COMPARATIVE ANALYSIS OF DIFFUSION MODELS FOR IMAGE GENERATION

Vaibhav Sharan
vsharan1@asu.edu

Krishnaprasad Palamattam Aji
kpalamat@asu.edu

Unnikrishnan Madhavan
umadhava@asu.edu

Ansh Sharma
ansh@asu.edu

1 Introduction

The project aims at implementing and comparing different diffusion models that are used for image generation. We will evaluate the performance of these different models in generating such high-quality images from textual prompts. The well known models like FLUX, Stable Diffusion will be studied in detail and compared with each other on the basis of their strengths, weaknesses and applicability in different scenarios. This study aims to shed light on the effectiveness of text to image diffusion models.

2 Background and Motivation

Artificial Intelligence has witnessed a paradigm shift in techniques for image generation over the past few years. Generative Adversarial Networks(GANs) have been the go to approach for synthetic image creation for a long time. Diffusion models recently emerged as a powerful alternative, especially in the domain of text-to-image conversion [1].

Diffusion models were introduced in 2015 by Sohl-Dickstein et al.[2] and have gained popularity due to their ability to generate diverse images with high quality and remarkable fidelity. GANs rely on a generator-discriminator (adversarial) architecture to generate images. Unlike GANs, diffusion models use a gradual denoising process to generate images which has shown remarkable stability during training and better control over the generation process.

Diffusion models gained popularity due to their exceptional performance in generating images from text. Models like DALL-E 2, Stable Diffusion, FLUX.1 and Midjourney have captured public imagination with their ability to form real like images from textual descriptions. Our project "Comparative Analysis of Diffusion Models for Image Generation", is motivated by the need to understand the strengths and weaknesses of different diffusion model architectures. These models are evolving rapidly, and a comprehensive comparison is crucial and beneficial to both researchers and practitioners in this field.

To conduct this comparison and analysis, we will be using a range of quantitative metrics that have become a standard in evaluating such image generation models. The metrics we use will include Inception Score(IS), Fréchet Inception Distance(FID), Contrastive Language-Image Pre-training(CLIP), Text-to-Image Faithfulness evaluation with question Answering(TIFA), and CLIP Maximum Mean Discrepancy(CMMD).

By utilizing these metrics, we aim to provide a nuanced understanding of how different diffusion models fare across different aspects of image generation. This will help researchers and practitioners in the field to select the right diffusion model for their work.

3 Related Work

Diffusion models for image generation were first introduced by Sohl-Dickstein et al. and later refined by Ho et al., demonstrating the potential of this approach for high-quality image synthesis. Building on this foundation, several text-to-image diffusion models have achieved impressive results, pushing the boundaries of what's possible in AI-generated imagery. Recent advancements in diffusion models have focused on improving various aspects such as efficiency, scale, architecture, training techniques, and controllability. This comparative study aims to evaluate several state-of-the-art diffusion-based text-to-image models across various metrics to assess their strengths, limitations, and potential synergies in the domain of text-to-image generation. This section provides an overview of key developments in diffusion-based text-to-image synthesis, with a focus on three prominent models: Stable Diffusion, FLUX.1, and Kwai Kolors.

Stable Diffusion, developed by Rombach et al., introduced the concept of latent diffusion, allowing for efficient training and inference while maintaining high image quality. This model has become widely adopted due to its open-source nature and ability to generate diverse, high-resolution images from text prompts. Stable diffusion focuses on improving and scaling rectified flow models for high-resolution text-to-image synthesis. A rectified flow formulation, which connects data and noise in a straight line, rather than traditional curved diffusion paths is used. A set of new noise samplers (Logit-Normal sampling, Mode sampling with heavy tails, CosMap sampling) that improve performance over previous samplers were introduced for rectified flow models. A novel transformer-based architecture called MM-DiT (Multimodal Diffusion Transformer) specifically designed for text-to-image tasks was used in stable diffusion. It uses separate weights for text and image modalities. The authors scale their model up to 8B parameters and demonstrate predictable scaling trends in validation loss and performance metrics. Their largest model (depth=38, 1024x1024 resolution) achieves a 0.74 overall score on GenEval, outperforming DALL-E 3 (0.67) and other state-of-the-art models. On human preference evaluations using Parti-prompts, their 8B model outperforms current closed and open-source SOTA models in visual quality, prompt following, and typography generation. Larger models can be sampled using fewer steps. For example, their depth=38 model shows only a 2.71% relative CLIP score decrease when using 5 sampling steps instead of 50, compared to 4.30% for the depth=15 model. They demonstrate a strong correlation between validation loss and comprehensive evaluation metrics like GenEval, T2I-CompBench, and human preference ratings.

FLUX.1, a more recent development by Blacck Forest labs, aims to improve upon Stable Diffusion's capabilities by incorporating advanced techniques such as rectified flow and parallel attention layers. FLUX claims to offer enhanced speed, efficiency, and prompt adherence compared to earlier diffusion models. All public FLUX.1 models are based on a hybrid architecture of multimodal and parallel diffusion transformer blocks and scaled to 12B parameters. The model claims to improve over previous state-of-the-art diffusion models by building on flow matching, a general and conceptually simple method for training generative models, which includes diffusion as a special case. In addition, the model claims to have increased model performance and improved hardware efficiency by incorporating rotary positional embeddings and parallel attention layers.

Source: How does Flux work? The new image generation AI that rivals Midjourney. medium link: https://medium.com/@drmarcosv/how-does-flux-work-the-new-image-generation-ai-that-rivals-midjourney-7f81f6f354da https://blackforestlabs.ai/announcing-black-forest-labs/

Kwai Kolors is a large-scale text-to-image generation model based on latent diffusion, developed by the Kuaishou Kolors team. Similar to other latent diffusion models, Kolors operates in a compressed latent space rather than pixel space, which offers advantages in computational efficiency and generation quality. The model was trained on billions of text-image pairs and demonstrates significant capabilities in visual quality, complex semantic accuracy, and text rendering for both Chinese and English characters. The Kolors model is based on a latent diffusion architecture, which operates in a compressed latent space rather than pixel space. This approach offers several advantages such as reduced computational complexity, improved generation quality, and enhanced robustness to noise. the model architecture consists of three main components: a variational encoder (VAE) for encoding and ddecoding images, a U-Net back bone for the diffusion process, and a noise-aware classifier-free guidance mechanism which allows for better control over the generation process and improved quality of generated images. Kolors employs an adaptive noise schedule that adjusts the noise level based on the input image quality. By operating in a compressed latent space, Kolors achieves faster training and inference times compared to pixel-space diffusion modelsThis feature enhances the model's robustness to various levels of noiseKolors achieves an FID score of 2.27 on the FFHQ dataset, outperforming other state-of-the-art model.

The evaluation of diffusion models for image generation is an active area of research with different metrics to assess the various aspects of images generated. The key metrics we will be using in our comparison are:

3.1 Inception Score (IS)

Introduced by Salimans et al. [3], the Inception Score measures the quality and diversity of the images generated. It was initially designed for GANs but it can be used for any image generation models. It uses a pre-trained Inception v3 network to evaluate how well the generated images can be classified into distinct categories. The IS is calculated as:

$$IS = \exp(\mathbb{E}_x[KL(p(y|x)||p(y))]) \tag{1}$$

where $p(y|x)$ is the conditional label distribution for generated images, and $p(y)$ is the marginal label distribution over all generated images. A higher IS score tells us that the image is more diverse and realistic.

3.2 Fréchet Inception Distance (FID)

Proposed by Heusel et al. [4], FID is a combination of Fréchet distance and features extracted from the Inception-v3 model. FID was introduced after IS and addresses some of its limitations. FID calculates the distance between the feature representations of the generated and real image distributions:

$$FID = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \tag{2}$$

where μ_r, Σ_r are the mean and covariance of the real image features, and μ_g, Σ_g are those of the generated images. Lower FID scores suggest that the generated images are closer to real images in terms of quality and diversity.

3.3 CLIP Score

Leveraging the CLIP (Contrastive Language-Image Pre-training) model developed by Radford et al. [5], the CLIP Score assesses how well images are generated and align with their text prompts:

$$CLIP\_Score = \cos(CLIP_{text}(prompt), CLIP_{image}(generated\_image)) \tag{3}$$

where $\cos$ is the cosine similarity between the CLIP embeddings of the text prompt and the generated image. A higher CLIP score means that the generated image matches better with the text prompt.

3.4 CLIP-Maximum Mean Discrepancy

CLIP-MMD(CMMD) is an extension of the CLIP score and uses Maximum Mean Discrepancy (MMD) to measure the distance between the CLIP embeddings of generated and real images [6]:

$$CLIP\text{-}MMD = MMD(CLIP_{image}(real\_images), CLIP_{image}(generated\_images)) \tag{4}$$

A lower CMMD score means that the generated images are closer to the real images in terms of feature space distribution. CMMD claims to fix some of the limitations of FID [7].

3.5 TIFA (Text-to-Image Faithfulness evaluation with question Answering)

Introduced by Yang et al. [8], TIFA measures how faitfully a generated image represents its text prompt. It uses visual question answering to evaluate how well the generated image captures the information from their textual descriptions:

$$TIFA\_Score = Accuracy(VQA_{model}(generated\_image, question\_from\_prompt)) \tag{5}$$

where the VQA model answers questions derived from the original text prompt based on the generated image. TIFA is a fairly new metric and compared to other traditional metrics, offers a more fine-grained assessment of text-image alignment.

When these metrics are used in combination to evaluate our diffusion models, we can assess the quality, diversity and faithfulness to text prompts of the images generated.

4 Progress

5 Execution Plan

6 Workload Distribution

6.1 Model Setup and Implementation (Vaibhav Sharan, Ansh Sharma)

- Set up local environments for Stable Diffusion, FLUX.1, and Kwai-Kolors
- Implement image generation scripts for each model

6.2 Quantitative Metrics Implementation (Krishnaprasad Palamattam Aji)

- Develop scripts for IS, FID, CLIP Score and CMMD

6.3 Data Collection and Processing (Unnikrishnan Madhavan)

- Create diverse text prompts dataset
- Develop scripts for TIFA metrics

6.4 Analysis and Report Writing (All Team Members)

- Analyze results and compare model performances
- Write findings and prepare visualizations

Weekly meetings will be held to track progress and address challenges.

7 Execution Plan

8 Headings: first level

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. See Section 8.

8.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j | y, v, w; \theta) = \frac{\alpha_i(t) a^{w_t}_{ij} \beta_j(t+1) b^{v_{t+1}}_j(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a^{w_t}_{ij} \beta_j(t+1) b^{v_{t+1}}_j(y_{t+1})} \tag{6}$$

8.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

9 Examples of citations, figures, tables, references

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget

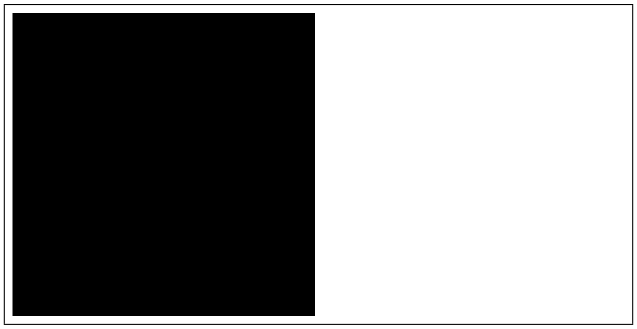

Figure 1: Sample figure caption.

lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui. [1, **?**] and see [**?**].

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
   \citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated…

`https://www.ctan.org/pkg/booktabs`

9.1 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure 1. Here is how you add footnotes. [1] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

9.2 Tables

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo. See awesome Table 1.

9.3 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

[1] Sample of the first footnote.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

10 Conclusion

Your conclusion here

Acknowledgments

This was was supported in part by......

References

[1] Dhariwal P and Nichol A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems, 34, 8780-8794*, 2021.

[2] Maheswaranathan N. Sohl-Dickstein J., Weiss E. A. and S Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning, 2256-2265*, 2015.

[3] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[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. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021.

[6] Ruocheng Gao, Xiaoyuan Guo, Kristen Grauman, and Matt Kusner. Measuring and mitigating unintended bias in image captioning. *arXiv preprint arXiv:2211.13449*, 2022.

[7] Andreas Veit Daniel Glasner Ayan Chakrabarti Sanjiv Kumar Sadeep Jayasumana, Srikumar Ramalingam. Rethinking fid: Towards a better evaluation metric for image generation. *arXiv preprint arXiv:2401.09603v2*, 2024. License: CC BY-NC-SA 4.0.

[8] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3335–3343, 2022.