COMPARATIVE ANALYSIS OF DIFFUSION MODELS FOR IMAGE GENERATION

GROUP NUMBER 34

Vaibhav Sharan vsharan1@asu.edu
Krishnaprasad Palamattam Aji kpalamat@asu.edu
Unnikrishnan Madhavan umadhava@asu.edu
Ansh Sharma ashar479@asu.edu

ABSTRACT

This study presents a comparative analysis of diffusion models in the context of text-to-image generation. This reports start with brief introduction of diffusion models, followed by a section on the literature and related work we reviewed. We then discuss about standard metrics used in quantifying diffusion model quality and performance. This section is followed by a detailed description of the methodology followed in this work. Results and analysis derived are discussed. We then summarize this report with challenges we faced and scope for future work.

Keywords: Diffusion Models, Text-to-Image Generation, Stable Diffusion, FLUX.1, Kwai-Kolors, Image Quality, Inception Score (IS), Fréchet Inception Distance (FID), CLIP Score, Text-to-Image Faithfulness (TIFA), Artificial Intelligence, Image Synthesis, Generative Models.

1 Introduction

In recent years, diffusion models have emerged as a powerful approach for text-to-image generation, offering significant advancements in generating high-quality and semantically aligned images. Humans naturally visualize images when reading stories, enhancing comprehension and enjoyment. Developing an automated system to generate realistic images from textual descriptions is a complex challenge and represents a significant step toward achieving human-like artificial intelligence. With advancements in deep learning, the text-to-image task has emerged as a remarkable application in Generative AI. This study evaluates the performance of several state-of-the-art diffusion models, including Stable Diffusion 3 Medium, Stable Diffusion 3.5 Medium, and FLUX.1-dev, across key standard metrics such as IS, CLIP, FID, TIFA, and CLIP MMD. The process consists of creating text prompts from a dataset, utilizing the models to create images, and evaluating the results using these metrics. The results help to clarify the capabilities of each model and help researchers choose the best model for a given application by highlighting its advantages, disadvantages, and determine their effectiveness and suitability for various applications.

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.

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

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 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 Implementation Challenges

We performed a comprehensive search and survey of the current open-source models available on the internet, more specifically on websites like HuggingFace.co [13], GitHub, civit.ai [14] to identify the models we are going to perform this comparative study on. Till the time we were working on our milestone one report we had selected FLUX.1-dev by Black Forest Lab, Stable Diffusion 3 Medium by StabilityAI, and Kolors by Kuaishou Kolors team. While implementing the image generation and processing pipeline described in the next section, we faced various challenges, the major one being frequent crashes because of low memory on the inference machine.

The biggest challenge we faced was finding a suitable machine for our task. Diffusion models consists of billions of parameters, spread across multiple layers which even after the model has been trained, is needed to be loaded into the memory for inference of images from the text-prompts. For example, the model size for FLUX.1-dev is 30GB. Other than the diffusion model, we need text-encoders and text-decoder models to process the prompts, variational autoencoders for helping in the image generation process. These are also loaded in the memory, and can take upto several GBs of memory based on the exact models used. We explored multiple cloud options, including ASU Sol supercomputer, but they had limitations on what we could run on them. Running a image generation pipeline of diffusion models is a heavy task and was not allowed. In the end, we decided to use a personal laptop with decent hardware and a dedicated GPU for our project.

The other major challenge we faced was the availability of open source weights for Kolors diffusion model. On doing further research we found out that different versions of Stable Diffusions performed differently. Therefore we chose Stable Diffusion 3.5 as our new bench mark instead of Kolors.

To overcome this challenge, we searched about techniques which allows for diffusion models to fit in our machine with 6GB VRAM and 16GB RAM. The method we used is called Quantization [15]. Quantization is a method used to decrease the computational and memory expenses associated with running inference by using low-precision data types, such as 8-bit integers (int8), instead of the standard 32-bit floating-point (float32) representation for weights and activations.

By lowering the bit count, the model requires less memory storage and theoretically consumes less energy. Additionally, operations like matrix multiplication can be executed more quickly using integer arithmetic. This technique also enables models to run on devices with low memory like our consumer grade laptop.

Quantization is a relatively new technique in the realm of diffusion models, and there are currently few publicly available quantized versions. In our research, we tested 16-bit, 8-bit, 5-bit, and 4-bit quantized models of FLUX.1-dev [16], Stable Diffusion 3 Medium [17], and Stable Diffusion 3.5 Medium [18]. We deviated from our initial plan of using Kolors model, and instead went with comparison of two versions of Stable Diffusion family of models, SD3 Medium and SD3.5 Medium.

We found that the 16-bit models for SD3 and SD3.5 could fit within our machine's memory limits, which can be calculated by multiplying the number of parameters by 16-bits, 4GB for SD3 and 5GB for SD3. However, FLUX.1-dev, which has 12 billion parameters, was too large to fit even in its 5-bit variant. We were able to run the 4-bit quantized model of FLUX.1-dev, but the images produced had significant quality issues, including distorted colors, noisy artifacts, and overall low quality. These results underscore the challenges associated with using aggressive quantization in diffusion models.

5.2 Image Generation Pipeline

To streamline the whole image generation process, we used ComfyUI [19] framework. The first step was acquiring the model weights for the diffusion models from HuggingFace. These weights were of GGUF format, to integrate them with our ComfyUI workflow, we used a custom node called ComfyUI-GGUF [20]. As quantized models do not inherently include built-in CLIP [21] models, it is necessary to download the appropriate CLIP models from the Stable Diffusion Text Encoders and store them in the CLIP directory of the ComfyUI folder.

The pipeline employs a CLIP text encoder to convert text prompts into embeddings that inform the image generation process. This task is facilitated by the `CLIPTextEncode` node within ComfyUI. We also used `t5xxl_fp16.safetensors` [22] from Google for text encoding of the prompts.

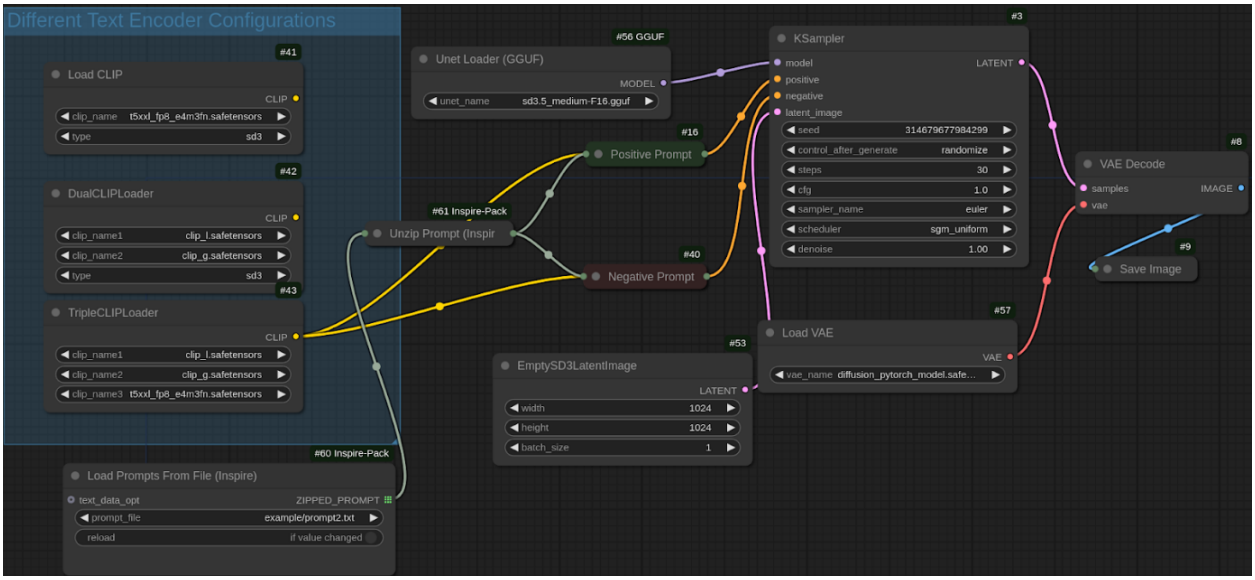

Figure 1: Representation of the ComfyUI workflow for SD3.5 image generation pipeline. The connections between the different nodes are marked by colored lines. The inference parameter details are shown in the KSampler node.

The Variational Autoencoder (VAE) [23] is integral to encoding images into latent space and subsequently decoding them back into pixel space. In ComfyUI, researchers can utilize the `VAE Encode` node to transform images into latent representations and the `VAE Decode` node to reconstruct images from these latent codes.

The workflow within ComfyUI as shown in figure 1 was structured to connect nodes responsible for loading the U-Net model, CLIP model, and VAE model effectively. The process consisted of utilizing an `EmptySD3LatentImage` node to initialize a latent image for as our starting noisy image. We employed the `KSampler` node to manage parameters like denoising, iterations, seeds, sampler for the diffusion process. We kept the iterations fixed to 30 steps and used the same parameters for all the models to ensure the comparative study is accurate. Finally, we connected a `Save Image` node at the end of the pipeline to store the generated images in a PNG format.

Another custom node `Inspire Pack` [24] was utilized in this project to make one-click generation of images from a text file containing multiple prompts possible. This custom node reads all the prompts in a zipped fashion from the text file and then unzips them and feeds them into the `Positive Prompt` and `Negative Prompt` nodes.

This pipeline was used to generate a set of 100 images for each model from the generated text prompts as discussed in the next section. The average inference time of each image for FLUX.1-dev, SD3, SD3.5 was 12 minutes, 4 minutes, and 6 minutes respectively.

5.3 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.

5.3.1 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.

5.3.2 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.

5.3.3 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.

5.3.4 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 [25]. 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.

5.4 CLIP MMD

CLIP MMD is an integration of MMD techniques into CLIP framework and helps in quantifying text-to-image generative model performance. FID assumes normality of distributions which is often violated. Whereas MMD does not rely on such assumptions. Also, MMD is less dependent on sample size compared to FID. These characteristics makes MMD more reliable and consistent across different evaluation settings. Studies also show that CLIP MMD aligns better with human judgments than FID, particularly in cases involving progressive distortions or subtle improvements in image quality. To measure the CLIP MMD score for the chosen models, same image pairs used for TIFA evaluation was used as CLIP MMD measures the distance between real and generated images. A python script with Embeddings from the `clip_mmd` libraries were used to calculate the score for each model.

6 Results and Analysis

Table 2 captures the results from metrics evaluation.

Model	IS	CLIP	FID	TIFA	CLIP MMD
Stable Diffusion 3 Medium	6.413	18.847	180.193	0.838	0.548
Stable Diffusion 3.5 Medium	6.127	18.744	183.266	0.824	0.497
FLUX.1-dev	6.247	18.805	181.133	0.917	0.438

Table 2: Metrics scores for models compared

From this data it can be inferred that for the used setup, Stable Diffusion 3 Medium shows high IS, CLIP and FID scores. Whereas FLUX.1-dev has better numbers in terms of TIFA and CLIP MMD scores among the models compared. FLUX.1-dev demonstrates excellent semantic alignment and fidelity to text prompts. The lowest CLIP MMD shows strong multi modal alignment FLUX.1-dev model holds. Stable Diffusion 3 Medium has slightly better FID than FLUX.1-dev, suggesting marginally more realistic image quality in terms of distribution similarity to real data. Stable Diffusion 3.5 Medium performs slightly worse across most metrics compared to its predecessor (Stable Diffusion 3 Medium). From this data we can conclude that Stable Diffusion 3 Medium and FLUX.1-dev are suitable for applications such as creative content generation, diverse dataset generation for training machine learning models. Also, the high TIFA score of FLUX.1-dev makes it ideal candidate for scientific or educational illustrations where text fidelity is crucial.

These scores can vary depending on the flow setup and other implementation intricacies. The table just demonstrates an overall picture from our comparative study.

Images generated for the text prompt: "The image is a collage of three photos showing how to make 10 minute DIY leather pulls. The first photo on the top left shows a close-up of a black cabinet door with two leather handles attached to it. The handles are brown and appear to be made of leather. The second photo in the top right shows the same leather handles hanging on the door. The third photo shows a black wardrobe with a coat rack and a potted plant on the floor. The text on the image reads "10 Minute DIY Leather Pulls"." are captured in figures 2, 3, and 4. The images illustrates the difference in images generated using the models compared to an good extent. We can see that the image generated using FLUX.1-dev (Fig 4) is realistic and it was able to embed the text accurately in the image.

Figure 2: SD3 Medium

Figure 3: SD3.5 Medium

Figure 4: FLUX.1-dev

7 Future Scope

This study opens up several avenues for future research:

- **Examining New Models:** Future studies can look at multimodal models and newer versions of Stable Diffusion (like SD 4.0) for more dynamic image production.
- **Domain-certain Fine-Tuning:** By optimizing models for specialized domains such as fashion design or medical imaging, results for certain industries can be improved.
- **Advanced Evaluation Criteria:** It might be feasible to better capture aspects of created images including texture, semantic accuracy, and originality by developing more intricate criteria.
- **Model Optimization:** These models are made more usable for real-time applications by reducing computation costs through techniques like knowledge distillation and pruning.
- **Ethical and Bias Considerations:** Future studies should address model biases and fairness to guarantee ethical use in text-to-image generation.

8 Challenges Faced

We ran into a number of issues during the research that hindered our ability to compare diffusion models in a consistent and seamless manner:

- **Problems with Model Configuration:** The study first employed the FLUX.1-dev, Kolors, and SD3. Later, we switched to SD3.5 Medium instead of Kolors, which created more overhead in terms of model configuration and implementation and to guarantee consistent input-output compatibility.
- **Performance Variability:** Direct comparisons were challenging due to the models' performance variability in areas like image quality, text-to-image alignment, and variety.
- **Data and Prompt Generation:** In order to improve the descriptiveness of the Conceptual Captions dataset, sophisticated models such as Microsoft Florence-2 were employed, which increased processing complexity and time.

- **Computational Constraints:** The increased complexity of newer models led to longer inference times, and even with quantized versions, the inference time remained a bottleneck for generating large sets of images.
- **Limitations of Open-Source Models:** The study was limited to open-source models, which, while accessible, lacked certain proprietary optimizations and features that could have improved performance and customization.

9 Conclusion

In this study, we conducted a comprehensive comparison of diffusion models for text-to-image generation, focusing on their performance across multiple metrics, including IS, CLIP, FID, TIFA, and CLIP MMD. We implemented a image generation pipeline of the selected models. These pipelines were fed by text prompts generated by the methods discussed to create the set of images for our metrics calculations. The results revealed distinct strengths and weaknesses among the models, providing valuable insights into their suitability for various applications. All the three models generate images with decent diversity and quality. Notably, FLUX.1-dev demonstrated superior text-image alignment, fidelity, and text embedding capabilities, as evidenced by its TIFA and CLIP MMD scores.

10 Team Member Contribution

Vaibhav Sharan

- Performed the model selection process
- Implemented the image generation pipeline
- Generated images from the text prompts for all the diffusion models

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

- Setup the script and environment for CLIP MMD Calculation
- Analyzed the results and performed a thorough study to understand the different use cases of each models and their trade offs

Ansh Sharma

- Collaborated with team members to interpret the results of each metric and identify trends or patterns in model performance.
- Assisted in preparing slides or visual representations of the data and findings for project presentations.

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] Hugging face – the ai community building the future. `https://huggingface.co/`.

[14] Civitai: The home of open-source generative ai. `https://civitai.com/`.

[15] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models, 2023.

[16] city96/flux.1-dev-gguf · hugging face. `https://huggingface.co/city96/FLUX.1-dev-gguf`.

[17] city96/stable-diffusion-3-medium-gguf · hugging face. `https://huggingface.co/city96/stable-diffusion-3-medium-gguf`.

[18] city96/stable-diffusion-3.5-medium-gguf · hugging face. `https://huggingface.co/city96/stable-diffusion-3.5-medium-gguf`.

[19] Comfy ui. `https://github.com/comfyanonymous/ComfyUI`.

[20] Comfy ui gguf. `https://github.com/city96/ComfyUI-GGUF`.

[21] Openai clip. `https://github.com/openai/CLIP`.

[22] google/t5-efficient-xxl · hugging face. `https://huggingface.co/google/t5-efficient-xxl`.

[23] Wikipedia contributors. Variational autoencoder — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Variational_autoencoder&oldid=1261565566`, 2024. [Online; accessed 12-December-2024].

[24] Comfyui inspire pack. `https://github.com/ltdrdata/ComfyUI-Inspire-Pack`.

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