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

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.

Model	Stable Diffusion	FLUX.1	Kwai-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 high, especially with dev/pro versions	High
Prompt Understanding	Keyword-based	Advanced NLP	Advanced NLP
Unique Features	Inpainting, outpainting, image-to-image	Excellent prompt adherence, detailed outputs	High-quality photorealistic synthesis, noise robust

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

Stable Diffusion, developed by Rombach et al. [3], introduced the concept of latent diffusion, allowing for efficient training and inference while maintaining high image quality. 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. 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 Black Forest labs, aims to improve upon Stable Diffusion's capabilities by incorporating advanced 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. 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 [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 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 [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 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 models. This feature enhances the model's robustness to various levels of noise.

Table 1 captures some of the characteristics of these models.

4 Evaluation Metrics

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:

4.1 Inception Score (IS)

Introduced by Salimans et al. [7], 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.

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 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.

5 Progress

For the Milestone 1 Report, we are about one week behind schedule as mentioned the timeline we estimated in the Project proposal because of more extensive literature review than expected. We have included the updated Execution Plan accordingly such that we are able to finish the project in stipulated time. The following subsections describe the tasks we have completed for the first milestone.

5.1 Survey of State of the Art Diffusion Models

We performed a comprehensive search and survey of the current open-source models available on the internet, more specifically on websites like HuggingFace.co, Github, civit.ai to identify the models we are going to perform this comparative study on. We selected FLUX.1[dev] by Black Forest Lab, Stable Diffusion 3 Medium by StabilityAI, and Kolors by Kuaishou Kolors team.

5.2 Identification of Evaluation Metrics

As mentioned in section 4, we identified Inception Score, Fréchet Inception Distance, CLIP Score, CLIP-MMD, and TIFA to be the most relevant and effective our comparative study.

5.3 Environment Setup for the Diffusion Models

We have created and activated virtual environments for each of the three models with all the required packages and libraries including PyTorch, Hugging Face Diffusers, CUDA. The model weights are ready to use for the evaluation step.

6 Execution Plan

This following timeline is the updated version taking into consideration the work done till first milestone.

Prompt Engineering (Week 6). We will be creating sample text prompts for image generation in different categories such as people, multiple objects, landscapes, portraits.

Collection of Evaluation Metrics Data (Weeks 7 and 8). In this phase, we will use identical prompts to generate images from all models and analyze them to evaluate the generated image output. We will implement and collect metrics for quantitative evaluation as discussed before.

Comparative Analysis (Week 9). During this part of project, we will use different metrics and compare models on them along with assessing their performance by testing various batch sizes and image resolutions.

Results Compilation and Conclusions (Week 10). Now we will start writing reports by collecting and analyzing the results data and perform statistical analysis to highlight significant differences between models. We will also use graphs and charts to further elaborate our results.

Report Writing and Presentation (Weeks 11 and 12). Here, we will write a detailed report on highlighting results, and conclusions and simultaneously develop a presentation, we will also include a gallery of generated images for comparison. In the last week, an internal peer review will be conducted to finalize the report and presentation based on everyone's feedback.

7 Workload Distribution

7.1 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

7.2 Quantitative Metrics Implementation (Krishnaprasad Palamattam Aji)

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

7.3 Data Collection and Processing (Unnikrishnan Madhavan)

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

7.4 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.

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.