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
ashar479@asu.edu

ABSTRACT

Abstract goes here.

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 became popular as they performed remarkably well in generating images from textual prompts. Models like DALL-E 2, Stable Diffusion, FLUX.1 and Midjourney are some of the most popular text to image diffusion models that generate high quality images. Our project "Comparative Analysis of Diffusion Models for Image Generation", is motivated by the need to understand both the strengths and weaknesses of different diffusion models. Research in the field of diffusion models is growing rapidly and a comprehensive comparison between models is beneficial to both researchers and practitioners in this field.

We will use a wide range of quantitative metrics to compare and analyze image generation diffusion 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 shed light on 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

Recent advancements in diffusion models have focused on improving various aspects such as architecture, scale, efficiency, training techniques, and controllability. This study aims to compare and evaluate state-of-the-art diffusion-based text-to-image models across various metrics to assess their strengths, limitations, and potential similarities in the

Model	Stable Diffusion	FLUX.1	Kolors
Architecture	Latent diffusion model	Rectified flow Transformer	Latent diffusion model
Parameter Size	860 million (SD 1.5)	12 billion	2.6 billion
Image Quality	High	Very highs	High
Prompt Understanding	Keyword-based	Advanced NLP	Advanced NLP
Unique Features	Inpainting, outpainting, image-to-image translation	Excellent prompt adherence, detailed outputs	High-quality photorealistic synthesis, noise robust

Table 1: Comparison of Stable Diffusion, FLUX.1, and Kolors

domain of text-to-image generation. An overview of key developments in diffusion-based text-to-image generation is discussed in this section, with a focus on three recent models; Stable Diffusion, FLUX.1, and Kwai-Kolors.

Stable Diffusion, developed by Rombach et al. [3], introduced the concept of latent diffusion, that allows efficient training and inference while maintaining high image quality. Rather than the traditional curved diffusion paths, a rectified flow formulation, which connects data and noise in straight line 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. This architecture uses separate weights for text and image modalities.

FLUX.1, a recent development by Black Forest labs, tries to improve upon Stable Diffusion's capabilities by incorporating techniques such as rectified flow and parallel attention layers [4]. FLUX claims to offer enhanced speed, efficiency, and prompt adherence compared to earlier diffusion models. FLUX.1 model is based on an architecture that consists of parallel and multimodal diffusion transformers with 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 [5]. In addition, the model claims to have increased model performance and improved hardware efficiency by incorporating rotary positional embeddings and parallel attention layers.

Kwai-Kolors is a text-to-image generation model based on latent diffusion, developed by the Kuaishou Kolors team. Kolors operates in a compressed latent space rather than pixel space, which offers advantages in computational efficiency and generation quality [6]. 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 latent diffusion 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 decoding images, a U-Net back bone for the diffusion process, and a noise-aware classifier-free guidance mechanism. Kolors uses an adaptive noise schedule that adjusts the noise level based on the input image quality.

Table 1 captures some of the characteristics of these models.

4 Evaluation Metrics

We will be using the following metrics to assess the various aspects of images generated by diffusion models from textual prompts:

4.1 Inception Score (IS)

Inception Score was introduced by Salimans et al. [7]. It measures the quality and diversity of images generated by a model. 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.

4.2 Fréchet Inception Distance (FID)

Proposed by Heusel et al. [8], 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.

4.3 CLIP Score

Leveraging the CLIP (Contrastive Language-Image Pre-training) model developed by Radford et al. [9], 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.

4.4 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 [10]:

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

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 [11].

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

Introduced by Yang et al. [12], TIFA measures how faithfully 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.

5 Methodology

5.1 Model Setup

Vaibhav Setup and Models go here

5.2 Evaluation Metrics Methodology

We downloaded 100 image caption pairs from `Conceptual Captions`, a Google Research Dataset. The captions provided along with the images were not descriptive enough to generate images using the diffusion models. An advanced vision model Microsoft `florence-2` integration was used on Hugging Face to generate more descriptive captions for the downloaded images. These prompts were used for text to image generation in the three selected models. This section describes the methodology used to evaluate the selected diffusion models.

5.2.1 Inception Score

The Inception Score (IS) evaluates the quality and diversity of generated images. A pre-trained Inception V3 model was used with `InceptionScore` from torchmetrics. The images were resized to 299 x 299 for compatability and then converted to tensors. The Inception Score and Standard Deviation was then calculated for the set of generated images of each diffusion model. The mean IS and SD were used for our comparison and analysis.

5.2.2 CLIP score

The CLIP score measures how well the generated images align with their textual prompts. The generated images were preprocessed into tensors and all the captions of the generated images were stored in a text file. These were the inputs to `CLIPScore` from torchmetrics, a pre-trained CLIP model(openai/clip-vit-base-patch16) which was used to compute the similarity scores between them. The CLIP score of a model was taken as the mean of CLIP scores of the 100 generated images of the model.

5.2.3 Fréchet Inception Distance

FID calculates the similarity between the distributions of real and generated images in feature space. It captures both the quality and realism of generated images. The real and their corresponding generated images were given as inputs to the `FrechetInceptionDistance` function from torchmetrics.

5.2.4 TIFA

TIFA measures how faithfully the generated images represent their textual prompts by leveraging a Visual Question Answering(VQA) model. The TIFA methodology used here is similar to the one mentioned in [13]. Question answer pairs generated for each textual prompt using GPT-3. Since these questions were generic and not specific to the textual description, the set of questions and answers were created manually for the 100 images. 5 question answer pairs which asked about the subject, color in the image, presence of people, weather, etc were created for each textual prompt. A BLIP VQA model (Salesforce/blip-vqa-base) was used to the answer the questions given the generated image. A pre-trained BERT model from Hugging Face was used to compute the similarity between expected answers and the answers given by the VQA model. If the similarity score was above 0.7, it was deemed to be the right answer. The total number of correct answers and the total number of questions were recorded and used to calculate the TIFA score.

6 Results and Analysis

Result and Analysis goes here

7 Team Member Contribution

Vaibhav Sharan

-

Krishnaprasad Palamattam Aji

- Prepared descriptive textual inputs for image generation in all 3 models
- Script generation for calculating IS, FID, CLIP Score and TIFA score of all the models

Unnikrishnan Madhavan

-

Ansh Sharma

-

8 Conclusion

Conclusion and any future work goes here

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] Esser Patrick, Kulal Sumith, Blattmann Andreas, Entezari Rahim, Muller Jonas, Saini Harry, Levi Yam, Lorenz Dominik, Sauer Axel, Boesel Frederic, Podell Dustin, Dockhorn Tim, English Zion, Lacey Kyle, Goodwin Alex, Marek Yannik, and Rombach Robin. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206v1*, 2024.

[4] Marcos V. Conde. Announcing black forest labs. `https://medium.com/@drmarcosv/how-does-flux-work-the-new-image-generation-ai-that-rivals-midjourney-7f81f6f354da`, 2024.

[5] BlackForestLabs. Announcing black forest labs. `https://blackforestlabs.ai/announcing-black-forest-labs/`, 2024.

[6] Kuaishou Technology. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. 2024.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] Yushi Hu. tifa. `https://github.com/Yushi-Hu/tifa`, 2023.