Improving Caption Similarity with GLIDE Generative Models

William Santosa[1] **Manik Narang**[1] **Harshil Bhullar**[1] **Shaira Alam**[1]
[1]UCLA
{wsantosa, maniknarang, hbhullar18, shairaalam}@cs.ucla.edu

Abstract

With the advent of artificial intelligence and machine learning, many have worked toward utilizing these technologies for image generation. Diffusion probabilistic models (often shortened to diffusion models) are one such generative model, used to produce data similar to the ones on which they are trained. OpenAI's GLIDE model, published in 2021, is an example of a diffusion model. When paired with a guidance technique, such as CLIP or CFG, it produces images with greater texture and detail relative to prior models. However, we noticed that because these guidance techniques trade diversity for fidelity, if there does not exist an image with the specified captions in the dataset, it often selects the next best-known characteristics, which in turn results in images that don't exactly match the captions. This report will introduce diffusion models in greater detail, discuss OpenAI's original GLIDE paper, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models", and discuss our efforts to improve caption similarity with GLIDE.

1 Introduction

OpenAI's work in the development of a text-to-image AI model, which combines language understanding with diffusion models, is the foundation of our project. Our objectives include utilizing the trained model created by GLIDE, with a specific emphasis on its language understanding capabilities through diffusion models. We gain an understanding of the pivotal roles played by Contrastive Language-Image Pretraining (CLIP) and Configurable Generation Functions (CFGs) and the model's distinctive capacity to enhance texture and detail in generated images.

1.1 Problem Statement

Within the context of our project, we utilize GLIDE and leverage OpenAI's pre-existing model. Our project unveils an intriguing dynamic and reveals where our problem statement emerges: while GLIDE substantially elevates overall image detail, specific details outlined in captions may not always be maintained. This discrepancy is exemplified in instances where textual prompts describing a scene result in images that deviate from the expected correspondence, revealing a nuanced challenge in the intersection of language understanding and image generation.

1.2 Objective

As we navigate through our exploration, our project focuses on the model's ability to augment texture and detail, which surpasses previous diffusion models. This not only represents an advancement but also positions GLIDE as a standout model in the landscape of AI-generated imagery. Our main objective focuses on fine-tuning the existing OpenAI model's features for increased accuracy in caption similarity in its generation of images.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

This report will provide a detailed account of the problem statement and objectives driving our investigation, shedding light on the strengths and nuances of the Text to Image Diffusion Model. Through this lens, we aim to contribute valuable insights that not only enhance our understanding of GLIDE but also pave the way for future refinements and advancements in the field of AI-generated imagery.

2 Literature Review

In December 2021, OpenAI authors published GLIDE, an advanced AI model excelling in generating high-quality, photorealistic images from textual descriptions. It was created to refine AI-driven image generation, leveraging language model strengths and advanced diffusion techniques. GLIDE improves Text-to-Image AI by combining innovative techniques, improving photorealism and detail, and offering the potential for wide-ranging applications, human-AI collaboration, and groundbreaking contributions to ethical AI development.

2.1 Diffusion Models

The paper included an exploration of generative models and discussed how diffusion models are different from prior models like GANs, VAEs, and more. There are two main processes to facilitate generating novel images: forward and reverse diffusion. In forward diffusion, gaussian noise is iteratively added to an image until the image is purely Gaussian noise. The reverse process, subsequently called reverse diffusion, attempts to undo the Gaussian noise via neural networks. By training a neural network to undo the noise, the model is then able to generate new, similar images from the noise. Generating latent vectors involves utilizing the normal distribution, with parameters like the diffusion rate (β_T) and its complement ($\alpha_T = 1 - \beta_T$) deciding the learning rate.

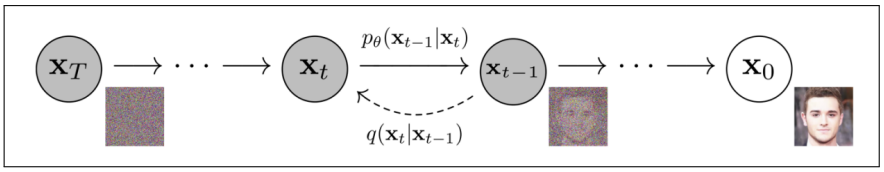

Figure 1: Forward and reverse diffusion process.

2.2 Training Data

The OpenAI team trained their model using a 3.5 billion parameter text-conditional diffusion model (64 x 64) and a 1.5 billion parameter text-conditional upsampling model. Guided Diffusion, employing both Contrastive Language Image Pretraining (CLIP) and Classifier-Free Guidance Scale (CFG) methods, predicts the correct class for input images, guiding the training procedure and creating samples. CLIP utilizes image and caption encoders to guide the training procedure while CFGs use an unconditional variable to randomly select what portion of the images become unconditional.

2.3 Findings

Their findings demonstrate that GLIDE with CFGs generalizes a wider variety of prompts than with CLIP guidance. Quantitatively, GLIDE outperforms other text-conditional generative image models and achieves competitive FID on the MS-COCO validation set. Additionally, human evaluators report that images generated with CFGs have increased caption similarity relative to CLIP.

2.4 Critical Analysis

Limitations such as slow sampling speed due to maintaining the same dimensionality and requiring 15 seconds to sample one image compared to the faster 1 forward pass in GANs, inhibit GLIDE (and in general, diffusion models) from being used in real-time applications. Future improvements should include addressing safety considerations. For example, the model's tendency to produce more pink toys for the prompt "Toys for girls" than for "Toys for boys" highlights a potential bias

that needs correction. and maintaining higher caption similarity when generating images. Points of contention arise when discussing guidance methods, as classifier-free guidance is unexpectedly preferred over CLIP guidance in terms of photorealism and caption-similarity. This finding prompts further exploration into the nuanced dynamics between different guidance methods and their impact on the model's output.

Figure 2: CFG vs CLIP guidance.

2.5 Concluding Notes

As efforts continue to enhance its capabilities, addressing limitations, ethical considerations, and points of contention will be pivotal in ensuring its responsible and impactful integration into various domains. The future trajectory of GLIDE involves not only refining existing functionalities but also exploring novel applications, scaling efficiency, and responding to emerging needs in the dynamic landscape of AI-driven image generation.

3 Methodology

OpenAI glide-text2im API We utilized the reduced model that OpenAI made public for GLIDE. This model is based on the same diffusion model architecture as the full-fledged GLIDE model. However, this model has significantly fewer parameters than the full model (which contains hundreds of millions of parameters). This model is also trained on a subset of the dataset used on the full model, and any potentially offensive and human-based data in the dataset has been removed. Using this API, we wanted to work on providing a text caption as input and generating an output image.

```
# Sampling parameters
prompt = "an oil painting of a corgi"
batch_size = 1
guidance_scale = 3.0

# Tune this parameter to control the sharpness of 256x256 images.
# A value of 1.0 is sharper, but sometimes results in grainy artifacts.
upsample_temp = 0.997
```

Figure 3: GLIDE Sampling Parameters.

The main parameter here is the prompt, which acts as our caption. The model then takes these parameters and feeds them into the model through the p-sample-loop method as seen below. Using this prompt, it generates a 64x64 image output.

3.1 Milestone 2: Approach

In our research, we used an innovative approach by utilizing OpenAI's GLIDE model as the foundation for our project. A hurdle we faced right off the bat was figuring out how to finetune OpenAI's GLIDE model. This proved difficult initially as the full model was not publicly released, so we instead turned our focus on the reduced model that was made available to the public. To enhance the performance of GLIDE, we strategically employed a fine-tuning API, which allowed us to tailor the model more closely to our specific requirements and inputs. While this resource proved very useful, it came with its complications. The repository contained unclear documentation, resulting in time invested to understand the repository structure. Additionally, the installation dependencies were outdated, which we resolved through downgrading Python and upgrading packages. Preprocessing our dataset also proved time-consuming, as we needed to determine what input parameters the model needed. Our methodology involved creating a specialized dataset, curated to augment the model's capability in yielding more effective results. This distinct dataset was used to aid in our main goal of improving the model's caption similarity to our use case. This necessitated the implementation of a customized dataloader, designed to handle both images and captions from this dataset. A critical aspect of setting up this data loader was the adaptation of our dataset's images, ensuring they were reshaped to align with the input parameters required for optimal model training. As for our training process, we fine-tuned our model across different configurations of hyperparameters and training epochs. Determining the number of epochs, learning rate, and the unconditional parameter rate for CFG were also significant steps that we focused on. As for evaluation, we decided to use FID scores. The overarching goal was to strike a balance between fostering the model's generalization capabilities and maintaining high standards of image fidelity and detail.

3.2 Dataset

For our project, we selected Google's Dreambooth dataset as our primary source, attracted by its collection of high-fidelity images. Our focus was particularly narrowed down to dog images within this dataset, aligning with our objective to enhance the model's proficiency in generating captions specifically tailored to dog-related prompts. To effectively utilize these images, we first extracted the dog pictures from the Dreambooth dataset. To mitigate the risk of caption bias, we divided the dataset equally among our team members, ensuring each person was responsible for captioning approximately 25% of the images. This division not only fostered a diverse range of inputs but also helped in maintaining objectivity in our captioning process. This dataset had images of dogs of diverse breeds engaged in a wide array of actions, set against varying backgrounds and contexts. This diversity was beneficial in training our model to recognize and accurately generate captions for a broad spectrum of dog-related scenarios and improve generalizability

3.3 Milestone 3: Limitations of GLIDE

While GLIDE marks a significant advancement over its predecessor, the original DALL-E model, it is not without its limitations, characteristic of generative modeling. One of the primary challenges is the generation of high-resolution images. While striving for high quality and coherence, maintaining resolution remains a hurdle, often demanding increased computational resources. Additionally, GLIDE exhibits biases and ethical concerns, particularly noticeable in its inconsistent capability to accurately generate specific breeds of dogs. A notable example is its tendency to produce images of dogs with black hair/fur when prompted for brown-haired dogs, indicating a potential bias in its training (exemplified in the Results section). This leads to another limitation: the quality and scope of the training data. Like many generative models, GLIDE's performance heavily relies on the diversity and breadth of its training dataset. A limited or skewed dataset can significantly impair the model's output quality and accuracy. Furthermore, the evaluation of generative models like GLIDE remains challenging. Current metrics such as the FID, Inception Score, and R-Precision, while useful, do not fully encapsulate the subjective nature of image assessment. This limitation underscores the ongoing reliance on manual evaluation methods, highlighting the need for more nuanced and comprehensive automatic metrics. Addressing these limitations forms a critical part of our Milestone 3, where we aim to propose and implement enhancements to GLIDE that would mitigate these issues and elevate its performance and applicability.

3.4 Denoising to Improve Quality and Resolution

Our primary approach for milestone 3 was to attempt to denoise images and increase quality through better resolution. We leveraged the upsampler within GLIDE, which is itself a diffusion model. This approach involves a process of gradually transforming a lower-resolution image into a more coherent, higher-resolution image over a series of iterative steps. Distinctively, our model conditions on a low-resolution image, using it as a guiding reference throughout the upsampling process. This contrasts with standard diffusion models, which typically begin with a random noise distribution instead of a concrete image. Our method focuses on iterative refinement, where the model progressively adds finer details and increases pixel density at each step. This method allows the model to incrementally transform the low-resolution base into a high-resolution image by predicting and incorporating high-frequency components absent in the original. It is also important that the upsampler preserves the fundamental content and meaning of the original image while adding new details, AKA semantic consistency. By implementing this technique, the upsampler model effectively reduces the pixelation and noise commonly associated with lower-resolution images. The result is a sharpened, more detailed output, rendering the high-resolution image significantly more natural and coherent, thus addressing one of the key limitations of GLIDE.

3.5 Training Data and Removing Bias

To address another critical limitation of GLIDE, particularly its dependence on training data which can lead to biases, we implemented two key strategies: diversifying the training data and noise augmentation. For diversifying our training dataset, we sourced data from an array of geographic locations, ensuring a broad spectrum of visual styles, environments, and contexts. This included incorporating a variety of content types such as urban and rural landscapes, and both indoor and outdoor settings. Although we were not able to implement these, there are further approaches that could help in improving training data. Firstly, we could implement a data auditing process to actively modify the dataset, addressing specifically identified biases, like the issue with brown and black fur in dogs, and auditing for other potentially skewed representations. Another method would be to introduce synthetic noise augmentation to our methodology. This would involve adding artificial noise to images, creating variations in aspects like lighting, and color, and introducing artificial occlusions such as blurring or pixelation. Further augmenting the dataset could involve techniques like rotations, scaling, and cropping to broaden the diversity of our data. Another method that has seen success in generative modeling improvement is to utilize image pairing, where we can include both a clean and a noise-augmented version of the same image in our dataset. This approach would be designed to aid the model in focusing on key features across different image versions, thereby encouraging it to denoise its outputs. These combined strategies aimed to robustly train the GLIDE model, enhancing its ability to generate more accurate, unbiased, and high-quality outputs.

3.6 Enhancing Text-to-Image Synthesis

When generating captions for images, linguistic discrepancies between captions of the same image can cause deviations from the intended or "ground truth" representation. There is a method that we researched, which provided a refinement process to provide consistency between captions and images: contrastive learning. Contrastive learning plays a crucial role in the pretraining stages. During pretraining, contrastive learning is employed to develop consistent textual representations for various captions describing the same image. This is achieved by creating embeddings of the captions that bring closer the representations of captions for the same image while distancing those for different images in the embedding space. This approach helps in generating images that are more similar to each other when they are created from different captions of the same image, leading to a more accurate and consistent interpretation of the image content. While we were not able to implement this method, we believe that this would further improve the caption similarity of our model.

3.7 Improving Automatic Evaluation Metrics

As mentioned earlier, a limitation of GLIDE, and generative models in general, is the lack of solid automatic evaluation metrics. Since the evaluation of image generation given a caption is subjective, manual methods usually are more reliable, but significantly less efficient. To enhance automatic evaluation metrics for GLIDE, an innovative approach called TIGEr is introduced, moving beyond

traditional text-matching methodologies found in metrics like BLEU and CIDEr. TIGEr addresses the limitations of current metrics that heavily rely on text matching between reference and machine-generated captions. These existing metrics often overlook the fact that reference captions might not fully encapsulate the image content and ignore the inherent ambiguity in natural language. TIGEr, by incorporating a machine-learned text-image grounding model, assesses caption quality by considering not just the textual representation but also how accurately the caption reflects the image content. Furthermore, TIGEr compares machine-generated captions with human-generated ones, offering a more comprehensive evaluation of caption quality, accounting for nuances in image representation and language use. This evaluation metric could prove useful in improving generative models.

4 Results and Discussion

Here, we will briefly go over our results and discuss those results.

4.1 Milestone 1

We chose GLIDE's text-to-image model as the baseline model. We ran the pre-trained model with captions from the GLIDE paper and achieved comparable results.

4.2 Milestones 2 and 3

Training the GLIDE model using Google's Dreambooth dataset resulted in images that resembled the caption better than GLIDE's model did. Using an upsampler trained with 20 epochs resulted in caption-similar but grainy images.

To reduce the grainy effect, we finetuned our model by training the upsampler with 40 epochs. This resulted in better-quality images when compared to images generated using the upsampler trained with 20 epochs in milestone 2.

Left to right: real photograph, GLIDE generated, 20-epoch finetuned-model generated, 40-epoch finetuned-model generated.

Figure 4: "a happy corgi with a big smile surrounded by cherry blossoms"

Figure 5: "small brown fluffy dog with black tongue standing on grass with autumn leaves on the ground"

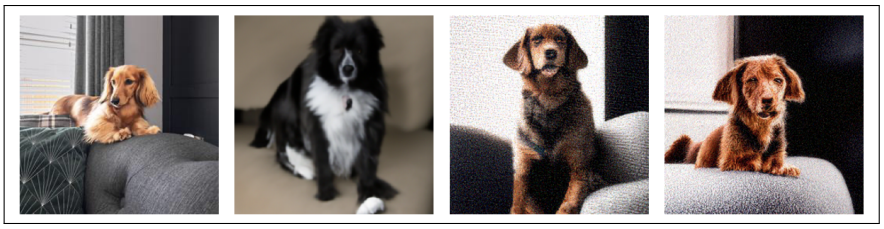

Figure 6: "medium sized brown fluffy dog posing while sitting on a couch"

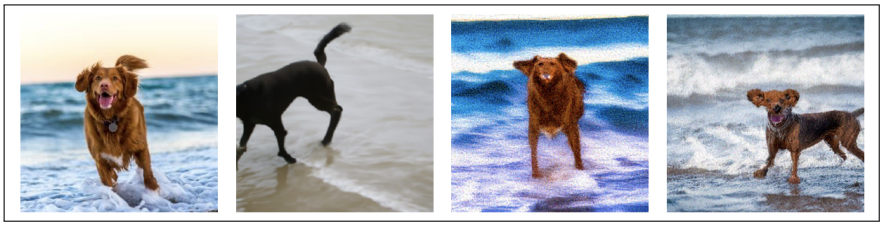

Figure 7: "brown dog with grey dog tags running through waves while smiling at camera"

Judging the images by caption similarity, our finetuned model generated more caption-similar photographs compared to the original GLIDE. Original GLIDE failed to generate photographs with the correct details (as shown in figures 6 and 7 where original GLIDE fails to get the color of the dog correctly).

Furthermore, we quantified the results by using FID scores, which assess visual quality and diversity of images. The lower the FID scores, the better. When we compared the FID scores of real images with those of original GLIDE and finetuned GLIDE, we found out that finetuned GLIDE images were more similar to the real images, as reflected by the lower scores.

Upsampler epochs	Models	64x64	256x256
20 epochs	Avg FID real images vs original GLIDE	281.75	237.73
20 epochs	Avg FID real images vs finetuned GLIDE	217.36	191.12
40 epochs	Avg FID real images vs original GLIDE	281.75	237.73
40 epochs	Avg FID real images vs finetuned GLIDE	217.36	189.55

Table 1: Image quality comparison using FID scores

5 Conclusion

As we conclude our project on the Text to Image Diffusion Model, we can consider the key insights gained and the impact of specific training strategies on the capabilities of GLIDE. Our exploration, encapsulated within the problem statement and objectives, not only displays the strengths of this model but also lays the groundwork for potential future advancements in the dynamic field of AI-generated imagery.

5.1 Summary

Summarizing our findings, the precision of training specificity emerges as a major factor in fine-tuning the model, enhancing caption similarity, and allowing the model to capture nuances and details specific to given contexts.

Our main model focused on providing images that increase caption similarity for images related to “dogs”. This provided a critical insight on the importance of the data contribution towards the accuracy of the image-caption similarity.

In terms of fine-tuning, we found that there was a direct influence of the number of training epochs on model performance underscores the importance of balancing computational resources for optimal efficacy. Integrating upsampling techniques has proven instrumental in refining image resolution, contributing to a more visually appealing output.

Additionally, our evaluation metric using FID affirms the success of our fine-tuning efforts, with lower FID values correlating with improved image generation. This holistic evaluation, combining qualitative and quantitative analysis, substantiates the model's efficacy in generating accurate and coherent textual-image pairs.

5.2 Contributions

Reflecting on our contributions, this report not only sheds light on the current capabilities of the Text to Image Glide Diffusion Model but also underscores the potential for future refinements and advancements. The concept of training specificity opens avenues for tailoring models to specific domains, while the balance between training epochs and model performance invites ongoing research for optimal strategies in real-world applications with vast amounts of data for better precision in output.

The integration of human influence in the form of image captioning and qualitative analysis, as discussed in the associated research paper, *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*, emphasizes the relationship between AI and human insight in refining the quality of generated images. Acknowledging this collaborative approach is crucial for the continued development of text-to-image generation.

5.3 Potential Future Work

Looking forward, potential future work includes exploring further fine-tuning models for specific domains, optimizing training strategies, and further advancing upsampling techniques to refine image resolution. The extension of the model's capabilities to generate textual descriptions from images represents an exciting avenue for broader utility.

In conclusion, our exploration not only encapsulates the nuances of the Text to Image Diffusion Model and the prowess of GLIDE but also lays the foundation for continued advancements. As we navigate the evolving landscape of AI-generated imagery, this report displays the potential of this model, guiding future endeavors and improvements that are bound to come

6 References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising diffusion probabilistic models*. Advances in Neural Information Processing Systems, 33:6840–6851, 2020.

[2] Luo, C. (2022). *Understanding Diffusion Models: A Unified Perspective*. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/2208.11970.

[3] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... Chen, M. (2022). *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2112.10741.

[4] Sohl-Dickstein, Jascha; Weiss, Eric; Maheswaranathan, Niru; Ganguli, Surya (2015-06-01). *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. Proceedings of the 32nd International Conference on Machine Learning. PMLR. 37: 2256–2265.

7 Appendices

7.1 GLIDE

Glide, a pioneering text-to-image model developed by OpenAI, stands at the forefront of AI-driven image generation. This appendix provides additional details about Glide, its architecture, and its underlying mechanisms. For a comprehensive understanding of Glide, refer to the official documentation provided by OpenAI.

7.2 AI (Artificial Intelligence)

Artificial Intelligence, a foundational concept in our exploration, plays a crucial role in the development and application of advanced models like Glide. This appendix offers a brief overview of key principles, methodologies, and applications of Artificial Intelligence, setting the stage for the context in which Glide operates.

7.3 CLIP (Contrastive Language-Image Pretraining)

CLIP is an integral component influencing Glide's capabilities. This appendix delves into the specifics of CLIP, its architecture, and its role in enhancing language understanding and image generation. For further details, refer to the dedicated documentation provided by OpenAI on CLIP.

7.4 CFGs (Configurable Generation Functions)

Configurable Generation Functions (CFGs) contribute significantly to the capabilities of Glide. This appendix provides supplementary information on CFGs, their role in the model, and their impact on the generation of visually detailed images. Further insights can be obtained from OpenAI's documentation on CFGs.

7.5 FID (Fréchet Inception Distance)

FID serves as a quantitative metric for evaluating the quality of generated images. This appendix offers a concise explanation of FID, its computation, and its relevance in assessing the success of our fine-tuning efforts. For a more in-depth understanding, refer to the literature on FID in image generation research.

7.6 Generative Model

A generative model is a type of statistical model that generates new samples that resemble a given dataset. In this context, generative models learn the underlying patterns and structures of the provided dataset to create new, similar data points. They can be utilized in a variation of real-world applications, image generation, text generation, and data synthesis.

Generative models aim to capture the probability distribution of the data, allowing them to produce realistic samples. Common examples of generative models include Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and autoregressive models.

7.7 Diffusion Model

A diffusion model is a probabilistic model used to describe the process by which a substance or quantity spreads or diffuses over time. In the context of machine learning and image processing, diffusion models are employed for tasks such as image denoising and generation.

The diffusion process in these models involves iteratively transforming an initial signal or image through a series of diffusion steps, slowly adding noise to the signal. By controlling the diffusion process, these models can be used to remove noise from images, generate realistic samples, or even perform tasks like image editing. The denoising diffusion probabilistic model is a specific type of diffusion model designed for denoising images.