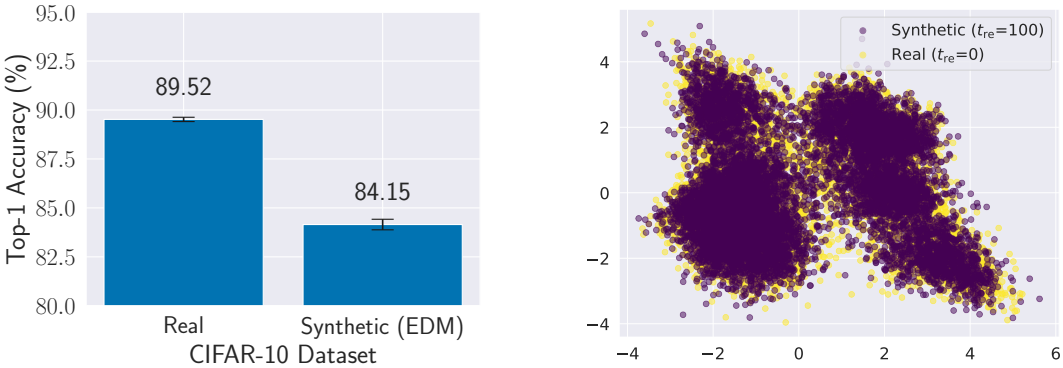
On the Limitation of Diffusion Models for Synthesizing Training Datasets

Shin'ya Yamaguchi[1,2]* **Takuma Fukuda**[3]†
[1]NTT [2]Kyoto University [3]Chiba University

Abstract

Synthetic samples from diffusion models are promising for leveraging in training discriminative models as replications of real training datasets. However, we found that the synthetic datasets degrade classification performance over real datasets even when using state-of-the-art diffusion models. This means that modern diffusion models do not perfectly represent the data distribution for the purpose of replicating datasets for training discriminative tasks. This paper investigates the gap between synthetic and real samples by analyzing the synthetic samples reconstructed from real samples through the diffusion and reverse process. By varying the time steps starting the reverse process in the reconstruction, we can control the trade-off between the information in the original real data and the information added by diffusion models. Through assessing the reconstructed samples and trained models, we found that the synthetic data are concentrated in modes of the training data distribution as the reverse step increases, and thus, they are difficult to cover the outer edges of the distribution. Our findings imply that modern diffusion models are insufficient to replicate training data distribution perfectly, and there is room for the improvement of generative modeling in the replication of training datasets.



(a) Top-1 Test Accuracy on CIFAR-10

(b) Feature Visualization (PCA) on ResNet-18

Figure 1: Our motivation and finding. (a): Synthetic datasets produced by a modern diffusion model (EDM [1]) degrade the classification performance when solely using them for training classifiers. (b) We input synthetic samples to a classifier trained on a real dataset, and found that the features of synthetic samples concentrate on the modes of real feature distribution and do not cover the outer edge of the distribution. This means that diffusion models are still insufficient to replicate datasets for training classifiers.

*Corresponding author. Email: `shinya.yamaguchi@ntt.com`
†Work done at NTT as an intern.

NeurIPS 2023 SyntheticData4ML Workshop

1 Introduction

In the past decade, deep generative models have witnessed remarkable advancements in generating high-quality synthetic samples that are human-indistinguishable from real data. Among these generative models, diffusion models [2] have attracted much attention because they can outperform the existing generative models (e.g., GANs [3] and VAEs [4]) by learning reverse (denoising) processes through score-based likelihood maximization [5, 6].

The high-quality samples from diffusion models naturally raise research interest in their applicability for training target discriminative models (e.g., classifiers), and recent studies intensively develop training techniques utilizing synthetic samples from diffusion models. For instance, He et al. [7] demonstrated that synthetic samples from text-image diffusion models (e.g., Stable Diffusion [6]) can achieve impressive zero-/few-shot learning performance by querying the synthetic training samples with crafted prompts representing target dataset categories. Moreover, Burg et al. [8], Azizi et al. [9], and Dunlap et al. [10] highlighted the potential of diffusion models for data augmentation application. They investigated diffusion-based data augmentation methods by modifying diffusion models with nearest neighbor exploration in sampling [11], scaling up models [9], and customizing text prompts for querying samples [10]. However, in contrast to these remarkable successes, we observed that training models solely on naïvely generated synthetic samples leads to inferior performance compared to models trained on real data (Figure 1a). This indicates that even state-of-the-art diffusion models are not able to replicate the entire training data distribution, and there is a gap between real and synthetic datasets in terms of training classifiers. In this paper, by analyzing diffusion models, we aim to answer the following important and open research question: *What is the cause of the gap between real and synthetic datasets?*

This paper analyzes the gap between synthetic and real datasets from two perspectives: (i) the quality of synthetic samples and (ii) the impact of synthetic samples on training classification models. To assess the gap, we introduce the concept of **real sample reconstruction** by focusing on the diffusion and reverse process. Real sample reconstruction involves corrupting real samples by the diffusion process up to pre-defined steps and then restoring the corrupted samples by the reverse process; We refer to the pre-defined step as the reverse step. By varying the reverse steps, we can continuously control the trade-off between the remaining information from the input real samples and the synthetic information injected by the reverse process (Figure 2). We empirically investigate how the synthetic information affects the sample quality and the classification performance.

Our experimental findings are summarized as follows:

- Diffusion models generate synthetic samples that are nearly indistinguishable as real or fake compared to competitive models such as GANs.
- Increasing the reverse steps (i.e., making sample properties closer to synthetic samples) leads to gradual degradation in the sample quality and classifier performance.
- Leveraging the synthetic samples for training classifiers does not adversely affect the tendency of classifier outputs (e.g., attention map).
- Synthetic samples are easier to classify than real samples.
- The synthetic samples concentrate near the modes of the data distribution in the feature space of classifiers (Fig. 1b), and a larger reverse step brings the sample closer to the mode.

These findings suggest that modern diffusion models have limitations in generating samples away from the modes. This can be explained by the fact that diffusion models learn to denoise samples in the direction that maximizes the likelihood at each step in the reverse process. That is, the reverse process may bring the sample closer to the typical mode. Therefore, replicating the training dataset by diffusion models can result in accuracy degradation due to the less information far from the modes.

2 Related Work

Diffusion model [12, 2] is a class of generative models inspired by thermodynamics. They learn iteratively denoising process called *reverse process* corresponding to the corruption process adding noises called *diffusion process*. Song et al. [13] revealed the relationship between diffusion models and denoising score matching with stochastic gradient Langevin dynamics and explained optimization

Algorithm 1 Real Sample Reconstruction

Require: Real sample x, reverse step $t_{\mathrm{re}} > 1$
Ensure: Reconstructed sample $\hat{x}$

```
// Corrupting x with diffusion process for t_re
x_0 ← x
for t = 1, ··· , t_re do
    x_t ← √α_t x_{t-1} + √(1 − α_t) ε_{t-1}
end for
// Restoring x_{t_re} with reverse process
for t = t_re, ··· , 1 do
    x̂_{t-1} ← (1/√α_t) ( x_t − ((1−α_t)/√(1−ᾱ_t)) ε_θ(x_t, t) + σ_t z )
end for
x̂ ← x̂_0
```

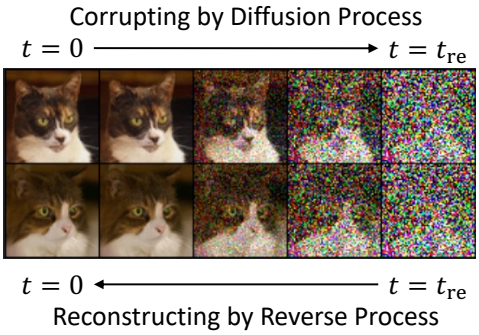


Figure 2: Real Sample Reconstruction

of the reverse process as score-based likelihood maximization. By introducing conditional guidance in the reverse process, diffusion models successfully control output by class labels [5, 14] and text embedding [15, 6], and a number of subsequent studies are still being published.

Since diffusion models can achieve high-quality synthetic samples in comparison to other generative models (e.g., GANs and VAEs) [5], recent studies investigated the capability of diffusion models as a source of training datasets [7, 8, 9, 10]. These studies utilized text-image pre-trained diffusion models such as Stable Diffusion [6] for generating synthetic training samples and succeeded in improving classification performance by adding the synthetic samples into training datasets. In contrast, this paper focuses on class conditional diffusion models trained only on target datasets for discriminative tasks and does not consider large pre-trained diffusion models to discard the effects of knowledge transfer from external pre-trained datasets.

3 Preliminary

Here, we briefly introduce the principles of diffusion models and real sample reconstruction, which is used for our main analysis.

3.1 Diffusion Models

A diffusion model learns a data distribution $p(x)$ by optimizing the parameterized reverse (denoising) process assuming Markov Chain with length T [12, 2], which corresponds to the forward diffusion process. Specifically, most modern diffusion models are optimized by minimizing the family of the following loss function with respect to the neural network parameter θ [2, 5, 6].

$$\mathcal{L}(\theta) = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right], \tag{1}$$

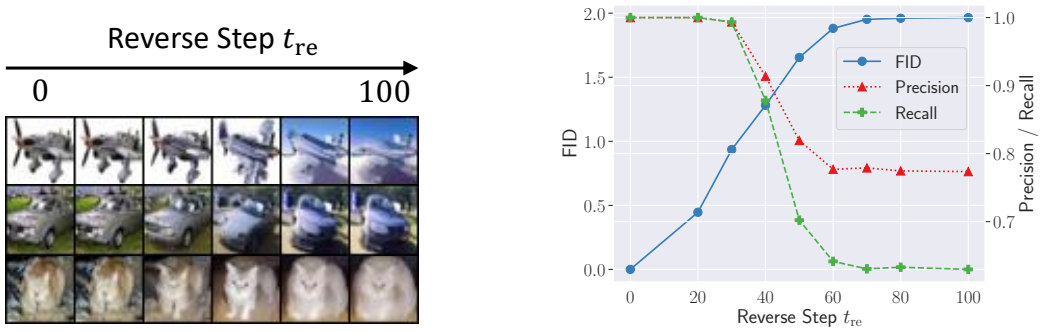
where ϵ_θ is the denoising autoencoder parameterized by θ, t is the time step randomly sampled from $\{1, \cdots, T\}$, x is the input, and x_t is a noisy variant of x. In inference time, a synthetic sample $\hat{x}$ is generated by sequentially applying the denoising function for each t from T to 1 as

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t) + \sigma_t z\right), \tag{2}$$

where $\alpha_t = 1 - \beta_t$, β_t is a scheduled variance in $\{\beta_1, \cdots, \beta_T\}$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, $\sigma_t = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}$ and $z \sim \mathcal{N}(0, 1)$. Song et al. [13] showed that this denoising process corresponds to stochastic gradient Langevin dynamics, which produces samples by iterative updating x_t with the score $\nabla_x \log p(x)$:

$$x_t = x_{t-1} + \frac{\delta}{2}\nabla_x \log p(x_{t-1}) + \sqrt{\delta}z, \tag{3}$$

where δ is a step size. In this paper, we implement diffusion models with conditional EDM [1] to generate a synthetic labeled dataset for training classifiers.



(a) Reconstructed Samples

(b) FID [19] and Precision/Recall [20]

Figure 3: Quality Assessments of Synthetic Sample

3.2 Real Sample Reconstruction

We introduce real sample reconstruction, which produces intermediate samples between real and synthetic by exploiting the diffusion and reverse process. Real sample reconstruction first corrupts the real samples by the diffusion process from 0 to a specified time step t_{re}, and then recovers the corrupted samples by the reverse process from t_{re} to 0. Given a real data point x, we produce a reconstructed sample $\hat{x}$ with a reverse time step t_{re} by following Algorithm 1. This reconstruction algorithm is similar to SDEdit [16], which is an image-editing method based on the reconstruction of guide images by diffusion models. Intuitively, $\hat{x}$ is fully real when $t_{\mathrm{re}} = 0$, a fully synthetic when $t_{\mathrm{re}} = T$, and half of real and synthetic when $t_{\mathrm{re}} = T/2$ as depicted in Figure 2. Unlike the purpose of SDEdit, we aim to produce intermediate samples of real and synthetic by simply inputting real images into the diffusion and reverse process.

4 Analysis

In this section, we report the experimental results assessing (i) the quality of reconstructed samples from diffusion models and (ii) the effects on classifiers trained on the reconstructed samples. We used the CIFAR-10 dataset [17] as the target dataset, the CIFAR-10 pre-trained conditional EDM [1] ($T = 100$) as the diffusion model, and ResNet-18 [18] as the classifier architecture.

4.1 Analysis on Synthetic Sample

Evaluation Protocol. To analyze reconstructed synthetic samples, we measured Frechèt inception distance (FID) [19], precision/recall [20], and fake detection accuracy [21]. Among them, FID and precision/recall are measured on the ImageNet pre-trained feature extractor. FID evaluates the gap between real and synthetic datasets, and precision/recall evaluate the probabilities that synthetic/real samples fall within the real/synthetic distributions. Fake detection accuracy is calculated on a classifier trained to distinguish real and synthetic samples on both the pixel and frequency domains. This is useful to find out how different synthetic and real samples are in terms of input to the classifier. We used 50,000 synthetic samples and 50,000 real samples to calculate the metrics.

Sample Quality. We first show the visualization of the reconstructed samples in Figure 3a. We reconstructed the samples on every 10 steps of $t_{\mathrm{re}} \in [20, 80]$ from EDM. As the reverse step t_{re} increases, the reconstructed samples gradually lose information on the input real sample, and represent information on the synthetic sample. Nevertheless, in visual quality, it is hard to distinguish between a synthetic sample and a real sample for every t_{re}. Next, we show the FID and precision/recall scores calculated on the reconstructed samples in Figure 3b. We see that increasing reverse steps progressively degrades all the quantitative metrics. This indicates that the reverse process may be harmful to maintain the information on the real samples. In particular, the recall scores are significantly degraded by the reverse process, indicating that the synthetic sample does not sufficiently cover the training data distribution.

Fake Detection Accuracy. We demonstrate the fake detection accuracy on the synthetic samples. To evaluate the worst quality case, we used $t_{\mathrm{re}} = 100$ in this experiment. Table 1 shows the fake detection accuracy in the pixel domain and frequency domain (DCT). For comparison, we also print the result of StyleGAN3 [22]. The higher scores mean easier samples to be detected as fake. While the StyleGAN3 samples were easily distinguished by the classifier, fewer EDM samples were detected as fake. These results suggest that although the synthetic datasets from diffusion models differ in quantitative measures such as FID, their properties as input to the classifier are almost the same as those of the real samples.

Table 1: Fake Detection Accuracy (CIFAR-10)

Generative Model	Accuracy (%)	
	Pixel	DCT
StyleGAN3 [22]	89.56	53.62
EDM [1]	56.15	58.91

Table 2: Top-1 Test Accuracy on CIFAR-10

Reverse Step t_{re}	Test Acc. (%)
0 (Fully Real)	$89.52^{\pm.11}$
30	$88.84^{\pm.08}$
50	$87.41^{\pm.32}$
70	$85.76^{\pm.49}$
100 (Fully Synthetic)	$84.15^{\pm.27}$

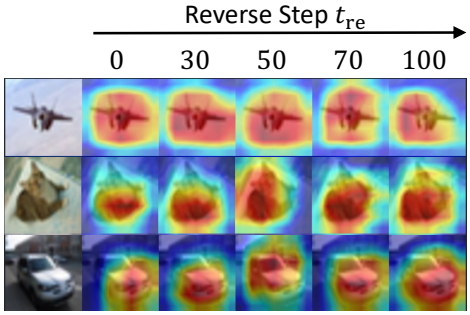


Figure 4: GradCAM Visualization

4.2 Analysis on Training Classifiers

Evaluation Protocol. We analyze trained classifiers on reconstructed synthetic samples by varying the reverse step t_{re}. We evaluate test classification accuracy, attention map visualization by Grad-CAM [23], output entropy, and feature visualization by principle component analysis (PCA). We trained ResNet-18 classifiers for 100 epochs on the synthetic CIFAR-10 datasets yielded by real sample reconstruction with $t_{\mathrm{re}} = 30, 50, 70$, and tested them on the real CIFAR-10 test set. We used the SGD optimizer with a learning rate of 0.01 dropping by multiplying 0.1 for every 30 epochs. We also show the results when using the real dataset (i.e., $t_{\mathrm{re}} = 0$) and the fully synthetic dataset by the reverse process with random noise (i.e., $t_{\mathrm{re}} = 100$). For GradCAM and feature visualization, we used the output of `block4` on ResNet-18. We calculate the marginal output entropy by

$$H_\theta(y) = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{C} p_\theta(y=j|x_i)\log(p_\theta(y=j|x_i)), \tag{4}$$

where N is a dataset size, C is a class number, $p_\theta(y=j|x_i) = \frac{\exp(f_\theta(x_i)[j])}{\sum_k^C \exp(f_\theta(x_i)[k])}$, f_θ is a classifier.

Classification Performance. Table 2 shows the top-1 test accuracy on the real CIFAR-10 test set for each reverse time step t_{re}. Similar to the sample quality shown in the previous section, we see that the performance of the classifier degrades as the reverse step increases. This implies that the reverse process of diffusion models eliminates information important for solving classification tasks from the original real sample.

Attention Map. Figure 4 shows the visualizations of GradCAM. We input real test samples of CIFAR-10 for each trained model. Interestingly, while the test accuracy is degraded by real sample reconstruction, the synthetic samples used for training do not change the attention of the trained models. This means that the synthetic sample itself has no noticeable negative impact on learning classification tasks.

Output Entropy Next, we assess the quantitative effects on the classifier's prediction. To this end, we used the classifier trained on the real CIFAR-10 because we can consider it an ideal classifier for the purpose of training dataset replication. Figure 5 plots the output entropy $H_\theta(y)$. We calculated $H_\theta(y)$ by inputting the synthetic samples into the classifier. In Figure 5, `Train Samples` and `Test Samples` mean the calculated entropy scores on the reconstructed samples from real samples of

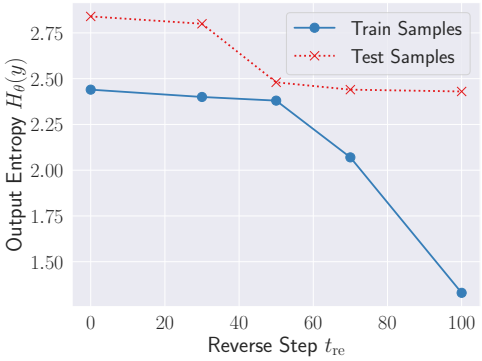


Figure 5: Output Entropy of Classifier

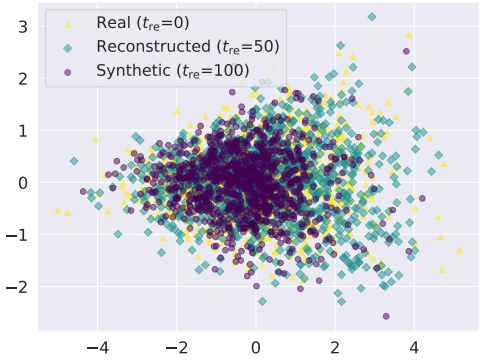


Figure 6: Feature Visualization (`truck` class)

the train/test set. Note that again, in this experiment, we used only the classifier trained on the real CIFAR-10 to assess the characteristics of synthetic samples. We see that increasing the reverse step makes the synthetic samples low entropy, indicating easy samples to be classified. Thus, the diffusion model tends to produce a typical sample that is representative of the class by the reverse process.

Feature Visualization. Finally, we visualize the features of synthetic samples to examine how synthetic samples behave on the classifier. Similar to the previous paragraph, we used the classifier trained on the real CIFAR-10 for feature visualization. We applied PCA to the extracted features of input synthetic samples and reduced the dimension to two. Figure 1b and 6 are the visualization results of all class samples and `truck` class samples, respectively. In Figure 1b, the synthetic sample is concentrated inside the distribution formed by the real samples, while its outer edges are not well covered. Meanwhile, in Figure 6, the reconstructed at $t_{\mathrm{re}} = 50$ appears to cover the region where the sample at $t_{\mathrm{re}} = 100$ is scarce. These results suggest that the synthetic samples from diffusion models tend to concentrate the center (mode) of training data distribution, and the reverse process gradually pulls the synthetic samples toward the modes of training distribution.

5 Discussion

Through the empirical analysis in the previous section, we observed that the modern diffusion models can produce quite realistic synthetic samples, but they still have insufficient generative performance for replicating training datasets for classifiers. In particular, the reverse process of diffusion models seems to gradually concentrate the synthetic samples toward the modes of the training data distribution. This can be explained by the interpretation of the diffusion model as a score-based generative model. As we discussed in Sec. 3.1, a reverse process corresponds to a step of stochastic gradient Langevin dynamics as shown in Eq. 3. That is, a reverse step contains the gradient of log-likelihood (score) $\nabla_x \log p(x)$. Therefore, the iterative denoising of samples by multiple reverse steps means that the samples are moving closer to the region of greater likelihood, i.e., the mode of the distribution. Eq. (3) has a disturbance term by z to prevent the concentration of sampling near the modes, but our experimental results suggest that this cannot be completely prevented for the purpose of replicating training datasets.

6 Conclusion and Takeaway

This paper empirically showed the limitations of diffusion models for synthesizing datasets for training classifiers. Modern diffusion models are not sufficient to replicate entire training datasets due to the sampling concentration near the data distribution modes. This can be caused by the reverse denoising process, which naturally moves the samples toward the modes. From these observations, one of the important takeaways is that we should improve diffusion models to cover the outside edges of training data distributions. Another one is that, currently, the data augmentation applications of diffusion models, which utilize both real and synthetic samples, can be more suitable to train high-performance classifiers than replicating entire training datasets and utilizing only synthetic samples. We believe that these observations and implications will be helpful for future research.

References

[1] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in neural information processing systems*, 2022.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, 2020.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, 2014.

[4] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[5] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[7] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

[8] Max F Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. A data augmentation perspective on diffusion models and retrieval. *arXiv preprint arXiv:2304.10253*, 2023.

[9] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

[10] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *arXiv preprint arXiv:2305.16289*, 2023.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.

[12] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2015.

[13] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representation*, 2021.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[16] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representation*, 2022.

[17] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

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

[19] 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 *Advances in Neural Information Processing Systems*, 2017.

[20] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 2019.

[21] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. 2020.

[22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 2021.

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 2017.