GAIA: A Benchmark of Analyzing User Rankings for Synthesized Images

Kriti Sharma[1], Thomas Sherk[1], Vatsa S. Patel[1], Minh-Triet Tran[2,3], and Tam V. Nguyen[1](✉)

[1] Department of Computer Science, University of Dayton, Dayton, USA
tamnguyen@udayton.edu
[2] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[3] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. Text-to-image models which are a part of Generative AI have become essential tools for digital artists and enthusiasts to create visually captivating images. These models have garnered significant attention and rapid advancements in recent years, enabling the creation of realistic and visually appealing images from textual descriptions. However, assessing the quality of these generated images remains a challenging task due to varying perceptions of image quality. Additionally, generated images often lack clear ground truth and the intricate details that capture human attention. To address these challenges in the study of artificially generated images, we introduce a novel approach with the Generative Artificial Image Assessment (GAIA) dataset. This dataset includes images from eight popular text-to-image AI models along with user rankings. GAIA is evaluated and predicted by pre-trained state-of-the-art networks using ranking classes and a regression technique to analyze the images. Our approach combines objective evaluation metrics, subjective human judgment, benchmark datasets with diverse ground truth annotations, and advancements in multimodal learning techniques. This comprehensive methodology provides a pathway to advancing the field of text-to-image generation.

Keywords: Generative AI · Image Synthesis · Deep Learning · Text-to-Image

1 Introduction

With the rapid advancements in artificial intelligence, text-image generation models such as Generative Adversarial Networks (GANs) [1] and the Contrastive Language-Image Pretraining (CLIP) [2] models have gained significant attention for their ability to create realistic and visually appealing images from textual descriptions. Evaluating the quality of these generated images, however, remains a difficult task, as these models are complex, some are proprietary, and the datasets are large. Moreover, the inclusion of neural networks/transformer architecture can also obfuscate the underlying decision process used to generate the image.

To address this challenge, this paper introduces a novel dataset, Generative Artificial Image Assessment (GAIA), and further analyses of user rankings for synthesized

G. Bebis et al. (Eds.): ISVC 2024, LNCS 15046, pp. 451–463, 2025.
https://doi.org/10.1007/978-3-031-77392-1_34

images generated by the following eight generative AI models: Adobe Firefly [3], DeepDream Generator [4], Artbreeder Mixer [5], DeepAI [6], Starry AI [7], Picsart AI Image Generator [8], Stability AI [9], and Midjourney [10]. While the dataset is one of the first efforts in this research problem, it attempts to provide a meaningful amount of data that is unbiased from the user generating. We will further provide details regarding the composition of the dataset and prescribed analyses on the newly collected dataset.

The remainder of this paper is organized as follows. Section 2 summarizes related work. Section 3 introduces the dataset collection and user ranking. Section 4 presents the analysis from the dataset. Finally, Sects. 5 concludes the paper and paves the way for future work.

2 Related Work

Numerous investigations have examined the assessment of images produced by artificial intelligence (AI) generators, offering significant perspectives on the caliber and attractiveness [11] of images generated by various generative AI instruments [12, 13]. Our study extends beyond these parameters by introducing diverse criteria rooted in human perception. Previous work [14, 15] focuses on using crowdsourced subjective ratings to assess the quality and appeal of AI-generated images. Notably, the study compares the effectiveness of various AI generators and emphasizes the value of crowdsourcing for a range of viewpoints. It is crucial to remember that the study does not provide a thorough examination of actual human perception; instead, it focuses primarily on quality and appeal [16].

Beyond the traditional method, we explore various dimensions including emotional resonance, contextual relevance, and visual coherence in AI-generated images. Prompt engineering [17] in prior studies highlights the importance of the practitioner's interaction with AI systems and raises questions about possible effects on human creativity in the process of creating art. In contrast, our assessment is carried out independently of prompt partitioning, which may result in the omission of subtleties in the generative process according to prompt complexity. This method provides a more comprehensive assessment of the generative process by considering the complexities of human perception and prompts.

In assessing image quality, there is much research into the prevalent focus on technical metrics like semantic object accuracy (SOA), inception score (IS) [18], and feature similarity index (FSIM) [19] to match the captions with the corresponding image generated using different evaluation techniques. Additionally, much research compares well-established IQA methods, including peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), mean structural similarity index (MSSIM), and feature similarity index (FSIM) [20]. Our research attempts to provide a more nuanced understanding of the generative AI image assessment landscape by adopting a comprehensive set of evaluative criteria, capturing subtleties that may have been missed in earlier studies. This method provides a more comprehensive assessment of the generative process by considering the complexities of human perception and prompts.

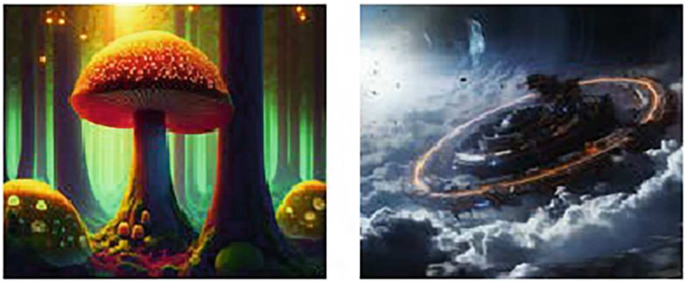


Fig. 1. Illustration of two images generated by different prompts: An enchanted forest with glowing mushrooms (left) by Deep Dream Generator [4] and prompt: A futuristic space station orbiting a gas giant, surrounded by swirling storms (right) generated by Midjourney [10].

3 Dataset Collection

The objective of the collected Generative Artificial Image Assessment (GAIA) dataset is to provide a set of images generated by what we consider readily available and publicly recognized text-to-image generation models. Although it is possible for other models to exist elsewhere on the internet, these are the models that we found and are suitable for this dataset.

The GAIA dataset consists of two sets of images, generated by prompts categorized as ‘simple’ or ‘complex’ as well as user rankings. In total, 100 prompts were used together with the 8 models to generate 800 images. An example of the two images generated from different types of prompts is shown in Fig. 1 in the paper. The labels are added to keep a record of the prompt number and the model used to download the images. To ensure label clarity, each tool and each prompt were given a unique identifier as well as qualifying information such as filename and prompt.

For the user ranking, many participants were asked to rank each prompt-image for the following criteria: prompt similarity, realism, aesthetics, and visual quality. The process of dataset collection is summarized in Fig. 2. To ensure label clarity, each tool and each prompt were given a unique identifier as well as qualifying information such as filename and prompt which was stored in tabular format.

3.1 Prompt Analysis

A specific examination of prompts is conducted to discern which ones are considered easy or difficult for the generative AI tools. This analysis provides insights into the challenges associated with certain textual prompts.

To obtain the prompts, we first used the popular AI language model, ChatGPT [21] to generate an initial pool of 200 prompts. Utilizing AI for prompt generation allowed us to ensure a focused and relevant dataset, which provides a solid foundation for subsequent analysis and experimentation in our research. After the initial pool was generated, we meticulously refined it to arrive at a final subset of 100 prompts by removing duplicate or short prompts. This process enabled us to systematically filter and curate prompts, ensuring a focused and relevant dataset for further analysis and experiments.

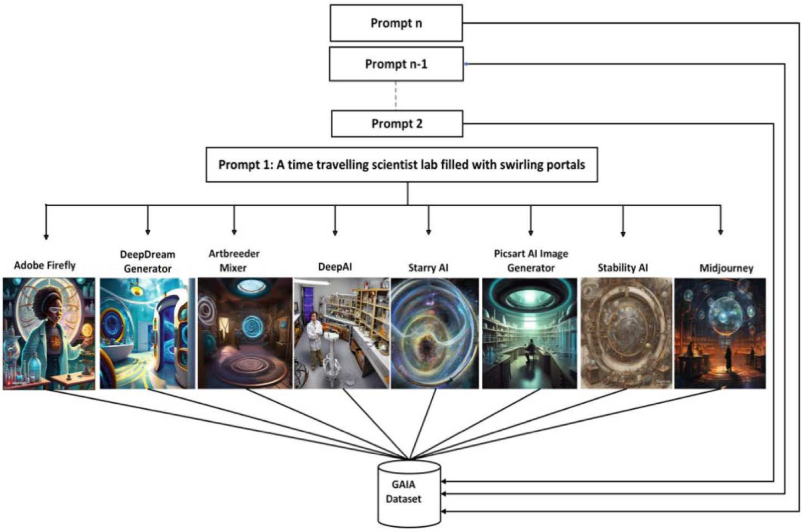


Fig. 2. The flowchart of GAIA dataset collection.

Table 1. Average lengths in different categories for Simple and complex Prompt

Category	Simple Prompt	Complex Prompt
Nature	42.0	61.58
Architecture	34.4	66.60
Historic	37.7	65.77
Human Related	44.5	63.52

To ensure diversity and robustness in the evaluation process, 100 prompts were categorized into two sets of 50 each – simple and complex. The complexity was determined by considering the following criteria: 1) Atypical Configurations 2) Creative Aspects and Complex Scenarios 3) Combining Various Concepts 4) Non-Traditional Mixtures 5) Imaginative Logic or Physics.

The main distinction is based on the respective lengths of simple and complex prompts. The complexity of the prompt is characterized by the increased length, contributing to the intricacy of the content it expresses. An analysis to compare the quantitative measure is conducted by comparing the average length of simple and complex prompt text where the average length of the complex prompts is almost double the average length of the simple prompts. This highlights the relationship between the inherent intricacies embedded in the conveyed information and the length of the prompts. To further compare the complexity of the prompts, the prompts are divided into 4 different categories named- Nature, Architecture, Historic, and Human related prompts where the nature category contains all the prompts related to flora and fauna, landscapes, wildlife, ecosystems, and environmental phenomena. Architecture prompts focus on architectural design, structures, and the built environment. Historic prompts delve into events, figures, and periods of historical significance. Finally, the human-related prompts consist of prompts with human inclusion in the scenarios for both simple and complex cases.

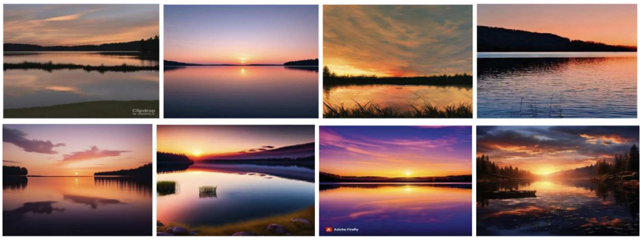


Fig. 3. Examples of images generated by *'simple'* prompt: *A serene sunset over a calm lake.* From left to right, top to bottom: Stability AI [9], Picsart [8], DeepAI [6], Artbreeder Mixer [5], Deepdream Generator [4], Starry AI [7], Adobe Firefly [3], and Midjourney [10].

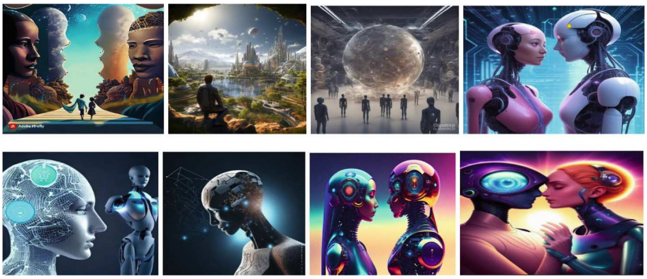


Fig. 4. Examples of Images Generated by *'complex'* prompt: *A post-singularity world where AI and humans coexist in harmony*. From left to right, top to bottom: Adobe Firefly [3], Midjourney [10], Stability AI [9], Picsart [8], DeepAI [6], Artbreeder Mixer [5], Deep-dream Generator [4], Starry AI [7].

In Table 1, the average lengths of the prompts are shown category-wise. Comparing the simple and complex averages within each category evaluates how complexity varies across different subject domains. It provides a comparison framework to investigate the relative levels of complexity within and across various theme areas.

The simple text is a straightforward scenario with standard setups and low complexity. It acts as a benchmark for assessing how generative AI reacts to traditional and understandable inputs. There are no imaginative mixtures or any combination of extreme concepts. In Fig. 3, there are examples of the images generated from a simple prompt: "A serene sunset over a calm lake." We can see the different results generated from different models.

The complex text is set with elevated levels of complexity, featuring imaginative and futuristic examples. The complicated prompt presents multiple levels of complexity, such as shown in Fig. 4: the idea of the post-singularity, the peaceful cohabitation of AI and humans, and the futuristic world's hypothetical nature. This example tests the ability of generative AI systems to handle complex situations, artistic elements, and unconventional combinations.

3.2 Assessment Criteria

The four criteria, namely, Prompt Similarity, Realism, Aesthetics, and Visual Quality, are individually evaluated to understand the strengths and weaknesses of each generative AI tool. The definition of each criterion is as follows:

Criteria 0: **Prompt Similarity** - This criterion measures how closely an image conforms to the provided prompt, whether it is a simple text or complex. The guidelines in the tool to rank the images are such that the user determines how closely the visual content complies with the themes and subject matter in the prompts.

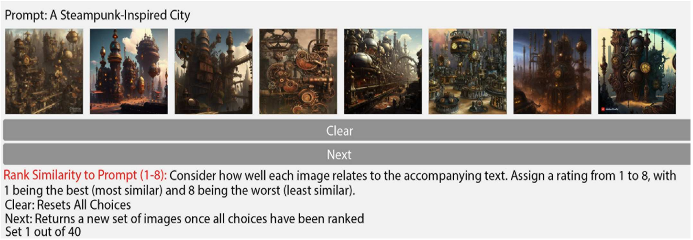
- Criterion 1: **Realism** - In the context of evaluating images, realism is the degree to which the content is authentic to real-world events. It measures how faithfully, without distortions or exaggeration, the visual representation captures identifiable and credible elements.
- Criterion 2: **Aesthetics** - An image's aesthetics refers to its artistic qualities and visual appeal. This criterion analyzes the visual elements' overall beauty, creativity, and composition, alongside their color equilibrium, uniformity, and the subjective artistic value that the human assessor feels.
- Criterion 3: **Visual Quality** - An image's visual quality includes its technical aspects, considering aspects like resolution, clarity, and overall image quality. It evaluates the image's sharpness, detail, and overall visual presentation, emphasizing the image's technical competence in terms of accuracy and clarity.

3.3 Participants

The study involved 20 participants consisting of staff and students affiliated with the computer science department. This selection aimed to ensure reliability and a comprehensive understanding of the ranking data's purpose. The ranking took 20 min for one participant to rank the images as per the given instructions. For each iteration of the ranking process (where a user ranked one prompt-image set), both the rankings and the order of image presentation were stored in a Firebase database [22].

3.4 User Interface for User Ranking

In the realm of different user inputs, diverse methodologies have been employed [23, 24], tailored to specific input characteristics. For the ranking collection for different criterion the ranking tool was designed. The ranking was administered online through a registered domain and the site were built using Fyne [25], and Firebase [22]. Upon loading, participants were prompted for their name only. A series of images were shown, randomized, with prompts and instructions. Users were asked to rank from best to worst by clicking each image which would become watermarked with the rank. Clear and submit buttons were presented below the series of images. To mitigate bias, per-mutations were employed to present images in a different order for a given prompt. Participants ranked the images based on the given criteria which changed depending on how far they were into the ranking.



Prompt: A Steampunk-Inspired City

Rank Similarity to Prompt (1-8): Consider how well each image relates to the accompanying text. Assign a rating from 1 to 8, with 1 being the best (most similar) and 8 being the worst (least similar).
Clear: Resets All Choices
Next: Returns a new set of images once all choices have been ranked
Set 1 out of 40

Fig. 5. The user interface of the human ranking tool.

The example of the tool shown in Fig. 5. The series of 8 images generated from different tools used for the project are there and below is the statement/ condition given on which the user needs to rank it. In the example shown the user needs to rank it based on similarity with the prompt that the user needs to see how much the images match the text provided to generate the image. The images disappear once it is clicked and selected. The rankings from the user are collected in the Firebase in the format shown in Table 2.

Table 2. Example of data gathered by human ranking tool, as it would appear in tabular form.

UID	Criterion	Prompt Type	Prompt	Criterion Rankings	Permutations
20	0	0	A mystical portal to another world.	[8, 7, 6, 5, 3, 4, 1, 2]	[1, 4, 3, 2, 6, 7, 5, 8]
20	0	0	A floating city in the sky.	[3, 2, 7, 8, 6, 5, 1, 4]	[2, 1, 5, 4, 3, 6, 7, 8]
20	1	0	A city on the back of a colossal turtle.	[6, 3, 2, 8, 1, 7, 4, 5]	[4, 1, 2, 8, 6, 7, 3, 5]

(*continued*)

Table 2. (*continued*)

UID	Criterion	Prompt Type	Prompt	Criterion Rankings	Permutations
20	1	0	A Viking longship on a stormy sea	[8, 5, 7, 2, 1, 3, 6, 4]	[1, 2, 4, 7, 8, 5, 3, 6]
20	2	1	An ancient Mayan temple hidden in the heart of a dense jungle	[6, 1, 7, 5, 4, 3, 2, 8]	[1, 7, 4, 3, 5, 2, 8, 6]
20	3	1	A haunted mansion with shifting corridors and ghostly apparitions	[7, 6, 5, 3, 4, 2, 1, 8]	[3, 5, 4, 7, 1, 8, 2, 6]

Checking the Average Values: The research analyzes the average scores given by participants for each image based on the four criteria. This analysis aims to identify which generative AI method consistently performs best in human rankings. As shown in Fig. 6, Adobe Firefly and Midjourney were particularly good, regularly scoring highly on all parameters in both categories. The user rankings are recorded given a scale where 1 represents the best choice and 8 corresponds to the least preferred option. Therefore, lower average scores indicate better performance as compared to the tools that attained high average scores.

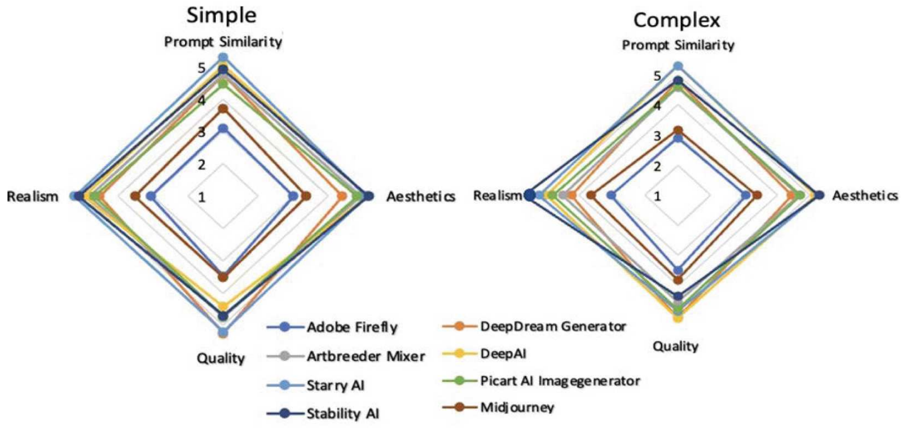


Fig. 6. Ranking from the user ranking on GAIA dataset (the lower the better).

The example of the tool shown in Fig. 5. The series of 8 images generated from different tools used for the project are there and below is the statement/condition given on which the user needs to rank it. In the example shown the user needs to rank it based on similarity with the prompt that the user needs to see how much the images match the text provided to generate the image. The images disappear once it is clicked and selected. The rankings from the user are collected in the Firebase in the format shown in Table 2.

Checking the Average Values: The research analyzes the average scores given by participants for each image based on the four criteria. This analysis aims to identify which generative AI method consistently performs best in human rankings. As shown in Fig. 6, Adobe Firefly and Midjourney were particularly good, regularly scoring highly on all parameters in both categories. The user rankings are recorded given a scale where 1 represents the best choice and 8 corresponds to the least preferred option. Therefore, lower average scores indicate better performance as compared to the tools that attained high average scores.

4 Dataset Analyses

This dataset is intended for use in evaluating image generation models and user preferences; however, considering that prompts, images, and user rankings are given, it is also possible that they can be used for other forms of machine learning applications such as inference.

For the purpose of this paper, we have decided to provide two analyses. For first analysis, we have used a deep learning approach with Convolutional Neural Networks (CNNs) for the feature extraction. The process is to integrate both visual and textual information. As shown in Fig. 7, we extract image features using VGG16 and while textual features are derived from prompts using Word2Vec [32].

Furthermore, we resize the image features to align with the dimensions of the prompt features. These resized features are then concatenated to form combined feature vectors, encapsulating both visual and semantic information. This integrated representation offers a richer feature space for subsequent analysis. These combined features, along with the target variable (average score), serve as input to regression models for prediction. In the model training and regression phase, the combined features are prepared for training regression models. Initially, the feature vectors are appended to the X list, with the singleton dimension removed using the squeeze() function. Simultaneously, the target variable, representing the average score, is appended to the Y list.

Following data preparation, the features are normalized using L_2 normalization to maintain consistency in their magnitudes. The Normalizer class from scikit-learn [26] is utilized for this purpose. Once the features are normalized, they are split into training and testing sets using the train test split with 20% of the data allocated for testing. This ensures the evaluation of model performance on unseen data.

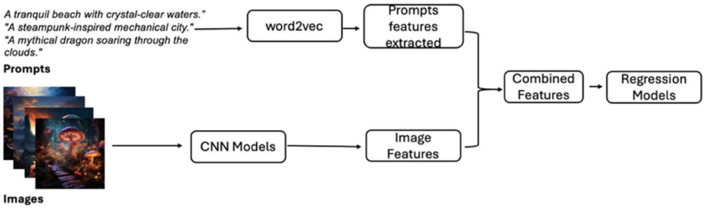

A tranquil beach with crystal-clear waters."
"A steampunk-inspired mechanical city."
"A mythical dragon soaring through the clouds."
Prompts
word2vec
Prompts features extracted
Images
CNN Models
Image Features
Combined Features
Regression Models

Fig. 7. The flowchart of training regression models from input features.

Table 3. Semantic Similarity: comparison of results generated by word2vec using original prompts and the output of the BLIP model. The highest scores are marked in **bold**.

Model	Simple	Complex	Both
Adobe Firefly	0.57	0.54	0.55
DeepDream Generator	0.57	0.57	0.57
Artbreeder Mixer	0.55	0.55	0.55
DeepAI	0.54	0.56	0.55
Starry AI	0.54	0.51	0.53
Picsart AI Image Generator	**0.60**	0.57	**0.58**
Stability AI, Clipdrop	0.58	**0.58**	**0.58**
Midjourney	0.56	0.54	0.55

The input for these models were the raw features extracted from the last dense layer of the VGG16 CNN [29] pre-trained on ImageNet dataset [30] as each image was forward propagated. The train/test split on GAIA was 80/20, respectively. We trained SVR regression model [27] using the sci-kit-learn [26] python library. The mean square error metric is shown in Table 3.

Specifically, support vector regression (SVR), trained and evaluated on the prepared datasets. The choice of regression models is due to efficiency in handling various data types, easy usage, and capturing underlying patterns. SVR works very well with high-dimensional feature spaces and complex patterns, making it well-suited for tasks where the relationships may not be easily linear.

Additionally, as we have categorized simple and complex prompts into different domains in the previous sections, therefore, to compare the prompts topic-wise, Support Vector Regression (SVR) with VGG16 feature extraction is to show mean square error (MSE) values. Our data, outlined in Table 4, indicates MSE metrics. The simple prompts excel in the History category but notably falter in the Human category, with a value of 3.112 in Similarity and 4.385 in Aesthetics. However, in complex prompts, the error increases for History, particularly with the Similarity criterion. For the Human category, the complex prompts exhibit higher MSE values in similarity, aesthetics, and quality criteria compared to simple prompts and drop significantly for complex prompts.

The second analysis is summarized in Table 3. We took our dataset of images and ran them through an image captioning deep learning model "BLIP" [31] and then calculated the semantic similarity between the output and the original prompt using word2vec [32]. We present the averages for simple, complex, and both for all models. The intention being to highlight how accurate the average generated image is to the original prompt. Note that the highest score for complex prompts is in terms of semantic similarity from the AI view (word2vec similarity). This shows the legitimate need of our dataset for actual human ranking on criterion such as realism, aesthetics and quality.

Table 4. Support Vector Regression Model Performance: Mean Squared Error (MSE) analysis of simple and complex prompts across categories

	Similarity	Aesthetics	Quality	Realism
Simple Prompts				
Nature	2.346	3.305	2.966	2.337
History	1.511	2.459	1.434	2.378
Human	3.112	4.385	2.368	1.952
Architecture	2.854	2.430	2.722	2.080
Complex Prompts				
Nature	2.789	2.517	2.036	1.621
History	3.168	2.481	1.935	3.109
Human	1.736	2.680	1.506	2.195
Architecture	2.660	3.138	1.554	2.414

5 Conclusion and Future Work

In this paper, we presented a novel dataset dubbed GAIA consisting of 800 images generated by 8 different models as well as the user rankings for the entire dataset. We conducted a comprehensive study on the newly collected GAIA dataset, including the evaluation and prediction of image models and user preferences. From this work, we learned the value of a variety of criteria when assessing generative AI methods. Most significantly, we showed our approach identified techniques such as Adobe Firefly and Midjourney as generally strong performers across a wide range of criteria ranked by humans, and we identified human rank patterns by using low average scores as clear evidence of better performance. We have explored two primary analyses to evaluate the performance of image generation models and user preferences based on both visual and textual data.

Future work should further investigate the interplay between prompt complexity and model performance, incorporating more diverse datasets and advanced models to refine our understanding of these dynamics. The expansion of the dataset could be considered in the future works. The expansion could involve increasing the number of prompts and can add more latest models to generate the synthesized images that can enrich the dataset. Text-to-image generation can be examined from an additional perspective by diversifying the models that were utilized to generate these images, allowing for an in-depth examination. Moreover, alternative to categorizing prompts as simple or complex we can implement a more nuanced complexity rating system. This could include assigning a numerical score or rating to each prompt based on various complexity dimensions, such as vocabulary difficulty, syntactic complexity, and conceptual intricacy.

Acknowledgments. This research was supported by the National Science Foundation (NSF) under Grant 2025234.

References

1. Ian, J., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. 2672–2680 (2014)
2. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 8748–8763 (2021)
3. Adobe firefly. https://www.adobe.com/products/firefly.html. Accessed July 2024
4. Deepdream generator. https://deepdreamgenerator.com/ddream/i0ark97dxvl. Accessed July 2024
5. Artbreeder mixer. https://www.artbreeder.com/create/mixer. Accessed July 2024
6. DeepAI: Text2Img Model. https://deepai.org/machine-learning-model/text2img. Accessed July 2024
7. Starry AI. https://starryai.com/app/. Accessed July 2024
8. Picsart AI image generator. https://picsart.com/ai-image-generator. Accessed July 2024
9. Stability AI, clipdrop. https://clipdrop.co/stable-diffusion,. Accessed July 2024
10. Midjourney. https://www.midjourney.com. Accessed July 2024
11. Flores Gallego, M.J., Perona, R., Puerta Callejon, J.M.: An application for aesthetic quality assessment in photography with interpretability features. Entropy **23**(11), 1389 (2021)
12. Adsumilli, B., Birkbeck, N., Bovik, A.C., Madhusudana, P.C., Wang, Y.: Image quality assessment using synthetic images. In: WACVW 2022 (2022)
13. Bovik, A.C., Sheikh, H.R.: Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
14. Preedanan, W., Kondo, T., Bunnun, P., Kumazawa, I.: A comparative study of image quality assessment. In: International Workshop on Advanced Image Technology, pp. 1–4 (2018)
15. Dumic, E., Grgic, S., Loncaric, M., Brzica, M., Tralic, D., Zaric, A.: Image quality assessment - comparison of objective measures with results of subjective test. In: Proceedings ELMAR-2010, pp. 113–118. Zadar, Croatia (2010)
16. Keimel, C., Hoßfeld, T., Korshunov, P., Mazza, F., Povoa, I., Redi, J.A.: Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal. In: Proceedings 2nd ACM International Workshop CrowdMM, pp. 29–34 (2013)
17. Oppenlaender, J.: The creativity of text-to-image generation. In: Proceedings of the 25th International Academic Mindtrek Conference (2022)
18. Deckers, N., et al.: The infinite index: Information retrieval on generative text-to-image models. In: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, pp. 172–186 (2023)
19. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: a feature similarity index for image quality assessment. IEEE Trans. Image Process. **20**(8), 2378–2386 (2011)
20. Raake, A., Ramachandra Rao, R.R., Merten, R., Goring, S.: Appeal and quality assessment for AI-generated images. In: International Conference Quality Multimedia Experience, pp. 115–118 (2023)
21. OpenAI, Chatgpt: Language model by openai. Accessed July 2024
22. Google Inc., Firebase Documentation. Accessed July 2024
23. Nguyen, T.V., Liu, S., Ni, B., Tan, J., Rui, Y., Yan, S.: Sense beauty via face, dressing, and/or voice. In: ACM Multimedia, pp. 239–248 (2012)
24. Nguyen, T.V., Liu, S., Ni, B., Tan, J., Rui, Y., Yan, S.: Towards decrypting attractiveness via multi-modality cues. ACM Trans. Multim. Comput. Commun. Appl. **9**(4), 28:1–28:20 (2013)
25. Williams, A., Fyne Contributors.: Fyne: Cross platform gui in go inspired by material design. Accessed July 2024
26. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

27. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification, 1396–1400 (2003)
28. Taunk, K., De, S., Verma, S., Swetapadma, A.: A brief review of nearest neighbor algorithm for learning and classification. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1255–1260 (2019)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
30. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
31. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. PMLR, pp. 12888–12900 (2022)
32. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013 (2013)
33. Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. IEEE Trans. Pattern Anal. Mach. Intell. **44**(3), 1552–1565 (2022)
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition, pp. 770–778 (2016)
35. Lévêque, L., et al.: Cuid: A new study of perceived image quality and its subjective assessment. In: Proceedings IEEE International Conference Image Processing (ICIP), pp. 116–120 (2020)
36. Nascimento, V., Florencio, D., Ribeiro, F.: Crowdsourcing subjective image quality evaluation. In: Proceedings of 18th IEEE International Conference on Image Processing, pp. 3097–3100 (2011)