), BigGAN (BIG) (Brock et al., 2018), MHGAN (MH) (Turner et al., 2019), ReACGAN (REAC) (Kang et al., 2021), ACGAN (AC) (Odena et al., 2017). The generative models we used for ImageNet are ADM (Dhariwal and Nichol, 2021), ADMG-ADMU (MG) (Dhariwal and Nichol, 2021), BigGAN (BIG) (Brock et al., 2018), DiT-XL-2 (DIT) (Peebles and Xie, 2023), LDM (Rombach et al., 2022), StyleGAN-XL (XL) (Sauer et al., 2022). The resolution of the generated images of ImageNet is 256×256. We use human error rate (Zhou et al., 2019) as a ground truth metric, as it reflects the evaluation of generative models based on human perception. When a generated dataset is highly realistic, it becomes challenging for humans to determine whether an image is real or fake, resulting in a high human error rate. The generated images and the results of the human error rate used in our experiments are sourced from Stein et al. (2023).

VCE Aligns Better with Human Perception than FID. We compare VCE and FID on CIFAR-10 and ImageNet in Figure 2. Figure 2(a) demonstrates that both VCE and FID can effectively evaluate generative models of CIFAR-10. However, Figure 2(b) reveals that FID on ImageNet correlates little with human perception, while VCE aligns exceptionally well. See details in Appendix C.

Notably, diffusion models generally outperform GANs in terms of human perception, while FID tends to favor GANs (Dhariwal and Nichol, 2021). VCE effectively compensates for this limitation. Based on the visualized results in Figure 1, we observe that GANs excel at capturing the crucial features of central objects while easily overlooking the background or other details. In contrast, diffusion models focus more on the entire image, aiming to capture the whole distribution. As a result, FID is easily deceived by GANs, but VCE, evaluating virtual classifiers on the real dataset, better reflects the generative models' ability to cover the real distribution.

VCE is More Robust than FID. The neural network exhibits a certain level of robustness to high-frequency noise and other perturbations during training. Therefore, we further investigate the impact of subtle perturbations on VCE and FID. Specifically, we add noise to the generated images or round



(a) Adding Noise

(b) Rounding Pixel

Figure 3: FID *v.s.* VCE when generated images of CIFAR-10 are slightly perturbed. E denotes error, while R represents the correlation coefficient between FID/VCE and the human error rate. The suffix “*” denoted a noisy dataset with Gaussian or round noise. (a): Adding Gaussian noise (mean=0, stdev=0.01) to generated datasets. (b): Keep 1 decimal place for normalized pixel values.

the pixel values, which are transformations that are challenging for the human eye to perceive (Visual examples in Appendix E). We evaluate the correlation of human error rates and metrics and calculate errors E after perturbation. In order to avoid the impact of different scales of FID and VCE, E_M is calculated using the formula: $E_M = \frac{1}{n}\frac{\sum|m_i' - m_i|}{\max_i(m_i, m_i')}$, where m_i and m_i' denote the metric values (FID/VCE) of the i-th generative model before and after perturbation, respectively. n represents the total number of generative models considered in the experiments. As shown in Figure 3, the error of VCE is significantly lower than that of FID, and it exhibits a stronger correlation with human perception than FID after being perturbed. This indicates that our method is robust.

Table 2: We evaluate six generative models trained from ImageNet with their correlation coefficients between VCE and the human error rate (higher the better). (a): VCE with different backbones under 30-epoch training. (b): VCE with MobileNet-V2 under different training epochs.

(a) Backbones

Backbone	ResNet18	MobileNet-V2	DenseNet121
Corr.	-0.69	-0.79	-0.66

(b) Training Epochs

Epochs	20	30	40
Corr.	-0.62	-0.79	-0.65

Configuration of VCE. At last, we analyze the behavior of VCE to different backbones and training epochs. As shown in Table 2, the best alignment between VCE and human perception is achieved using MobileNetV2 as the backbone and training for 30 epochs. Therefore, we recommend using 50k generated images as the training dataset, selecting MobileNetV2 as the backbone, and training for 30 epochs when computing VCE. Overall, VCE is roughly consistent under different choices of backbones and training epochs.

4 Discussions

In this paper, we propose a new procedure to evaluate image quality by training models on generated data while evaluating them on natural test data. The resulting metric, Virtual Classifier Error, shows promising performance on benchmark datasets for evaluating state-of-the-art generative models and exhibits better alignment with human preference, while being more robust against noises.

Meanwhile, we also acknowledge that compared to FID and its counterparts, VCE has a few practice challenges. First, it requires training a model with generated data that is more computationally expensive. Regarding this issue, we show that training only for a few epochs (*e.g.,* 20 epochs on CIFAR-10) already gives a fairly good metric. Further techniques like initializing from a pretrained model, or using advanced training techniques, may further improve its efficiency. Second, the training of VCE may depend on various hyperparameters (such as, learning rates, optimizers). As shown above, these choices do not have much influence on the final performance among a proper range, and one may still develop a benchmark metric with a standard training recipe. Therefore, despite these challenges, we believe that VCE still has the potential to become a promising alternative or compliment to existing image evaluation criteria.

Acknowledgement

Yisen Wang was supported by National Key R&D Program of China (2022ZD0160300), National Natural Science Foundation of China (62376010, 92370129), Beijing Nova Program (20230484344), and CCF-BaiChuan-Ebtech Foundation Model Fund.

References

Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9), 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Chunsan Hong, Byunghee Cha, Bohyung Kim, and Tae-Hyun Oh. Enhancing classification accuracy on limited data via unconditional gan. In *CVPR*, 2023.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11).

Minguk Kang, Woohyeon Shim, Minsu Cho, and Jaesik Park. Rebooting acgan: Auxiliary classifier gans with stable training. *Advances in neural information processing systems*, 34:23505–23518, 2021.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

Memoona Khanum, Tahira Mahboob, Warda Imtiaz, Humaraia Abdul Ghafoor, and Rabeea Sehar. A survey on unsupervised machine learning algorithms for automation, classification and maintenance. *International Journal of Computer Applications*, 119, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Shuangliang Li, Siwei Li, and Lihao Zhang. Hyperspectral and panchromatic images fusion based on the dual conditional diffusion models. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 2023.

Daniel Mas Montserrat, Carlos Bustamante, and Alexander Ioannidis. Class-conditional vae-gan for local-ancestry simulation. *arXiv preprint arXiv:1911.13220*, 2019.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *CVPR*, 2023.

Andrés D Pérez, Oscar Perdomo, Hernán Rios, Francisco Rodríguez, and Fabio A González. A conditional generative adversarial network-based method for eye fundus image quality enhancement. In *International Workshop on Ophthalmic Medical Image Analysis*, 2020.

Sai Saketh Rambhatla and Ishan Misra. Selfeval: Leveraging the discriminative nature of generative models for evaluation. *arXiv preprint arXiv:2311.10708*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 2022.

George Stein, Jesse C Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, J Eric T Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *arXiv preprint arXiv:2306.04675*, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

Ryan Turner, Jane Hung, Eric Frank, Yunus Saatchi, and Jason Yosinski. Metropolis-hastings generative adversarial networks. In *ICML*, 2019.

Lokesh Veeramacheneni, Moritz Wolter, and Juergen Gall. Wavelet packet power spectrum kullback-leibler divergence: A new metric for image synthesis. *arXiv preprint arXiv:2312.15289*, 2023.

Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, and Tommi Jaakkola. PFGM++: Unlocking the potential of physics-inspired generative models. *arXiv preprint arXiv:2302.04265*, 2023.

Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. Hype: A benchmark for human eye perceptual evaluation of generative models. In *NeurIPS*, 2019.

A EXPERIMENTAL DETAILS

During the training of the classifier for VCE, we utilize a dataset consisting of 50,000 generated images. The classifier is trained for 30 epochs using the SGD optimizer with an initial learning rate of 0.1 and batch size of 128.

B OMITTED PROOF

B.1 PROOF OF THEOREM 2.1

Proof.

$$\begin{aligned}
&D_{\mathrm{KL}}(P_d(Y|X)\|P_g(Y|X))\\
=&\mathbb{E}_{P_d(X,Y)}\log\frac{P_d(Y|X)}{P_g(Y|X)}\\
=&\mathbb{E}_{P_d(X,Y)}\log\frac{P_d(Y|X)P_d(X)P_g(X)}{P_g(Y|X)P_d(X)P_g(X)}\\
=&\mathbb{E}_{P_d(X,Y)}\log\frac{P_d(X,Y)}{P_g(X,Y)}-\mathbb{E}_{P_d(X,Y)}\log\frac{P_d(X)}{P_g(X)}\\
=&D_{\mathrm{KL}}(P_d(X,Y)\|P_g(X,Y))-D_{\mathrm{KL}}(P_d(X)\|P_g(X))\\
\leq&D_{\mathrm{KL}}(P_d(X,Y)\|P_g(X,Y))=D_{\mathrm{KL}}(P_d(X|Y)\|P_g(X|Y)).
\end{aligned}$$

□

C ADDITIONAL RESULS

Table 3 presents the results of VCE and FID on CIFAR-10, while Table 4 displays the results of VCE and FID on ImageNet. The result of human error sourced from Stein et al. (2023).

Table 3: VCE and FID of generative models on CIFAR-10.

	PFGM++	ADA	BigGAN	MHGAN	ReACGAN	ACGAN
Human Error Rate	43.58	39.3	38.68	33.62	33.56	14.82
VCE	9.92	12.9	14.79	13.88	15.22	52.87
FID	1.81	2.53	3.87	4.21	4.4	35.47

Table 4: VCE and FID of generative models on ImageNet.

	StyleGAN-XL	LDM	ADMG-ADMU	Dit-XL	BigGAN	ADM
Human Error Rate	15.34	30.88	26.88	28.62	15.8	26.64
VCE	77.59	74.49	77.36	75.534	79.76	77.61
FID	2.53	4.13	4.16	2.92	7.34	10.27

D RELATED WORKS

Due to the generative model reflects the learned distribution implicitly, we can only indirectly assess the quality of the generated model by sampling from the distribution, and cannot efficiently evaluate the likelihood. This poses significant challenges in evaluating generative models. Inception Score uses Inception V3 to classify generated images, while FID calculates the Wasserstein distance between real and generated images using Inception V3's feature representations. But, their results do not align with human perception (Stein et al., 2023) and tend to favor giving higher scores to GANs compared to diffusion models, despite the higher image quality achieved by diffusion models

(Dhariwal and Nichol, 2021). These methods are also highly sensitive to tiny differences. Performing operations such as resizing or compressing images (Parmar et al., 2022), rounding pixel values (Veeramacheneni et al., 2023), or adding subtle noise can cause FID to change greatly. Additionally, images which almost look like noise can also achieve very desirable IS and FID when compared to real images (Barratt and Sharma, 2018; Veeramacheneni et al., 2023). Overall, the approach of utilizing a pretrained classifier has several shortcomings. In contrast, our proposed method of employing a Virtual Classifier not only aligns more closely with human perception but also exhibits robustness to subtle variations.

E Visualization of the images

For a concrete understanding, we provide perturbed examples of the generated images from BigGAN. It is evident that despite the imperceptibility of the perturbations to human perception, there is a substantial change in FID while VCE remains relatively unchanged. This observation highlights the robustness of our proposed method.

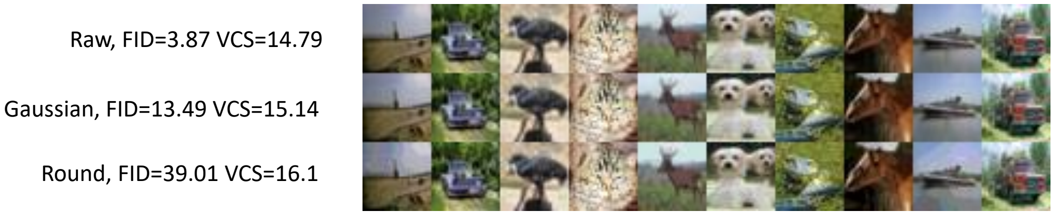


Figure 4: Examples of perturbed generated images from BigGAN.