

miniGLIDE: Towards a Photo-Realistic Text to Image Model With a Reduction in Parameters

Aaron Lozhkin
Rutgers University
al1336
al1336@rutgers.edu

Srinihar Bondalapati
Rutgers University
sb1686
sb1686@rutgers.edu

Jinal Shah
Rutgers University
js2865
js2865@rutgers.edu

Abstract

A big milestone for image generation has been the production of neural networks that are capable of producing images from text. A class of image generation models known as guided diffusion models have been proven to generate high-quality synthetic images. In this paper, we seek to reconstruct [Nichol et al. \(2021\)](#)’s GLIDE model while reducing parameter size. We present two models: miniGLIDE a general text to image model trained on the sbucaptions dataset ([Ordonez et al., 2011](#)) and dogGLIDE which was trained on the Stanford Dog Dataset ([Khosla et al., 2011](#)). We investigate how improvements to the diffusion training loop can lead to significantly better image quality in miniGLIDE, along with detailing how classifier free guidance impacts image accuracy in dogGLIDE. Given enough training time and data, we believe that locally learned guided diffusion models are more than attainable. [Our code can be found here on Github.](#)

1 Introduction

Recently, text to image models have risen in popularity due to their generalizability and use cases. The ability to create images based on textual information is valuable and text to image models have the potential to revolutionize the way we generate and use visual content. Training text to image models usually consists of a plethora of image-caption pairs and requires immense processing power. Previous solutions that have been presented for this problem include Generative Adversarial Networks ([Xu et al., 2017](#); [Zhou et al., 2021](#)) and transformer based models ([Ramesh et al., 2021](#)), but these approaches struggle to generate photo-realistic outputs.

Diffusion models have emerged as a more photo-realistic approach that are able to generate high-quality, synthetic images ([Nichol and Dhariwal,](#)

[2021](#); [Nichol et al., 2021](#); [Dhariwal and Nichol, 2021](#)). It has been shown that guidance techniques such as classifier-free guidance can help enhance image quality in exchange for image diversity. Our study seeks to investigate the GLIDE model as presented by [Nichol et al. \(2021\)](#) and explore the problem of text-conditional image synthesis through parameter reduced diffusion models using classifier free guidance¹.

2 Background and Related Work

2.1 Diffusion Models

During the training loop for diffusion models, we add some noise, $\epsilon \sim N(0, I)$ sampled from a normal distribution, to the starting image, x_0 , creating a noised image, x_t . This process of progressively adding Gaussian noise creates a Markov chain of latent variables that can be modeled as $q(x|x_t) := N(x_t; \sqrt{a} * x_{t-1}, (1 - \alpha_t) * I)$. We can subsequently learn a model of the form $p_\theta(x_{t-1}|x_t) := N(\mu_\theta(x_t), \sum_\theta(x_t))$ that approximates the true posterior of the image at t-1 before the noise was added to get to step t ([Ho et al., 2020](#)). We can model the loss as,

$$L_{simple} := E_{t \sim [1, T], x_0, \epsilon \sim N(0, I)} [||\epsilon - \epsilon_\theta(x_t, t)||^2]$$

where $\epsilon_\theta(x_t, t)$ is the models prediction of ϵ at time t. After enough diffusion steps, all the points making up the image are comparable to a normal noise distribution. Subsequently in the backward pass, we start with a random normal distribution and sample from the learned model to repeatedly estimate an image and its noise, as it was at its respective time position. This is done from step T to step 0.

¹Although [Nichol et al. \(2021\)](#) also utilizes CLIP guidance for image generation (and we did as well), we decided to omit our results from this paper as we did not train our own noisy CLIP model.

2.2 Improving Diffusion Models

Dhariwal and Nichol (2021) improved the training loss of their guided diffusion model by implementing a hybrid loss that trained the covariance matrix, \sum_{θ} , as oppose to Ho et al. (2020) which kept covariance constant.

$$L_{hybrid} = L_{simple} + \lambda L_{vlb}$$

where L_{vlb} is the loss of the variational lower bound and can be thought of as the KL Divergence between q and p , and λ is a small constant (1/1000) so as to not impact L_{simple} . We test the results of this improved loss between two iterations of our model in Table 3.

2.3 Guided Diffusion

Following the baseline diffusion model, we utilize the concept of guided diffusion, specifically classifier-free guidance (Ho and Salimans, 2021). Classifier free guidance is utilized during sampling and can be represented as,

$$\hat{\epsilon}_{\theta}(x_t|c) = \epsilon_{\theta}(x_t|\emptyset) + s \cdot (\epsilon_{\theta}(x_t|c) - \epsilon_{\theta}(x_t|\emptyset))$$

where $\epsilon_{\theta}(x_t|\emptyset)$ is the model’s prediction of the noise given no text information, \emptyset , and $\epsilon_{\theta}(x_t|c)$ is the model’s prediction of the noise given a text embedding, c . Here, s represents the guidance scale, which determines how far to push the model in the direction of the text embedding.

The mechanics behind classifier free guidance operate based on an internal representation of the models’ learned understanding of concepts and text embeddings. As the number of (image, caption) pairs increases, the model’s intuition behind what a certain caption should look like grows in tandem.

3 Method

3.1 Data Pipeline

For data preprocessing, we ”squished” our images down to a 64 x 64 resolution and scaled all the values such that the red green ,and blue (RGB) values would be between -1 and 1. Such a scaling allows for the model to train faster since computations would be performed on smaller values.

3.2 Model Training

Due to computational limitations, our model went through many progressive iterations before being optimized for training time and image quality. We

present two models of the same size: one trained on 85,000 images from the sbucaptions dataset with 100 diffusion steps (miniGLIDE) and another on 20,580 images from the Stanford dog dataset with 1000 diffusion steps (dogGLIDE). Both are text-conditional diffusion models at 64x64 resolution and consisted of 1.5 million parameters. This is in stark contrast to the 3.5 billion parameter model trained in the original GLIDE paper (Nichol et al., 2021).


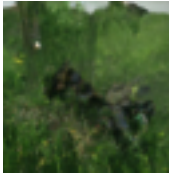
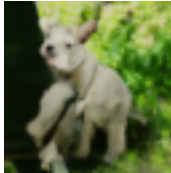
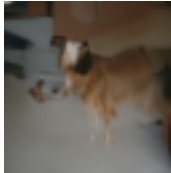
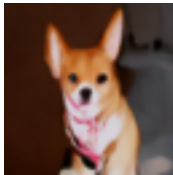
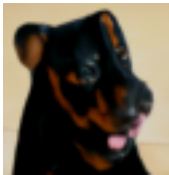
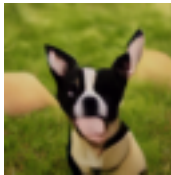
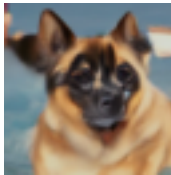
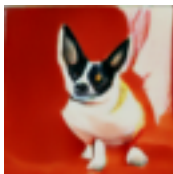

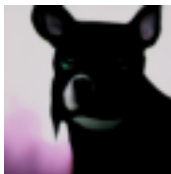
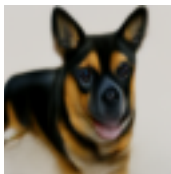
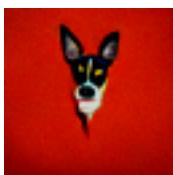

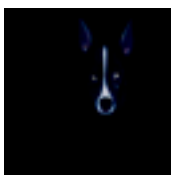

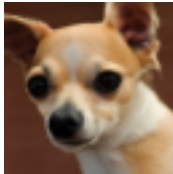
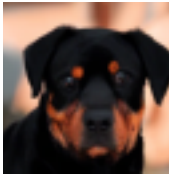
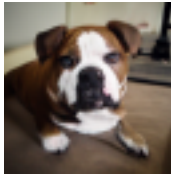

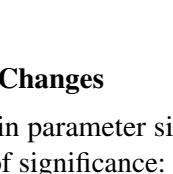
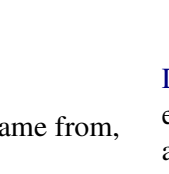
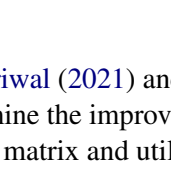
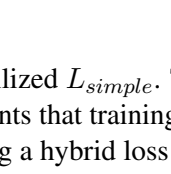
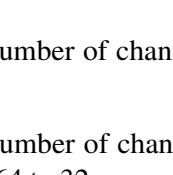
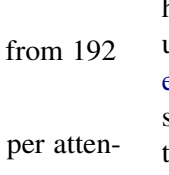
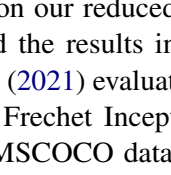
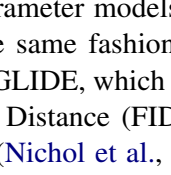
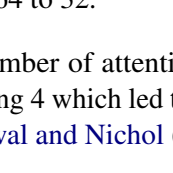
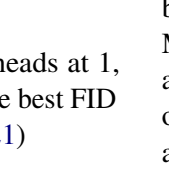
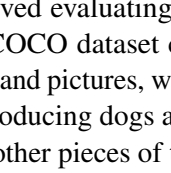
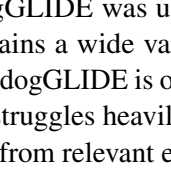
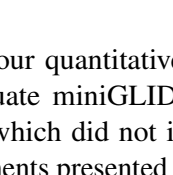
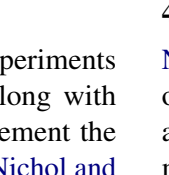
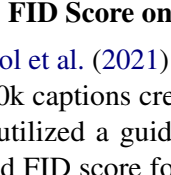
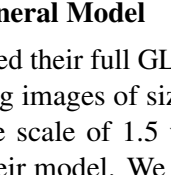








Since miniGLIDE was only able to train on approximately 85,000 (image, caption) pairs with a small batch size, we were limited by space and time to create a more general model that understood a wide variety of concepts. Thus, we decided to dedicate time to training dogGLIDE whose training set consisted of images and captions of standard household dog breeds. Our reasoning for choosing the Stanford dog dataset is because it consists of only 120 captions (unique dog breeds) which allowed classifier free guidance to have a much better understanding of the variety of images that can come from a single caption.

miniGLIDE was trained on iLAB for 36 hours with a batch size of 12 and elapsed 36 epochs, while dogGLIDE was trained on a local RTX 4700ti for 38 hours with a batch size of 8 and elapsed 125 epochs. We utilized the AdamW optimizer with a learning rate of 0.0001 for both models.

3.3 Model Architecture

Ho et al. (2020) implemented a stable diffusion architecture that we employ here in tandem with a text augmentation method proposed by Nichol et al. (2021). The main model which is able to predict noise added for a given diffusion step, t , is an ablated diffusion model (ADM) that is built upon the UNET architecture (Dhariwal and Nichol, 2021). The ADM model increases the number of attention heads, uses a variety of attention resolutions, and increases depth compared to the original UNET architecture (Dhariwal and Nichol, 2021). Nichol et al. (2021) found that the ADM model can be augmented with text by passing in tokens into a transformer model which outputs embeddings. These embeddings are then appended to each attention layer throughout the ADM along with being used in place of the class embeddings.

Table 1: Comparing different guidance values for Classifier Free Guidance on dogGLIDE. As we increase the guidance value, we approach the models understanding of what that specific dog breed should look like. Samples from the original GLIDE paper are included as well with a guidance scale of 3.0.

Guidance Scale	Chihuahua	Rottweiler	French Bulldog	German Shepherd
1.0				
				
				
				
3.0				
				
				
				
6.0				
				
				
12.0				
3.0 GLIDE				

3.4 Architectural Changes

The main reduction in parameter size came from, in decreasing order of significance:

- Changing the number of channels from 192 to 96
- Changing the number of channels per attention head from 64 to 32.
- Keeping the number of attention heads at 1, as oppose to using 4 which led to the best FID score by [Dhariwal and Nichol \(2021\)](#)

4 Experiments

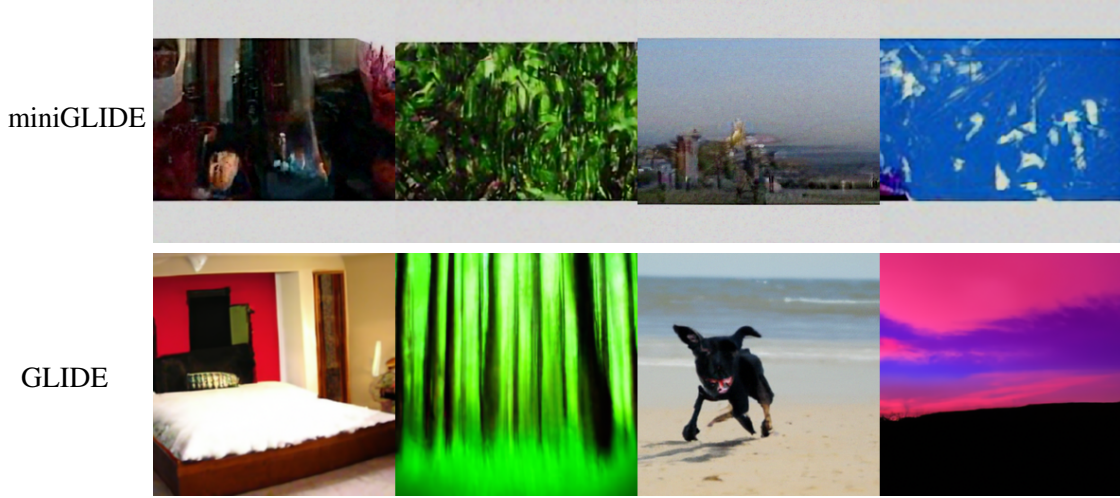
For the purpose of our quantitative experiments we decided to evaluate miniGLIDE along with miniGLIDESimple which did not implement the number of improvements presented by [Nichol and](#)

[Dhariwal \(2021\)](#) and utilized L_{simple} . This was to examine the improvements that training the covariance matrix and utilizing a hybrid loss calculation had on our reduced parameter models. We evaluated the results in the same fashion as [Nichol et al. \(2021\)](#) evaluated GLIDE, which tested zero-shot Frechet Inception Distance (FID) score on the MSCOCO dataset ([Nichol et al., 2021](#)). We believed evaluating dogGLIDE was unfair as the MSCOCO dataset contains a wide variety of images and pictures, while dogGLIDE is only capable of producing dogs and struggles heavily to extract any other pieces of text from relevant embeddings.

4.1 FID Score on General Model

[Nichol et al. \(2021\)](#) tested their full GLIDE model on 30k captions creating images of size 256x256 and utilized a guidance scale of 1.5 which optimized FID score for their model. We tested both

Table 2: Comparing different samples on 4 prompts using miniGLIDE and GLIDE. From left to right the inputted captions were 'Bedroom with a sleeping mother', 'A green forest', 'Dogs playing on the beach', and 'A dark purple landscape with a red sky'. Upsampled to 256x256.



models on 5k captions with a guidance scale of 3.0 as anything lower than 2.0 became unrecognizable relative to the original text captions when using miniGLIDE. Our model is currently only capable of producing images of size 64x64, so we decided to upsample them using the provided upsampler by Nichol et al. (2021) to match the experiment as closely as possible.

Table 3 shows that the iterations and improvements done by Dhariwal and Nichol (2021) lead to a 56.4 point decrease in FID score, which is a great improvement. We suspect that this mostly had to do with learning the covariance matrix \sum_{θ} and utilizing L_{hybrid} . This is because the covariance matrix is heavily utilized during sampling, so learning it led to much better quality samples that were also closer to the intended image.

Table 3: A comparison of FID on MSCOCO 256x256. Sampled with 5k captions with guidance scale equal to 3.0.

Model	Zero-Shot FID ↓
DALL-E (Ramesh et al., 2021)	28
LAFITE (Zhou et al., 2021)	26.94
GLIDE (Nichol et al., 2021)	12.24
miniGLIDEsimple	202.17
miniGLIDE	145.77

4.2 Qualitative Results

The first of our qualitative experiments can be seen in Table 1 where we compare various guidance

scales on dogGLIDE. Increasing guidance scales seems to work well up until 3.0 or 6.0. Afterwards, the results seem to be very color dominated or distorted. We included 4 samples from GLIDE (Nichol et al., 2021) as a reference to the ability of dogGLIDE. The second of our qualitative experiments can be seen in 2 and is comparing miniGLIDE to GLIDE, which did not see as spectacular of results. miniGLIDE struggled to understand certain concepts, which allowed other portions of the caption to completely overtake the image. For example, the fourth image was weighted very heavily on the word "sky" and the third on "beach". However, it did see some great results with the second image trying to generate "A green forest" which could be argued to be of comparable quality to GLIDE's output.

5 Conclusions

miniGLIDE and dogGLIDE are support to the notion that guidance based diffusion models can be replicated and learned on much smaller hardware at a fraction of the parameters. Although our models come with a sacrifice in quality, we hypothesize that given enough time and image-caption pairs, a model like miniGLIDE could learn to create a vast array of photo-realistic topics from text. The power of classifier free guidance on reduced parameter models is eminent here and can especially be seen in dogGLIDE. Future work could include running miniGLIDE on a bigger dataset such as MSCOCO or even LAION-400M for longer periods of time.

References

- Prafulla Dhariwal and Alex Nichol. 2021. [Diffusion models beat gans on image synthesis](#). *CoRR*, abs/2105.05233.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. [Denoising diffusion probabilistic models](#). *CoRR*, abs/2006.11239.
- Jonathan Ho and Tim Salimans. 2021. [Classifier-free diffusion guidance](#). In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.
- Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. [Novel dataset for fine-grained image categorization](#). In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO.
- Alex Nichol and Prafulla Dhariwal. 2021. [Improved denoising diffusion probabilistic models](#). *CoRR*, abs/2102.09672.
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. [GLIDE: towards photorealistic image generation and editing with text-guided diffusion models](#). *CoRR*, abs/2112.10741.
- Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. [Im2text: Describing images using 1 million captioned photographs](#). In *Neural Information Processing Systems (NIPS)*.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. [Zero-shot text-to-image generation](#). *CoRR*, abs/2102.12092.
- Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2017. [Attngan: Fine-grained text to image generation with attentional generative adversarial networks](#). *CoRR*, abs/1711.10485.
- Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. 2021. [LAFITE: towards language-free training for text-to-image generation](#). *CoRR*, abs/2111.13792.