Issues in Information Systems
Volume 25, Issue 2, pp. 277-292, 2024

DOI: https://doi.org/10.48009/2_iis_2024_123

Perception and evaluation of text-to-image generative AI models: a comparative study of DALL-E, Google Imagen, GROK, and Stable diffusion

Suhaima Jamal, *Georgia Southern University, sj14077@georgiasouthern.edu*
Hayden Wimmer, *Georgia Southern University, hwimmer@georgiasouthern.edu*
Carl M. Rebman, Jr., *University of San Diego, carlr@sandiego.edu*

Abstract

Generative Artificial Intelligence (AI) model is a revolutionary type of AI capable of producing high quality images based on textual inputs. These models utilize natural language processing (NLP) techniques and computer vision to understand and interpret the textual descriptions and then generate images that align with the given descriptions. This study evaluates four prominent text-to-image generative models-DALL-E, Google Imagen, Stable Diffusion, and GROK AI emphasizing on the text-to-image diffusion models. Using a comprehensive evaluation approach, we employ three mathematical formulas the Fréchet Inception Distance (FID), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) to assess image quality and realism across datasets collected from these AI platforms. Additionally, human evaluations are conducted to compare the perceptual impact of AI-generated images with mathematical metrics. Our findings highlight the varying degrees of perceived realism among different image generative With AI models, DALL-E and Imagen generally being perceived as more realistic than Stable Diffusion and GROK. As examined, human evaluation is the current gold standard in text-to-image evaluation; however, mathematical based metrics also have promise and value. Our findings contribute to the advancement of text-to-image synthesis and our results provide support for FID as the gold standard evaluation method as it most closely represented human evaluation.

Keywords: generative AI, diffusion model, text-to-image generative model, OpenAI DALL-E, Google Imagen, GROK AI, Stable Diffusion

Introduction

The emergence of generative AI frameworks has transformed the landscape of the creation of digital image creation, marking a new era of AI to generate diverse and realistic images from text prompts. Text to image generative models are AI models that leverage Natural Language Processing (NLP) techniques while interpreting the textual inputs to visual representation (Oppenlaender, 2023). Such generative models enable users to articulate ideas and visual concepts thorough natural language, offering numerous opportunities for various applications, i.e., entertainment, art, design, and visual communication. Some renowned text-to-image generation models include DALLE, Google Imagen, GROK, and Stable Diffusion, while recent advancements in text-to-image generative models have unlocked new opportunities in creative fashion and practical applications. However, despite the remarkable potentials of AI-generated images, challenges and considerations still exist regarding their quality, realism, and societal impact.

Text-to-image synthesis is the process of generating images from textual prompts and offers both promises and complexities. For example, this technology holds potential across various domains such as design, art, entertainment, and visual storytelling. Yet there are significant concerns and challenges ensuring the accuracy, coherence, and ethical implications of AI-generated images. The reliance on textual prompts introduces complexities in accurately conveying desired visual concepts, leading to potential discrepancies between the intended and generated images. Furthermore, other concerns such as misinformation, bias, and manipulation highlight the significance of rigorous evaluation techniques and safeguards in developing and deploying text-to-image synthesis systems.

While mathematical methods can provide quantitative assessments of text-to-image synthesis, metrics such as the Fréchet Inception Distance (FID), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) offer objective benchmarks for evaluating image fidelity and similarity to real-world counterparts. These metrics employ computational algorithms to assess various aspects of image quality, including perceptual similarity, structural integrity, and noise levels. By quantifying these attributes, mathematical metrics offer valuable insights into the technical performance of AI-generated images and enable comparisons across different models and datasets.

However, despite their utility, mathematical metrics have inherent limitations. Mathematical algorithms may not fully capture the subjective nature of human perception. Human observers possess the ability to discern subtle details, interpret contextual cues, and make qualitative judgments that extend beyond numerical measurements. Thus, human evaluation remains indispensable in the assessment of image realism and perceptual fidelity based on certain criteria such as contextual coherence, emotional resonance, and aesthetic appeal. By incorporating human evaluation alongside mathematical metrics, researchers can validate the technical accuracy of AI-generated images while contextualizing their perceptual impact within real-world contexts. In essence, the integration of mathematical metrics and human evaluation represents a synergistic approach to assessing the quality and realism of text-to-image synthesis.

This study evaluates four popular AI tools, including DALL-E, Google Imagen, Stable Diffusion, and GROK AI, focusing on their text-to-image diffusion models. Ten real images were collected from diverse sources to serve as benchmarks for evaluation. Our research addresses the challenges and opportunities inherent in text-to-image synthesis through a robust evaluation approach. We explored three mathematical formulas- FID, SSIM, and PSNR metrics across the image datasets, collected from four AI platforms to measure image quality and realism. Additionally, we conducted human evaluations in which participants assessed the realism and quality of AI-generated images, enabling a comparative analysis between mathematical metrics and human perception.

In this work, a two-fold gap in the current state of the art was identified. To address this, modern diffusion models were first compared based on their performance in generating realistic text-to-image outputs. Secondly, mathematical evaluation metrics were contrasted with human assessments, which remain the gold standard. A mathematical comparison of the evaluation metrics was conducted, and a human survey was carried out to understand the performance variations of the AI tools. Through this interdisciplinary approach, we aim to contribute to understanding and advancing text-to-image synthesis while promoting responsible and ethical AI development.

Literature Review

The research landscape in text-to-image generation is still relatively nascent, with a limited number of works exploring this emerging field. While interest in text-to-image synthesis has been growing,

particularly in recent years, the volume of literature remains relatively modest compared to more established areas of machine learning and computer vision. In the context of image tuning, et al. Xu et al. (2023) introduced UniTune, an innovative approach to image editing that accepts any image along with a textual description of the desired edit. It is capable of executing the modification while preserving the original image's quality. Unlike other methods, UniTune doesn't rely on additional inputs like masks or sketches and can handle multiple edits without retraining. In evaluations against another similar potential model, SDEdit, UniTune demonstrated a clear advantage, with a 72% preference over SDEdit's 28%. These results indicate that while both methods excel when edits are minor, UniTune outperforms significantly in scenarios requiring substantial pixel alterations, such as object duplication, movement, or resizing.

Xu et al. (2023) extended the original single-flow diffusion pipeline into a versatile multi-task multimodal network called Versatile Diffusion (VD), capable of handling various tasks such as text-to-image and image-to-text conversions within a single unified model. VD comprises three key components: a diffuser that operates within a multi-flow multimodal framework, variational autoencoders (VAEs) for converting data samples into latent representations, and context encoders for embedding contextual information. In comparison to existing models like CogView, LAFITE, GLIDE, and Make-a-Scene, VD demonstrates superior FID performance, with a score of 11.21 ±0.03 compared to 11.10 ±0.09.

Elarabawy et al. (2022) developed a method for featuring multiple easily adjustable hyperparameters for enabling a diverse array of real image edits. This method, termed as optimization-free and zero fine-tuning, relies on text-based semantic instructions for flexible editing. Unlike approaches generating numerous outputs and relying on additional mechanisms for filtering, this method allows for systematic modulation of target edits.

Moreover, Ruiz et al. (2023) introduced a method called Dream Booth, which synthesizes new renditions of a subject using a small set of subject images and a text prompt as guidance. This framework is developed by fine-tuning a text-to-image model with the input images and text prompts containing a unique identifier followed by the class name of the subject. Two metrics, CLIP-I and DINO were used to evaluate the performance. Dream Booth (based on the Imagen model) achieves higher scores for both subject and prompt fidelity compared to Dream Booth (based on Stable Diffusion), nearing the upper limit of subject fidelity achievable with real images.

For image synthesis, several research have been carried on blended diffusion models which combine ideas from diffusion models and autoregressive models to achieve high-quality image generation (Avrahami et al., 2023; Avrahami et al., 2022). A latent diffusion model (LDM) was designed by (Avrahami et al., 2023; Avrahami et al., 2022) to accelerate the diffusion process by functioning within a lower-dimensional latent space. This design eliminates the necessity for resource-intensive CLIP gradient computations at each diffusion step, thereby enhancing efficiency without compromising on the quality of image synthesis.

Numerous researchers have directed their attention toward transformer-based applications to achieve rapid and high-resolution image synthesis (Esser et al., 2021; Gafni et al., 2022; Gal et al., 2022). Ding et al. (2022) designed a solution leveraging hierarchical transformers and local parallel autoregressive generation. They pretrained a 6-billion-parameter transformer using a straightforward and adaptable self-supervised task, namely a cross-modal general language model called CogLM, and further refined it for swift super-resolution tasks. Their novel text-to-image system, CogView2, exhibits highly competitive generation capabilities when compared to contemporary state-of-the-art models like DALL-E-2, while also inherently supporting interactive text-guided editing of images.

There is another potential technique from NLP, called transformer-based model, a revolutionary advancement in natural language processing, has significantly enhanced the capabilities of language understanding and generation. Using transformer-based models, text-to-image generation has seen remarkable progress, enabling the creation of vivid and realistic visual content from textual descriptions (Ding et al., 2021; Naveen et al., 2021). Transformer-based models hold immense promise across diverse fields, revolutionizing natural language processing, image generation, and beyond (Jamal & Wimmer, 2023; Saifullah et al., 2024).

Wu et al. (2022) worked on object-guided joint decoder type transformer to generate images from natural language commands. Another two-stage model has been developed by Ramesh et al. (2022) where the first stage generates a CLIP image embedding from a provided text caption, and the second stage consists of a decoder that generates an image conditioned on this embedding. Moreover, Latent diffusion models offer significant advantages for tasks such as image inpainting, class-conditional image synthesis, as well as competitive performance across various other tasks like unconditional image generation, text-to-image synthesis, and super-resolution. In one such study, Rombach et al. (2022) analyzed the behavior of their latent diffusion models (LDMs) across different down sampling factors. These models are denoted as LDM-f, where LDM-1 corresponds to pixel-based diffusion models. This LDM (Latent Diffusion Model) has achieved new state-of-the-art scores for tasks such as image inpainting and class-conditional image synthesis. Moreover, it demonstrates highly competitive performance across a range of tasks, including unconditional image generation, text-to-image synthesis, and super-resolution.

Despite the growing interest and advancements in text-to-image generation, the research landscape in this domain remains in the early stage, with several areas requiring further exploration. Current literature highlights significant progress in the development of various models and techniques for text-to-image synthesis, including UniTune for image editing, Versatile Diffusion (VD) for multi-modal tasks, and Dream Booth for subject-specific renditions. Additionally, methods such as blended diffusion models and transformer-based applications have shown promising results in achieving high-quality image generation. However, a notable gap exists in comprehensive comparative evaluations of these models using both mathematical and human-centered approaches. To address this gap, our study presents novel contributions in conducting comprehensive comparison of text-to-image generative diffusion models while providing a holistic understanding of their weaknesses and strengths. While previous research predominantly focused on mathematical metrics, we incorporated human evaluations to assess the perceptual quality and realism of AI-generated images. Participants in our study evaluated the images based on contextual coherence, emotional resonance, and aesthetic appeal, providing a comprehensive understanding of the performance variations of different AI tools. This comprehensive evaluation not only addresses the identified research gaps but also sets a new standard for future research in the field.

Methods

This study explores text-to-image diffusion models, focusing on four prominent text-to-image generation models: DALL-E, Google Imagen, Stable Diffusion, and GROK AI. Ten real images were collected from diverse sources to serve as benchmarks for evaluation. Utilizing identical text prompts, ten sets of images were generated using each of the four platforms. The null hypothesis suggests that there is no significant distinction between the generated images and real images. To investigate this, another alternative research hypotheses has been initially formulated. That suggests that DALLE and Google Imagen produce more realistic images than the other platforms. Two techniques were employed to evaluate the generated images: computer evaluation and human evaluation via a survey. For computer evaluation, metrics such as FID

score, PSNR, and SSIM were computed for each set of generated images in comparison to the real ones. The results were analyzed and further validated through a survey involving thirty participants. The overall methodology is presented in figure 5.1.

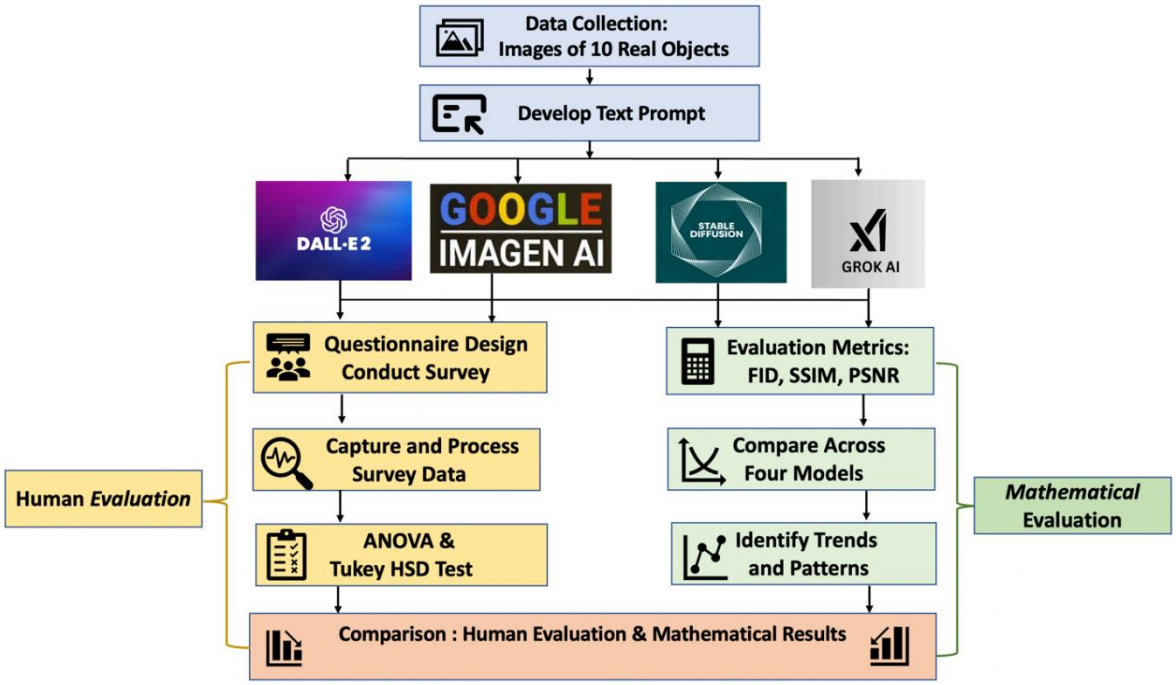

Figure 1: Overall Method Flow Diagram

Text to image diffusion models

The diffusion process in image generation refers to a method where noise is gradually added to an initial image through multiple steps. The noise is smoothed or diffused at each step, resulting in a new image that becomes increasingly distorted or randomized (Yang et al., 2023). This process is typically guided by a diffusion model, which defines how the noise is applied and how it evolves over time. An ideal text-to-image diffusion model follows the below general steps:

Step 1: Text Embeddings:

- At first the text descriptions are converted into a dense vector representation, which is known as text embedding, using techniques such as pre-trained language models (i.e., BERT, RoBERTA) or word embeddings.

Step 2: Diffusion Process:

- The diffusion process involves iteratively applying noise to the input text embeddings to generate intermediate noisy embeddings (Ho et al., 2022).
- These noisy embeddings are then passed through a diffusion model, which gradually improves the image features to generate more realistic images.

Step 3: Image Generation:

- The final refined embeddings are passed through a decoder network to generate images which correspond to the original textual description.

- Theis decoder network may consist of convolutional neural networks (CNNs), or other architectures tailored for particular image generation (Zhu et al., 2023).

Step 4: Training:

- During this training process, the model learns to minimize the difference between the generated and ground truth images corresponding to the input text descriptions.
- The diffusion model and decoder network parameters are optimized using techniques like maximum likelihood estimation or adversarial training.

Step 5: Evaluation:

- The quality of the generated images is evaluated using metrics such as FID score, SSIM, or human perceptual studies to assess realism and coherence with the input text descriptions.

DALL-E

DALL-E, stands for "Distribute Aggregate Linear Latent Encoder," is a groundbreaking text-to-image generation model developed by OpenAI. DALL-E is built upon the transformer-based architecture, which is similar to the one used in the GPT series of models, to encode the textual description into a dense vector representation (Marcus et al., 2022). This text embedding serves as conditioning information for any image generation process. Unlike traditional image generation models that rely on continuous latent spaces, DALL-E introduces a discrete latent space via a Vector Quantized Variational Autoencoder (VQ-VAE-2) model. Such discrete latent space allows for the representation of diverse image features in a compact and interpretable manner. One of the key strengths of DALL-E is producing highly diverse and semantically meaningful images based on a wide range of textual prompts. By leveraging the rich semantic representations encoded in the text embeddings, DALL-E can capture intricate details and nuances in the generated images, ranging from simple objects to complex scenes and abstract concepts. Additionally, the discrete latent space introduced by the VQ-VAE-2 model allows for fine-grained control over the generated images, enabling users to manipulate various visual attributes such as style, color, and compositions.

Google Imagen

Google Imagen is a text-to-image generation platform developed by Google, which leverages advanced machine-learning techniques for image synthesis from textual descriptions (Wang et al., 2023). While specific details of the architecture are not publicly disclosed, Google Imagen likely employs transformer-based models similar to those used in DALL-E or GPT for text encoding and generation. The platform may incorporate additional components for image synthesis, such as attention mechanisms or generative adversarial networks (GANs), to enhance the realism and diversity of generated images.

Stable Diffusion

Stable Diffusion, introduced in 2022, stands as a prominent deep learning model within the ongoing surge of AI advancements. Utilizing diffusion techniques, it primarily functions as a text-to-image model, capable of generating intricate visuals based on textual inputs. Beyond image generation, Stable Diffusion demonstrates versatility across tasks such as inpainting, outpainting, and facilitating image-to-image translations guided by text prompts. Additionally, it offers an "img2img" sampling script, which takes a text prompt, an existing image path, and a strength value ranging from 0.0 to 1.0. This script outputs a new image derived from the original one, integrating elements specified in the text prompt. The strength parameter controls the level of noise injected into the output image, influencing the degree of variation. Higher strength values yield more diverse images but may stray from semantic coherence with the provided prompt.

GROK AI

GROK AI is another prominent text-to-image generation platform that utilizes deep learning techniques based on textual descriptions for image synthesis. GROK AI employs transformer-based models or similar architectures for text encoding and image generation. The platform incorporates custom components or optimizations tailored to the text-to-image generation task to enhance the quality and fidelity of generated images. GROK AI aims to provide users with a seamless experience for creating realistic images from textual prompts, leveraging state-of-the-art machine learning algorithms for efficient and effective image synthesis.

Experiment

We collected real images from 10 subjects. Then, using a specific similar text prompt, similar images were generated using AI-based image generation tools: DALL-E, Imagen, Stable Diffusion, and GROK. Consequently, for each subject, a set of five images was compiled, comprising the original real image alongside four AI-generated images. For instance, Figure 2a and 2b shown below diplays a real image featuring a cat swimming, accompanied by four additional images generated by AI tools. This process was repeated across all subjects, resulting in a comprehensive collection of images for evaluation and analysis.

Figure 2a: Input Image of a Cat Swimming

Figure 3b: Input Image of a Cat Swimming

Evaluation

Two distinct evaluation methods were employed to assess the realism of the generated images. Method A involved mathematical evaluation, where metrics such as Fréchet Inception Distance (FID) score, Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) ratios were calculated using Python frameworks and libraries. Method B consisted of a human study conducted via a survey, with data gathered

from 28 subjects. Statistical evaluation was performed on the collected data, including ANOVA tests and Tukey post hoc analysis, to determine the significance of each group. These two evaluation methods will be elaborated upon in the following sections.

Method A: Mathematical Evaluation

The Fréchet Inception Distance (FID) score is a metric commonly used to evaluate the quality of generated images in generative adversarial networks (GANs) and other image generation models. It measures the similarity between the distributions of real and generated images in a feature space learned by an Inception-v3 neural network (Obukhov & Krasnyanskiy, 2020). The FID score is computed based on the mean and covariance of feature representations of real and generated images. Given two sets of feature representations μ_r and μ_g (mean); $\sum r$ and $\sum g$ (covariance) for real and generated images respectively, the FID score is calculated using the following formula:

$$FID = \left|\mu_r - \mu_g\right|^2{}_2 + T_i\left(\sum r + \sum g - 2\left(\left(\sum r \sum g\right)^{\frac{1}{2}}\right)\right)$$

Where:

- $\left|\mu_r - \mu_g\right|^2{}_2$ presents the squared Euclidean distance between the mean feature vectors.
- T_i represents the trace of matrix.
- $\sum r$ and $\sum g$ are the covariance matrices of real and generated feature representations, respectively.

The Structural Similarity Index (SSIM)

It is a metric used to quantify the similarity between two images, considering their luminance, contrast, and structure. It measures the perceptual difference between the original and processed images (Ponomarenko et al., 2018). SSIM is calculated using three components: luminance comparison, contrast comparison, and structure comparison. The overall SSIM score is the product of these three components.

The formula for SSIM is as follows:

$$SSIM(x,y) = \frac{l(x,y)\, c(x,y)}{M1\, M2}$$

Here, $l(x,y) = \frac{2\,\mu_x\,\mu_y + C1}{{\mu_x}^2 + {\mu_y}^2 + C1}$

$$C(x,y) = \frac{2\,\sigma_x\,\sigma_y + C2}{{\sigma_x}^2 + {\sigma_y}^2 + C2}$$

$C1$ and $C2$ are constants here and the values are chosen to avoid zero instability.
x and y are the two images that are being compared.
A code snippet of this calculation using Python programming is attached in Figure 3.

Peak Signal-to-Noise Ratio (PSNR)

This metric is commonly used to evaluate the quality of reconstructed or compressed images. It measures

the difference between the original image and its approximation, considering the difference's magnitude and the image's maximum possible range. The PSNR is calculated based on the Mean Squared Error (MSE) between the original image I and the reconstructed I' in decibels (dB). When the maximum possible pixel value is MAX, i and j are image dimensions, the formula for PSNR is as follows:

$$PSNR = 10.log_{10}\left(\frac{MAX^2}{MSE}\right)$$

Here, $MSE = \frac{1}{ij}\sum_{k=0}^{i=-1}\sum_{l=0}^{n=-1}[I(k,n) - I'(k,n)]^2$

```
def calculate_frechet_distance(mu1, sigma1, mu2, sigma2):
    epsilon = 1e-6

    # Numerical stability.. tried this to handle negative value issue!
    sqrtm_term = np.real(linalg.sqrtm(np.dot(sigma1, sigma2)))

    # Handle small eigenvalues
    if np.iscomplexobj(sqrtm_term):
        sqrtm_term = sqrtm_term.real

    fid = np.linalg.norm(mu1 - mu2) + np.trace(sigma1 + sigma2 - 2 * sqrtm_term + epsilon)

    return fid
```

```
[ ] def calculate_fid(real_images, generated_images, model):
    real_mu, real_sigma = calculate_activation_statistics(real_images, model)
    generated_mu, generated_sigma = calculate_activation_statistics(generated_images, model)
    fid = calculate_frechet_distance(real_mu, real_sigma, generated_mu, generated_sigma)
    return fid
```

Figure 4: Code snippet of FID Score Calculation

Necessary libraries are imported, including numpy for numerical computations, skimage.metrics for SSIM calculation, and tensorflow for PSNR calculation. A function called calculate_metrics is defined for calculating the metrics, which takes the real image and generated image as input. Inside the function, the FID score is first calculated using a function called calculate_fid. This function, assumed to be defined elsewhere, takes real and generated images along with an inception model as input. The SSIM score is then obtained using the SSIM function from skimage.metrics. This function computes the structural similarity between two images. Finally, the PSNR score is evaluated using the tf.image.psnr function from TensorFlow. PSNR measures the quality of an image in terms of signal-to-noise ratio. These scores are provided as quantitative assessments of the realism and quality of the generated images, complementing the human evaluation conducted in the study. The average FID, SSIM and PSNR scores are noted in table 1 below.

Table 1: Mathematical Evaluation Metrics

AI Tool\ Metrics	FID	SSIM	PSNR
DALLE	9.00%	1.35%	9.88
IMAGEN	10.43%	0.86%	10.20
GROK	10.69%	1.50%	10.51
Stable Diffusion	15.95%	0.95%	9.21

These three values have been illustrated in the graph Figure 4 to get the comparison results among three AI tools. As the lower percentages of FID indicate better similarity between real and generated images, implying higher quality and realism, DALL-E has the lowest FID score (9.0%) indicating better similarity with real images. On the other hand, the FID score is higher in Stable Diffusion (15.95%) than the other three. Moreover, the higher SSIM and PSNR score mean the better similarity with the real image. In this case, DALL-E has the highest value compared to other three suggesting its generated images have better quality and realism .

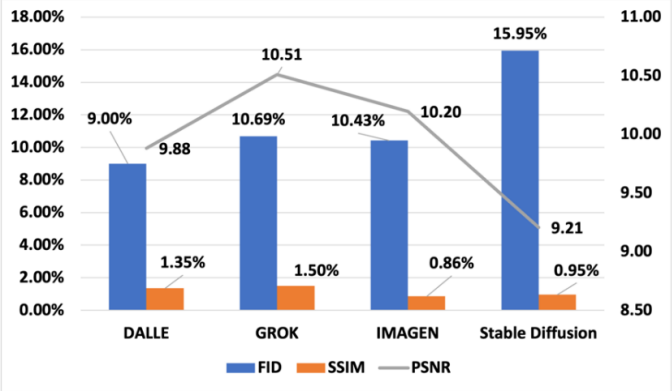

Figure 5: Comparison of Mathematical Metrics

Method B: Human Evaluation

In this study, as a part of human evaluation, a survey was developed and administered using the Qualtrics online survey platform to gather subjective evaluations of the images generated by human participants. The participants were asked to provide ratings for image realism on a scale of 1 to 7. The survey consisted of 10 sets of questions, with each set containing five sets of images. Participants were asked to rate the realism of the images, with 1 indicating less realistic and 7 indicating more realistic. The responses collected from the survey were structured and grouped based on five categories, including real images and images generated by DALL-E, Imagen, Stable Diffusion, and GROK AI. Each category represented a different image generation model.

Survey Result and Statistical Analysis

A total of 28 subjects responded during the survey comprising adults aged 18 to 44, with a gender distribution of 17 males and 11 females, all of whom were citizens of the United States of America. Each participant was randomly allocated four out of ten sets of questions, providing a diverse perspective on image realism assessment across different age groups and genders. Statistical analyses were conducted using ANOVA tests and Tukey's Honest Significant Difference (HSD) test to determine the significance of differences among the groups. Figure 5 presents a snapshot of survey question.

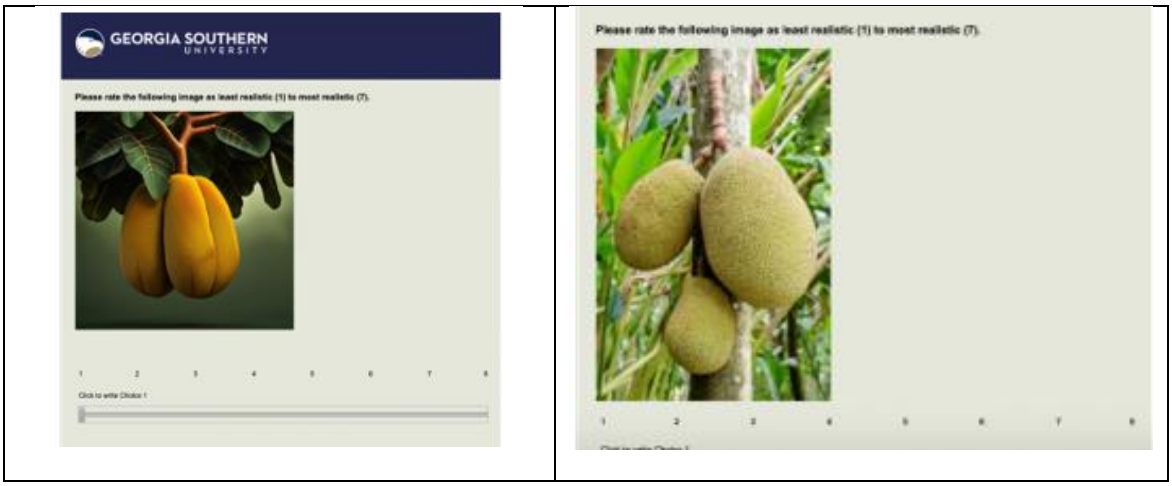

Figure 5: Survey Questions Snapshot

Table 2 summarizes the one-way ANOVA test result, including each group's average and variance.

Table 2: Group Summary

Groups	Count	Sum	Avg	Var
DALL-E	112	639	5.71	4.03
Stable Diffusion	112	535	4.78	4.90
Imagen	112	642	5.73	3.68
GROK	112	458	4.09	4.95
Real Image	112	650	5.80	4.11

The difference between groups from ANOVA test is tabulated in table 3. A significance value less than the chosen alpha level (e.g., 0.05) indicates that the differences between group means are statistically significant. In this case, F= 14.856, with a significance value of 0.000, suggests that the differences between group means are statistically significant. Therefore, we can reject the null hypothesis and conclude that there are significant differences in image realism ratings among the different image generation models.

Table 3: Group Differences

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	257.619	4	64.405	14.856	0.000
Within Groups	2401.755	554	4.335		
Total	2659.374	558			

For further understanding the variance among groups, Tukey's Honest Significant Difference (HSD) Test have been conducted. The results are noted in table 4. Each pair of groups is compared in this test, and the mean difference, standard error, significance level, and confidence interval are provided.

Table 4. Tukey HSD Result

Multiple Comparisons						
Tukey HSD						
(I) 1.000000		**Mean Difference (I-J)**	**Std. Error**	**Sig.**	**95% Confidence Interval**	
					Lower Bound	Upper Bound
1 DALL-E	2	**.917***	0.279	0.009	0.15	1.68
	3	-0.038	0.279	1.000	-0.80	0.72
	4	**1.604***	0.279	0.000	0.84	2.37
	5	-0.110	0.279	0.995	-0.87	0.65
		Mean Difference (I-J)	**Std. Error**	**Sig.**	**95% Confidence Interval**	
					Lower Bound	**Upper Bound**
2 Stable Diffusion	1	**-.917***	0.279	0.009	-1.68	-0.15
	3	**-.955***	0.278	0.006	-1.72	-0.19
	4	0.688	0.278	0.099	-0.07	1.45
	5	**-1.027***	0.278	0.002	-1.79	-0.27
3 Imagen	1	0.038	0.279	1.000	-0.72	0.80
	2	**.955***	0.278	0.006	0.19	1.72
	4	**1.643***	0.278	0.000	0.88	2.40
	5	-0.071	0.278	0.999	-0.83	0.69
4 GROK	1	**-1.604***	0.279	0.000	-2.37	-0.84
	2	-0.688	0.278	0.099	-1.45	0.07
	3	**-1.643***	0.278	0.000	-2.40	-0.88
	5	**-1.714***	0.278	0.000	-2.48	-0.95
5 Real Image	1	0.110	0.279	0.995	-0.65	0.87
	2	**1.027***	0.278	0.002	0.27	1.79
	3	0.071	0.278	0.999	-0.69	0.83
	4	**1.714***	0.278	0.000	0.95	2.48
* The mean difference is significant at the 0.05 level.						

The highlighted values are less than 0.05, indicating that the group difference is significant in these cases and are illustrated as comparative graphs in figures 5.6 and 5.7.

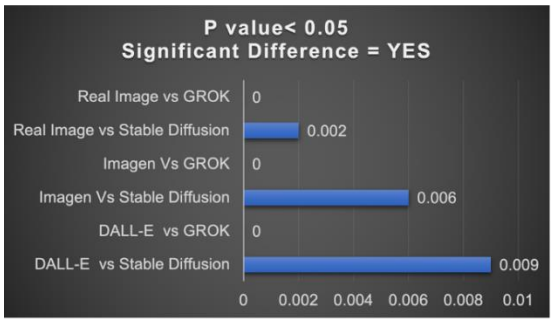

Figure 6: Groups with Significant Difference

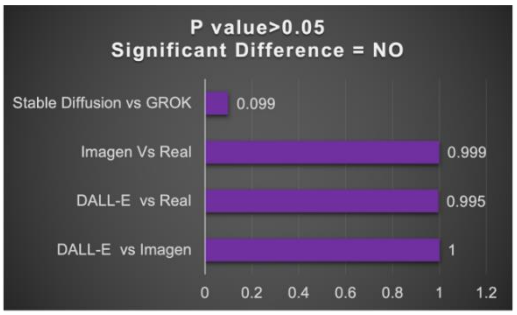

Figure 7: Groups with No Significant Difference

Discussion

Our study presents a thorough analysis of the performance of different text-to-image generative AI models, positioning these valuable findings within the existing research context. In terms of mathematical evaluation, lower FID scores and higher SSIM and PSNR values generally indicate better image quality and similarity to real images. Referring to table 1, DALL-E exhibited the most promising performance, with the lowest FID score (9.0%) and highest SSIM (1.35%) and PSNR values (9.88), suggesting its capability to produce images that closely resemble real images.

Conversely, Stable Diffusion demonstrated the highest FID score (15.95%) and lowest PSNR (9.21%) value, indicating potential limitations in generating realistic images compared to other models. These results provide valuable insights into the relative performance of these AI models in generating high-quality images.

Furthermore, the statistical analysis of the human perception survey data using Tukey's HSD test revealed significant differences in perceived realism between certain pairs of image-generative AI models. Like the mathematical evaluation, DALL-E showed significant differences in perceived realism compared to Stable Diffusion ($p = 0.009$), GROK ($p = 0$), and Real Image ($p = 0.002$), indicating that images generated by DALL-E were perceived as more realistic than those generated by these models. Similarly, Imagen did not exhibit significant differences in perceived realism compared to Real Image ($p = 0.999$), indicating that the perceived realism of images generated by Imagen was comparable to those generated by Real Image.

In summary, according to the participants' perceptions, the findings highlight the varying degrees of perceived realism among different image generative AI models, with DALL-E and Imagen generally being perceived as more realistic than Stable Diffusion and GROK. As examined, human evaluation is the current gold standard in text-to-image evaluation; however, mathematical based metrics also have promise and value. FID is growing as the standard evaluation method, and our results illustrate it most closely represented human evaluation. However, this underscores the importance of considering both objective metrics and subjective human perception in evaluating the performance of image generative AI models.

The implications of these findings for future research are substantial. First, there is a clear need for further studies that integrate both mathematical and human evaluations to provide a holistic assessment of AI models' performance. Future research should continue exploring how these evaluation methods can be

combined to offer a more comprehensive understanding of model capabilities. Additionally, the performance gap identified between models such as DALL-E and Stable Diffusion highlights the importance of ongoing model refinement and innovation.

For practical applications, these results underscore the value of using models like DALL-E and Imagen for tasks that require high levels of realism, such as digital art, entertainment, and design. Practitioners should consider the strengths and weaknesses of different models based on both objective metrics and subjective human evaluations to select the most appropriate tool for their specific needs.

Future studies should aim to include larger and more diverse participant groups to enhance the robustness of the results. Secondly, while we employed well-established mathematical metrics, these do not capture all aspects of image quality and may overlook subtleties that humans can perceive. Finally, the study focused on a limited set of AI models; expanding the range of models evaluated could provide a more comprehensive understanding of the current state of text-to-image generation technologies.

Conclusion

The demand for AI-generated images continues to rise across various domains, understanding the capabilities and limitations of these models is crucial. By combining mathematical evaluation with human perception studies, we have comprehensively understood the relative performance of prominent text-to-image generative AI models. The mathematical evaluation of the image generative AI models indicates that lower FID scores and higher SSIM and PSNR values generally correspond to better image quality and similarity to real images. FID has emerged as a robust metric, often aligning closely with human perception as observed in our survey data.

Therefore, FID could be considered a superior metric for mathematical evaluation compared to SSIM and PSNR. Additionally, the statistical analysis of the human perception survey data using Tukey's HSD test unveiled significant differences in perceived realism between certain pairs of image-generative AI models. DALL-E showed significant differences in perceived realism compared to Stable Diffusion ($p = 0.009$), GROK ($p = 0$), and Real Image ($p = 0.002$), suggesting that images generated by DALL-E were perceived as more realistic. Conversely, Imagen did not exhibit significant differences in perceived realism compared to Real Image ($p = 0.999$), indicating comparable perceived realism between images generated by Imagen and Real Image.

The findings highlight varying degrees of perceived realism among different image-generative AI models, with DALL-E and Imagen generally perceived as more realistic than Stable Diffusion and GROK. This underscores the importance of considering both objective metrics and subjective human perception in evaluating the performance of image-generative AI models.

Acknowledgements

This work was supported in part by the National Science Foundation (USA) under Grant no. 2321939

References

Abdi, H., & Williams, L. J. (2010). Tukey’s honestly significant difference (HSD) test. *Encyclopedia of research design, 3*(1), 1-5.

Avrahami, O., Fried, O., & Lischinski, D. (2023). Blended latent diffusion. *ACM Transactions on Graphics (TOG), 42*(4), 1-11.

Avrahami, O., Lischinski, D., & Fried, O. (2022). Blended diffusion for text-driven editing of natural images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., & Yang, H. (2021). Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems, 34*, 19822-19835.

Ding, M., Zheng, W., Hong, W., & Tang, J. (2022). Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems, 35*, 16890-16902.

Elarabawy, A., Kamath, H., & Denton, S. (2022). Direct Inversion: Optimization-Free Text-Driven Real Image Editing with Diffusion Models. *arXiv e-prints*, arXiv: 2211.07825.

Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., & Taigman, Y. (2022). Make-a-scene: Scene-based text-to-image generation with human priors. European Conference on Computer Vision,

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., & Salimans, T. (2022). Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research, 23*(1), 2249-2281.

Jamal, S., & Wimmer, H. (2023). An improved transformer-based model for detecting phishing, spam, and ham: A large language model approach. *arXiv preprint arXiv:2311.04913*.

Kaufmann, J., & Schering, A. (2007). Analysis of variance ANOVA. *Wiley Encyclopedia of Clinical Trials*.

Marcus, G., Davis, E., & Aaronson, S. (2022). A very preliminary analysis of DALL-E 2. *arXiv preprint arXiv:2204.13807*.

Naveen, S., Kiran, M. S. R., Indupriya, M., Manikanta, T., & Sudeep, P. (2021). Transformer models for enhancing AttnGAN based text to image generation. *Image and Vision Computing, 115*, 104284.

Obukhov, A., & Krasnyanskiy, M. (2020). Quality assessment method for GAN based on modified metrics inception score and Fréchet inception distance. Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4,

Oppenlaender, J. (2023). A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, 1-14.

Ponomarenko, M., Egiazarian, K., Lukin, V., & Abramova, V. (2018). Structural similarity index with predictability of image blocks. 2018 IEEE 17th International Conference on Mathematical Methods in Electromagnetic Theory (MMET),

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125, 1*(2), 3.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Saifullah, K., Khan, M. I., Jamal, S., & Sarker, I. H. (2024). Cyberbullying Text Identification based on Deep Learning and Transformer-based Language Models. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 11*(1), e5-e5.

Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D. J., & Soricut, R. (2023). Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Wu, F., Liu, L., Hao, F., He, F., & Cheng, J. (2022). Text-to-image synthesis based on object-guided joint-decoding transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Xu, X., Wang, Z., Zhang, G., Wang, K., & Shi, H. (2023). Versatile diffusion: Text, images and variations all in one diffusion model. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys, 56*(4), 1-39.

.Zhu, Y., Li, Z., Wang, T., He, M., & Yao, C. (2023). Conditional Text Image Generation with Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,