An Analysis of the Currently Available Text-to-Image Models

Thales Rogério Sales Almeida, Rodrigo Frasseto Nogueira, Hélio Pedrini

Relatório Técnico - IC-PFG-22-36
Projeto Final de Graduação
2022 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

The contents of this report are the sole responsibility of the authors.
O conteúdo deste relatório é de única responsabilidade dos autores.

An Analysis of the Currently Available Text-to-Image Models

Thales Rogério Sales Almeida* Rodrigo Frasseto Nogueira† Hélio Pedrini‡

Abstract

This work aims to study some of the most prominent publicly available models for the text-to-image generation task. In addition, we investigated whether an ensemble of these models can achieve better results using a CLIP model as a ranker. To perform these experiments, we selected two available models that performed well on the public MS-COCO benchmark. We also experimented with Stable Diffusion, a diffusion model that became popular due to the quality of the images it generates. We evaluated each model and the ensembles in subsets of the MS-COCO and FLICKR datasets.

1 Introduction

The desire to create an image simply by describing it is not new. A significant amount of research has been done by exploring different models and architectures to perform text-to-image generation. In the last months, we have been bombarded with breakthroughs in the field. Google presented Imagen [1], a diffusion model that was able to generate incredibly realistic results and achieved state-of-the-art results in the MS-COCO [2] evaluation dataset. A few weeks later, Google presented Parti [3], an autoregressive model with nearly 20 billion parameters that surpassed the Imagen results in MS-COCO and showed that autoregressive models could achieve astonishing results as well.

While the results of the Google models were impressive, the Google team did not provide any way for the community to interact with the model, since the code, dataset and model weights were not turned public, most likely due to fear of the malicious applications that such models could have. In this scenario, a couple of weeks after Parti appeared, a new diffusion model called “Stable Diffusion” was published by stabilityAI [1]. It showed a remarkable capability to generate convincing images; however, no benchmarks were published on the performance of the model in traditional datasets. Stable Diffusion weights were later made public and numerous applications derived from it.

Regarding text-to-image generation, we see in the field a huge gap between the best performing models and the models that are publicly available, since most developers of the SOTA models end up choosing not to make the models public for diverse reasons, such as fear of misuse or commercial implications. This situation motivates the research presented

*Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP, 13083-852.

†NeuralMind, Campinas, SP, 13083-898

‡Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP, 13083-852.

[1] https://stability.ai/

here, where we aim to study and discuss how well the currently available models perform in traditional Fréchet Inception Distance (FID) [4, 5] evaluations.

2 Related Work

Recently, LAFITE [6] has introduced a new way to train text-to-image models. Instead of relying in manually annotated text-to-image pairs, it invests in the high correlation between text and images predicted by CLIP [7] and uses a pretrained clip model to generate text features from the image features. It achieved impressive in-domain results, surpassing even Dall-E [8] in the MS-COCO dataset under the FID metric. The authors made the weights publicly available.

GLIDE [9] is a diffusion model that achieved notable results in various datasets. The main version of GLIDE was composed of two parts: a text-conditioned diffusion model that generated images at a resolution of 64×64, and a text-conditioned upsampling diffusion model that increases the generation to the 256×256 resolution. The associated paper showed the capability of GLIDE to generate realistic images. Although GLIDE did not make the weights public, it provided a smaller version of the model trained on a filtered dataset that did not contain humans or hatred symbols.

Stable Diffusion [10] is a latent text-to-image diffusion model that takes advantage of a CLIP ViT-L/14 encoder in order to condition the model generation. The model was trained in a subset of the LAION 5B [11] dataset. The training was done with images with a resolution of 256×256 pixels. The model was then finetuned in images with 512×512 pixels.

3 Methodology

This section describes the main stages of the methodology proposed in this work.

3.1 Selected Models

In total, three different models were used in this study: LAFITE, GLIDE and Stable Diffusion, which are briefly described as follows.

LAFITE was selected due to the high FID reported on the MS-COCO benchmark when finetuned on the dataset, rivaling much larger models such as Parti and Imagen. We choose to use the version of LAFITE that was finetuned on the MS-COCO dataset.

GLIDE was selected because it was the next best performing model in the MS-COCO benchmark with available checkpoints. However, the available weights were not the same ones that achieved the best performing results in MS-COCO. In this study, we use the filtered version of GLIDE that was made public. This version is smaller and was trained on a filtered subset of the original GLIDE. This subset excluded figures of humans and sensitive symbols and, even with these limitations, it was shown that this model could still generate reasonable results [9]. However, the model is unable to generate certain objects and humans.

Finally, we selected Stable Diffusion as one of our candidates due to its recent popularity. We used Stable Diffusion to generate 256×256 images; however, we observed later in the study that the quality of the generations from Stable Diffusion degraded a lot when generating images in a different resolution than the one used for finetuning the model, which was 512×512 pixels. It is likely that we would have found much more promising results by using Stable Diffusion to generate 512×512 images and then downscale to 256×256 pixels.

3.2 Ensembles

This work also aims to study how well the selected models can cooperate to achieve better performance. In order to do that we will experiment with two singular configurations:

- Ensemble 1 - A cooperation between all 3 models.
- Ensemble 2 - A cooperation between only GLIDE and Stable Diffusion.

The ensemble process is simple: we generate a certain amount of candidates images for a given prompt with each model and then use a CLIP model to score each candidate. The candidate with the best score is selected and used as the "output" of the ensemble. For the remainder of this report, we will refer to the process just described as CLIP reranking. Note that this process demands inference from all the models involved in the ensemble, which can be quite computationally expensive.

In this work, mainly due to computational constraints, we generated eight candidates for each model and prompt. We will later also discuss how much each one of the selected models performance improves when using CLIP reranking in eight samples.

3.3 Selected Datasets

In order to evaluate the selected models and ensembles, we choose to use the MS-COCO and FLICKR datasets. MS-COCO was selected since it is a traditional benchmark for text-to-image generation; however, some of the selected models (namely LAFITE) were finetuned on the dataset, meaning that we needed another dataset in order to make more accurate comparisons in a zero-shot scenario. For this purpose, we selected FLICKR.

Due to computational constraints, this work was unable to use the full validation set for MS-COCO (41000 images and text pairs) and the full FLICKR dataset (30000 image and text pairs). Instead, we selected a subset of one thousand images from MS-COCO validation set and from FLICKR to perform our evaluations. For the same reason, we could not calculate the FID metric in the standard 50 thousand images [12]. Therefore, we calculate each FID in a set of a thousand images and compute some extra metrics, such as LPIPS [13] and SSIM [14], to improve our analysis.

Furthermore, while the MS-COCO dataset has only one caption for each image in the dataset, the FLICKR dataset provides on average 5 different captions for each image. In this work, we choose to use for each image the caption with the highest length, assuming that it would be a more detailed description of the image.

4 Results

This section presents and discusses the experimental results achieved with this work.

4.1 Comparing models

We start our study by exploring a couple of samples from each model. Figure 1 shows different prompts and the images generated by each model.

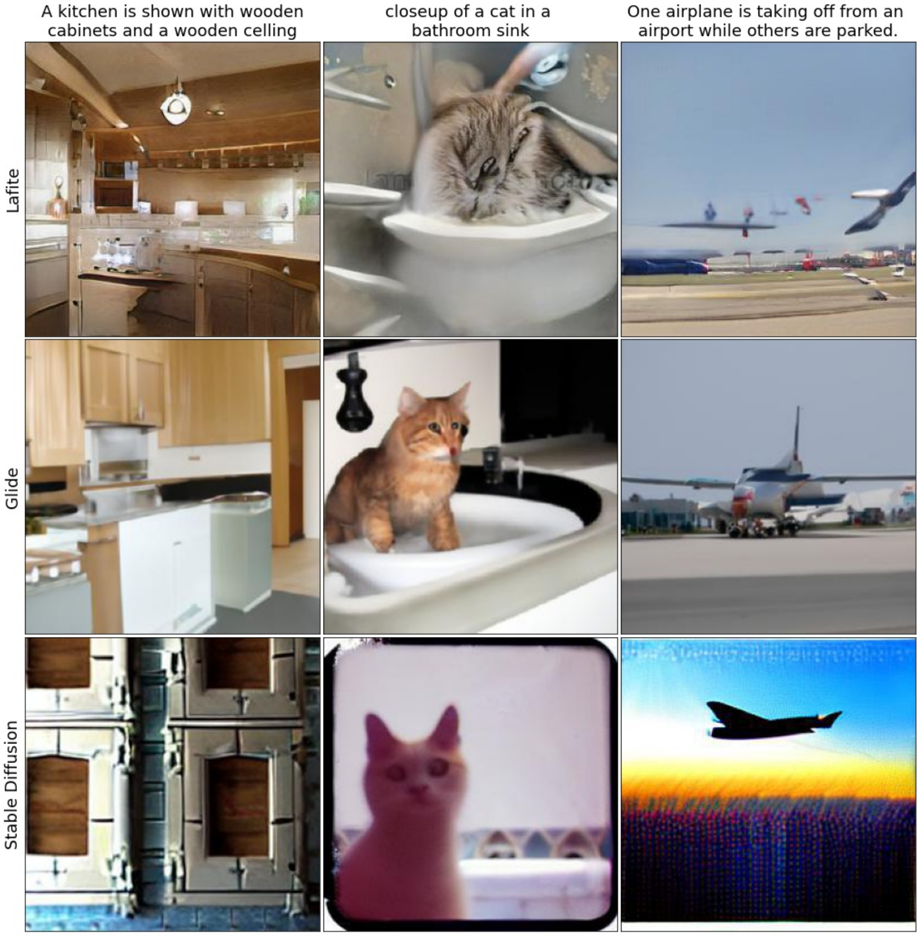

A kitchen is shown with wooden cabinets and a wooden celling
closeup of a cat in a bathroom sink
One airplane is taking off from an airport while others are parked.
Lafite
Glide
Stable Diffusion

Figure 1: Random samples from each model using prompts from MS-COCO.

From these samples, we can see that GLIDE and LAFITE generate images closer to reality, while Stable Diffusion's 256×256 images have more visual artifacts.

Another interesting aspect is how diverse the generations of each model are. In other words, how diverse are the samples of each model for the same prompt. Figures 2 and 3 show three different generations for the same prompt. It is possible to observe that GLIDE and Stable Diffusion are able to generate different images each time, whereas LAFITE generates almost identical images, only with small alterations such as flipping some objects in the scene.

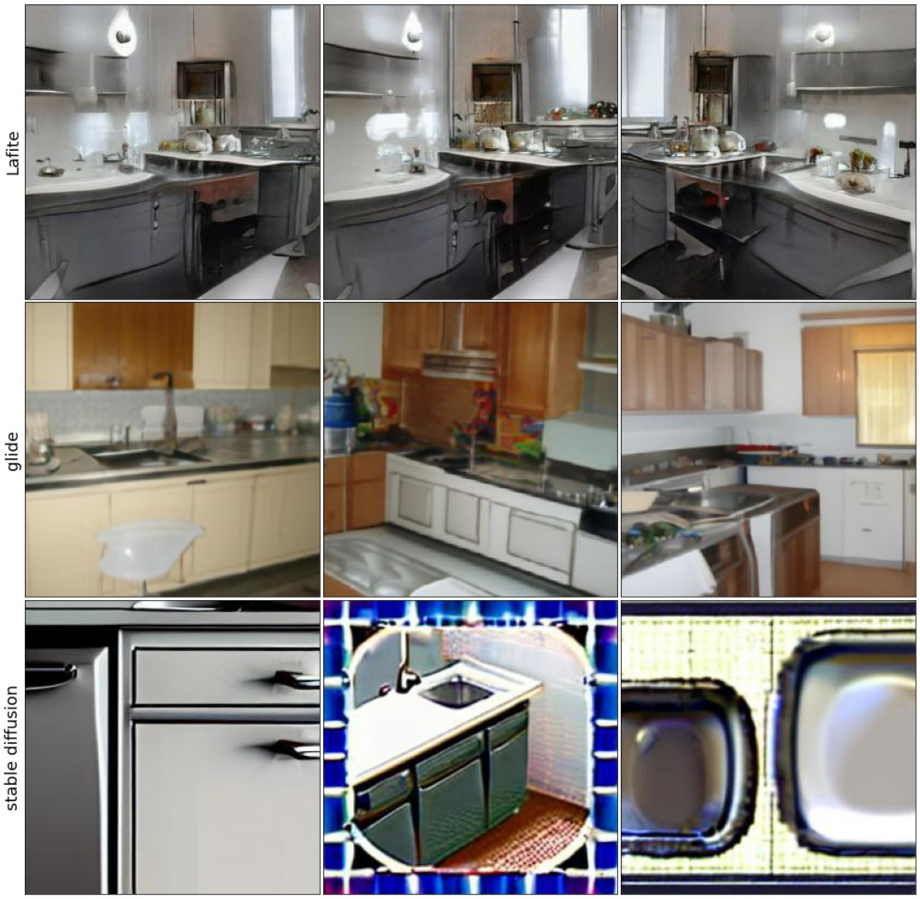


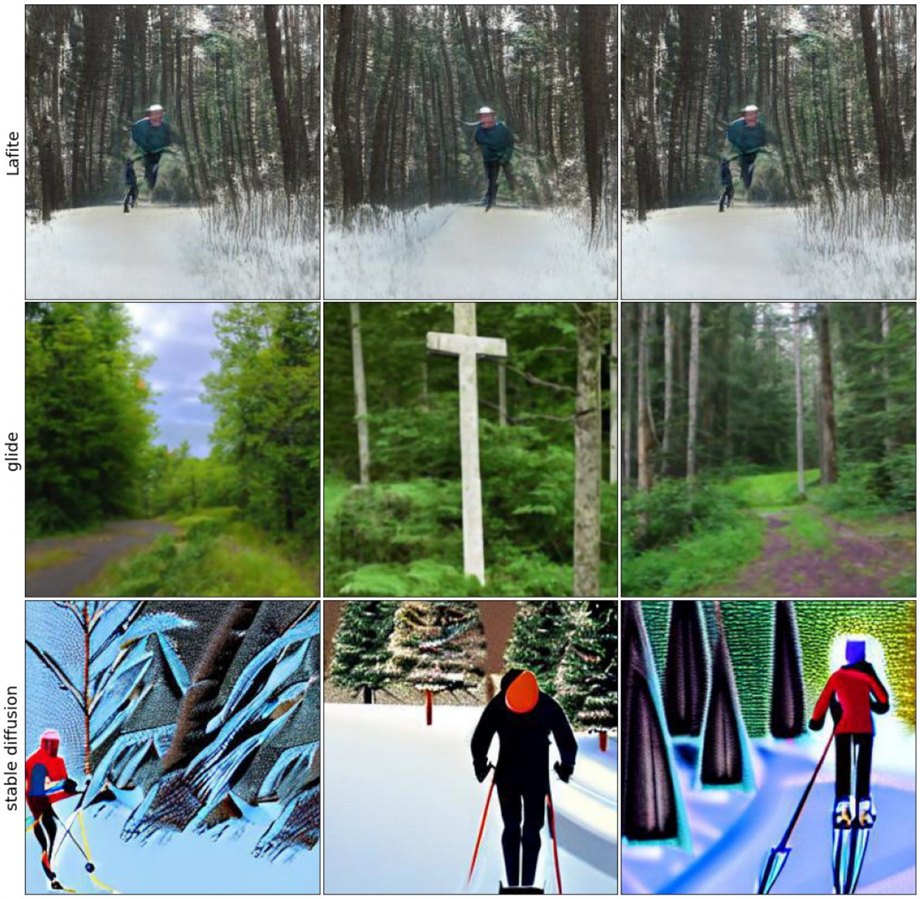
Figure 2: Generations for the prompt "this is a kitchen with dishes and a silver sink".

We also observe that the prompt shown in Figure 3 involves the notion of an "old man", Stable Diffusion and LAFITE are able to handle the concept and generate somewhat cohesive images. However, this version of GLIDE (since it was not trained in any images containing humans) is unable to understand the notion and completely ignores it in its generations, focusing only on other aspects of the prompt such as "cross" and "through the forest".

4.2 Quantitative Comparison

We evaluated each model in our respective datasets in order to know how well each of them performs on their own. Table 1 shows the FID obtained in each dataset.

We can see that LAFITE achieves the best FID value on both datasets. GLIDE is the second best in MS-COCO, while Stable Diffusion is the second best in FLICKR.



]

Figure 3: Generations for the prompt “the old man cross country skis through the forest”.

Table 1: Results for each model generating only 1 sample.

Model	**MS-COCO (FID)**	**FLICKR(FID)**
LAFITE	**7.209**	**4.422**
GLIDE	9.509	7.984
Stable Diffusion	12.912	7.957

4.3 CLIP Reranking

As mentioned previously, we are also interested in exploring how much we can improve the results of a given model using CLIP Reranking. We show the resulting images when generating only one sample for a couple of prompts in Figure 4 and the resulting images for the same prompts after CLIP Reranking with 8 samples in Figure 5.

We show in Table 2 the FID results for each model when generating only one sample and when generating 8 samples by applying CLIP reranking. We can see that LAFITE shows a very similar FID both when generating only 1 sample and when generating 8, which is

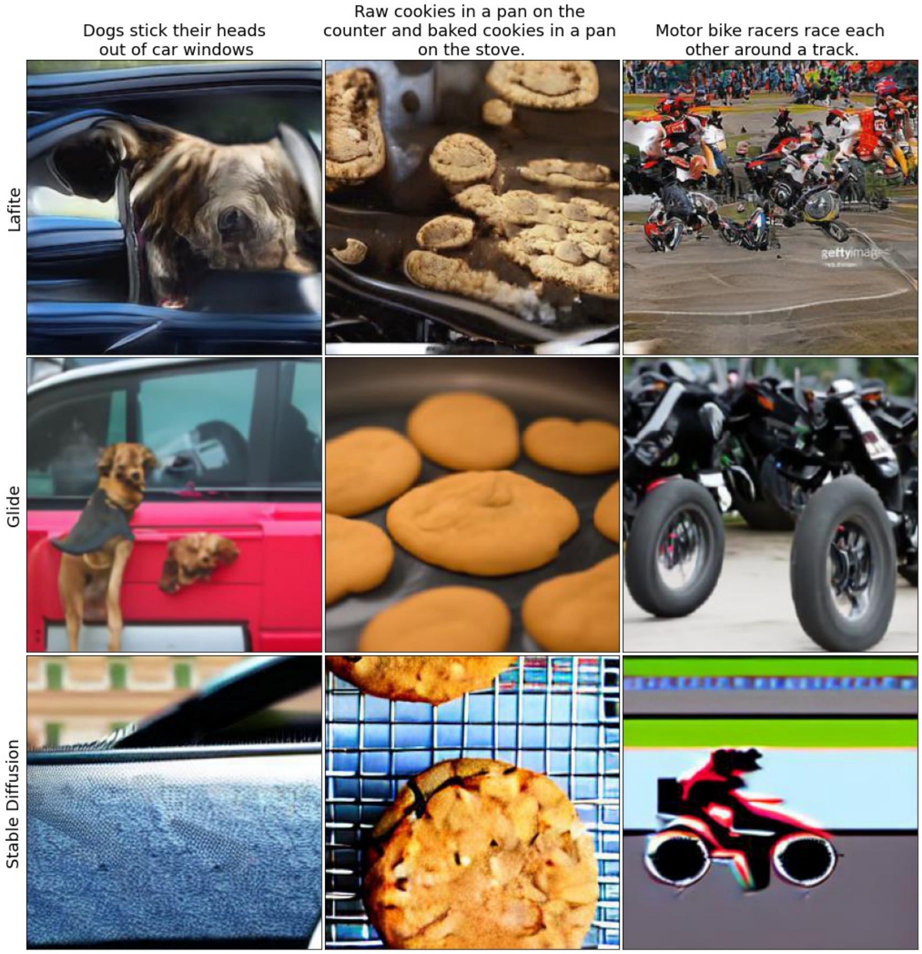

Dogs stick their heads out of car windows
Raw cookies in a pan on the counter and baked cookies in a pan on the stove.
Motor bike racers race each other around a track.
Lafite
Glide
Stable Diffusion

Figure 4: Resulting images when generating 1 sample.

expected since, as observed previously, LAFITE cannot generate a very diverse set of images for the same prompt. Both GLIDE and Stable Diffusion showed an improvement of around 0.7 FID point when generating 8 samples in MS-COCO; however, the same did not happen in the FLICKR dataset, where both GLIDE and Stable Diffusion were slightly worse.

Table 2: Comparison between the FID achieved when generating only one sample (FID(1)) and when applying CLIP reranking in 8 samples (FID(8)).

	MS-COCO			**FLICKR**		
Model	**FID(1)**	**FID(8)**	**Difference**	**FID(1)**	**FID(8)**	**Difference**
LAFITE	7.209	7.267	-0.057	4.422	4.378	0.044
GLIDE	9,509	8.708	0.800	7.984	8.112	-0.128
Stable Diffusion	12.912	12.271	0.640	7.957	7.993	-0.042

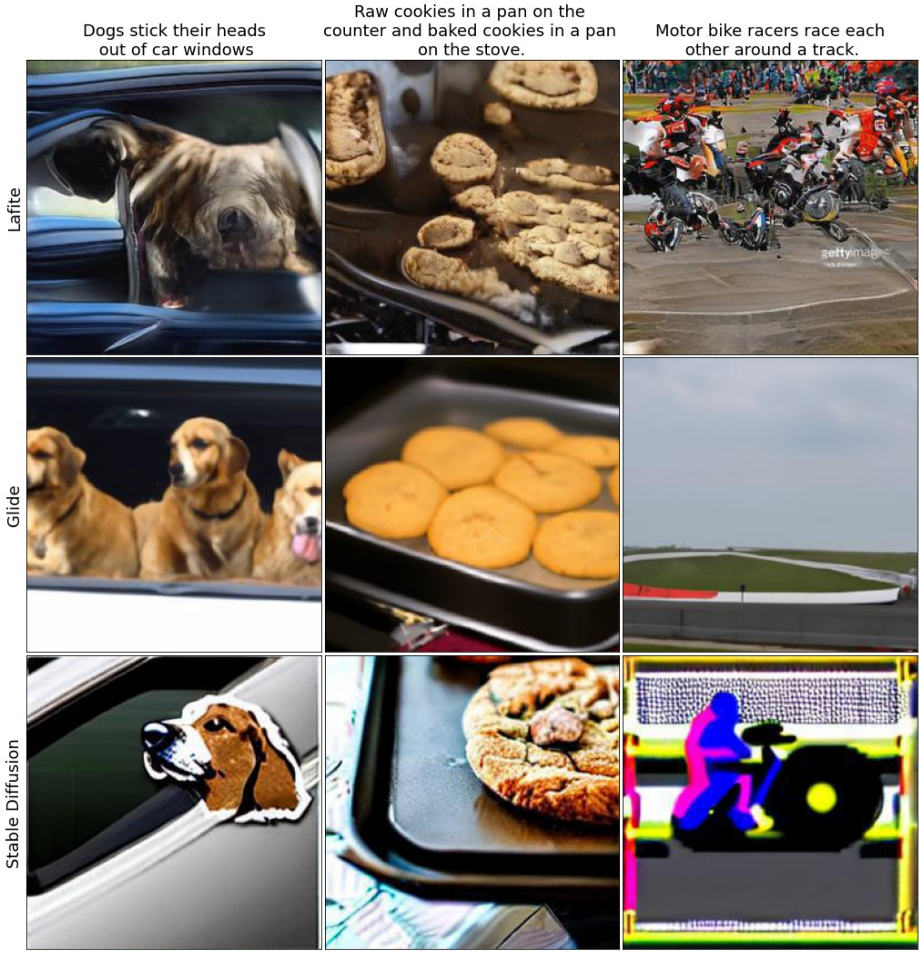

Dogs stick their heads out of car windows
Raw cookies in a pan on the counter and baked cookies in a pan on the stove.
Motor bike racers race each other around a track.
Lafite
Glide
Stable Diffusion
gettyimages

Figure 5: Resulting images when generating 8 samples and selecting with CLIP.

4.4 Ensembles

We showed previously that each of the models being studied has significant problems: GLIDE cannot generate humans, Stable Diffusion in the studied configuration had trouble in generating realistic images, whereas LAFITE cannot generate diverse samples. Given these problems, we investigated the capacity of collaboration among the samples generated by each model.

The ensemble process used is straightforward, where each model generates 8 candidate samples, making a total of 24 candidates. Each candidate is scored by CLIP and the image with the highest score is selected as the "output" of the ensemble. In other words, we just use CLIP reranking between the samples generated from all the models involved in the ensemble.

We study two different ensembles:

- Ensemble 1 - A cooperation between all 3 models.

- Ensemble 2 - A cooperation between only GLIDE and Stable Diffusion.

The motivation for studying the collaboration between only GLIDE and Stable Diffusion is that they should be able to cover each other's flaws. Stable Diffusion can generate symbols and humanoid figures, but GLIDE cannot, whereas GLIDE can provide more realistic generations of scenes and ambient. In addition, we also intended to see whether the combination of both models could surpass LAFITE. Table 3 shows the metrics of both ensembles and for each model.

Table 3: Metrics for each of the selected models and ensembles.

	MS-COCO				**FLICKR**			
Model	**FID**	**LPIPS**	**PNSR**	**SSIM**	**FID**	**LPIPS**	**PNSR**	**SSIM**
LAFITE	**7.267**	**0.6485**	8.4223	**0,1486**	4.387	**0.6681**	8.7369	0.1668
GLIDE	8.708	0.6671	8.7779	0.2361	8.112	0.7045	8.7600	0.2221
Stable Diffusion	12.271	0.7167	**7.2750**	0.1674	8.000	0.7359	**7.5508**	**0.1429**
Ensemble 1	7.307	0.6559	8.8111	0.1497	**4.316**	0.6773	8.6088	0.1662
Ensemble 2	9.162	0.6804	8.2234	0.1783	6.848	0.7224	7.9396	0.1635

Regarding the MS-COCO dataset, we can see from the results that neither ensemble 1 nor ensemble 2 was able to achieve better performance than the individual models. Ensemble 1 reached metrics similar to LAFITE, while ensemble 2 achieved metrics that are a couple of points lower than GLIDE. Apparently, since the models had a significant disparity in the quality of the results, the collaboration was not successful.

However, when we observe the results for FLICKR dataset, while ensemble 1 again shows similar metrics to LAFITE, ensemble 2 shows an interesting fact where this ensemble outperformed both original models under FID metric by more than 1 point. It is also worth mentioning that, on this dataset, both Stable Diffusion and GLIDE have similar capabilities. These results indicate that, in order to see improvements by using CLIP reranking-based ensembles, the original models should have a similar performance. When one model is much better than the others, the ensemble seems to be only as good as the best in the ensemble model.

5 Conclusions

In this work, we studied publicly available text-to-image models that showed promising results in the MS-COCO benchmark in the last few years, evaluating then in subsets of the MS-COCO and FLICKR datasets. We evaluated each model in each subset, studied the effects of CLIP reranking in 8 samples from each model, and studied the gains in performance of CLIP reranking-based ensembles in two different configurations. We observe that CLIP reranking provided gains to models that are able to generate diverse samples in the majority of cases, but there were exceptions.

Finally, we observed that our CLIP reranking ensembles were always limited to the performance of the best individual model when the models composing the ensembles have a high disparity between them. However, when the individual models had similar capabilities, the ensemble tended to be successful.

As directions for future work, we could explore more models or other variations, such as Stable Diffusion for generating 512×512 images. It would also be interesting to extend the evaluations to the entire dataset and see if the results found in this research still hold. Finally, we could evaluate other reranking methods, such as other versions of CLIP or even other methods that measure text and image correlation.

References

[1] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL `https://arxiv.org/abs/2205.11487`.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. URL `https://arxiv.org/abs/1405.0312`.

[3] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. URL `https://arxiv.org/abs/2206.10789`.

[4] 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. 2017. URL `https://arxiv.org/abs/1706.08500`.

[5] Alexander Mathiasen and Frederik Hvilshøj. Backpropagating through fréchet inception distance, 2020. URL `https://arxiv.org/abs/2009.14075`.

[6] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation, 2021. URL `https://arxiv.org/abs/2111.13792`.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

[8] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL `https://arxiv.org/abs/2102.12092`.

[9] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021. URL `https://arxiv.org/abs/2112.10741`.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. URL `https://arxiv.org/abs/2112.10752`.

[11] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models, 2022. URL `https://arxiv.org/abs/2210.08402`.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2018. URL `https://arxiv.org/abs/1812.04948`.

[13] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, 2018. URL `https://arxiv.org/abs/1801.03924`.

[14] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.