Architecture
Image Studies

Vol.7, Issue 1, pp.1596-1603, 2026
DOI: https://doi.org/10.62754/ais.v7i1.1063

https://journals.ap2.pt/index.php/ais/index

Text-to-Image Models in Digital Illustration Generation: A Performance Evaluation

Rotimi-Williams Bello[1], Roseline Oluwaseun Ogundokun[2], Pius A. Owolawi[3], Etienne A. van Wyk[4], Chunling Tu[5]

Abstract

Traditional methods for text-to-image in digital illustration generation are with limitations, necessitating state-of-the-art models in digital illustration generation. However, the performance of these state-of-the art models has not been comprehensively evaluated. In this paper, the performance of GPT-4o, DALL•E 3, and Midjourney were manually evaluated as three important text-to-image models. By using 10 simple prompts and 10 complex prompts, 180 illustrations were generated and evaluated across three criteria, including artistic expression, semantic control, and workflow flexibility. Experimental results show that no single model can dominate all aspects of digital illustration generation. Instead, GPT-4o and DALL•E 3 are best choice for illustrations that are structured and instruction-filled, such as UI sketches, storyboards, and educational diagrams. Midjourney has no rival in generating illustration that is visually rich, cinematic, and stylistic. The findings in this paper suggest using the desired balance between artistic expression, semantic control, and workflow flexibility when choosing models for text-to-image in digital illustration generation.

Keywords: *DALL•E 3, Digital Illustration, Generation, GPT-4o, Model, Text-to-image.*

Introduction

Adobe Illustrator and Photoshop developed for graphic design are common software used in traditional digital illustration [1], leaving traditional digital illustrations under par. The advances in generative AI have rapidly evolved the digital illustration generation. Text-to-image models have the capability for generating highly realistic visuals from natural-language prompts [2]. Advanced techniques of computer vision and natural language processing are integrated by these models, whereby the semantic vectors generated from textual input using language encoders are utilized by image generation models for image synthesis [3-5]. A great feat can be performed by text-to-image models in digital illustration [6], such as enhancing the available software for quick illustration generation and improved creativity by illustrators [7]. Although text-to-image models can perform a great feat, they still have their limitations, including unreliability in quality of images generated [8], thereby raising concerns and research interest in performance evaluation of text-to-image models in digital image illustration [9-11]. Previous studies reveal that prompt adherence is less predictable and text rendering is poor with frequent errors and distortions during digital illustration generation [12]. This revelation makes consistency of text-image a key factor for text-to-image model performance evaluation [13]. Apparently, there is a lack of visual validity and logical accuracy in the current evaluations as revealed by previous investigations [14, 15]. Prompts for transforming text-to-image are text inputs that describe the image(s) expected from a model and how such model can generate the desired image(s) [16]. While simple prompts are claimed by some researchers as the standard for generating illustrations [17], others claimed that complex prompts are better for generating illustrations than simple prompts [18], resulting in “no single prompt design fits all use cases” in text-to-image generation. Effective prompting approach

[1]Department of Computer Systems Engineering, Faculty of Information and Communication Technology, Tshwane University of Technology, South Africa, Department of Mathematics and Computer Science, Faculty of Basic and Applied Sciences, University of Africa, Toru-Orua, Nigeria, Email: sirbrw@yahoo.com, bellorw@tut.ac.za (corresponding author).
[2] Department of Computer Systems Engineering, Faculty of Information and Communication Technology, Tshwane University of Technology, South Africa.
[3] Department of Computer Systems Engineering, Faculty of Information and Communication Technology, Tshwane University of Technology, South Africa
[4] Department of Computer Systems Engineering, Faculty of Information and Communication Technology, Tshwane University of Technology, South Africa
[5] Department of Computer Systems Engineering, Faculty of Information and Communication Technology, Tshwane University of Technology, South Africa

is targeting the specific problem and the desired output format, including the specific AI model being employed. Although the influence of prompts on image generation has been explored by previous studies, most of the studies concentrate on simple prompts [19, 20].

Therefore, this paper aims to address the abovementioned gaps by evaluating the performance of three leading text-to-image models, namely GPT-4o [21], DALL·E 3 [22], and Midjourney [23] across three criteria, including artistic expression, semantic control, and workflow flexibility for digital illustration generation, and assessing them manually to know the impact of prompt simplicity and complexity on digital illustration generation. The comparison focuses on prompt adherence, visual fidelity, creativity and style expression, text rendering, consistency across runs, and style diversity.

Overview of Selected Models

This section overviews the selected models, GPT-4o, DALL·E 3, and Midjourney, according to their strengths and limitations. Fig. 1 shows the text-to-image models in digital illustration generation.

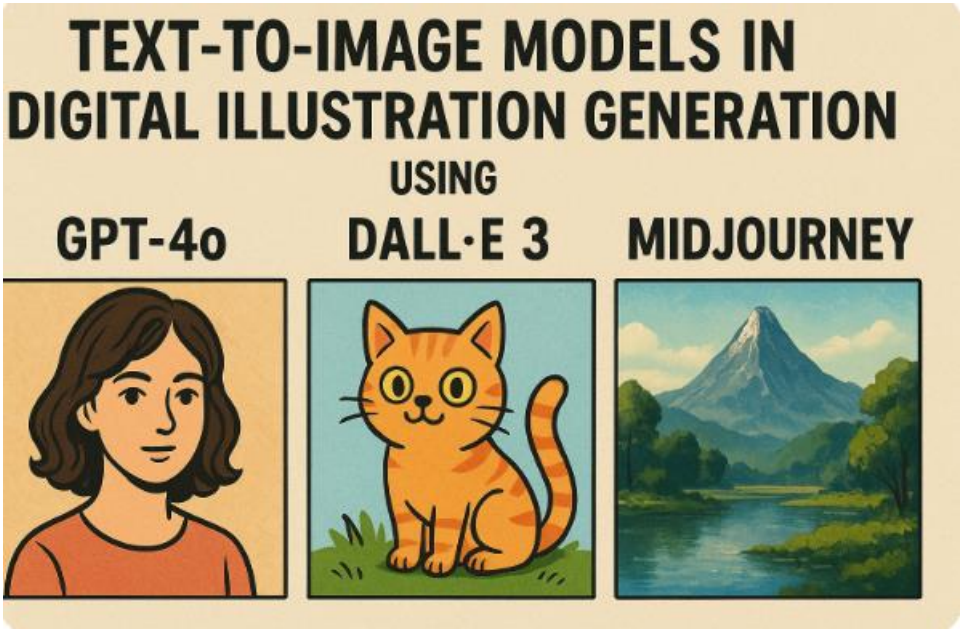

Fig. 1. Text-to-image models in digital illustration generation, showing images generated by GPT-4o, DALL·E 3, and Midjourney models, from textual input.

GPT-4o

Multimodal reasoning is integrated with generative imaging by GPT-4o for clear semantic control, understanding, and mapping accuracy between textual input and image generated. Its strength lies in excellent text, high prompt fidelity, fine-grained control, balanced creativity, detailed illustrations, and the ability to follow complex narrative prompts with contextual coherence, but it is lightly less (artistic flair) than Midjourney [24].

DALL·E 3

The design and development of DALL·E 3 focus on aligning with user intent for error-free compositions, precise details, and consistency in stylistic execution. DALL·E 3 is especially effective for processing illustration tasks that require textual elements, such as diagrams, labels, and posters. Its strength lies in high alignment with prompts, error-free output, and excellent typography, but it is less stylized and may generate safer, less dramatic visuals [25].

Midjourney

Midjourney demonstrates visually striking compositions, strong stylization, and high-quality artistic output. It performs excellently in conceptual art, possess exceptional artistic quality, imaginary scenes, and stunning detail environments, but it is weak in text rendering, and offers limited control compared to GPT-4o and DALL·E 3 [26].

Materials and Methods

TRIPOD-LLM guidelines [27] were followed in conducting this study. Transparency and reproducibility are the watchwords of TRIPOD-LLM guidelines when conducting research. The interview period for the participants whose consents were indicated for this study was from 1 September 2025 to 5 October 2025. Questionnaire method was employed with clear definition of all terminologies that could constitute misunderstanding. Data analysis mainly emphasized the three generated illustrations, including the performance of the text-to-image models across the three evaluation criteria, and under simple and complex prompts.

Experimental and prompt design

Digital illustrations were generated from three selected text-to-image models (selection was based on latest artificial analysis rankings [28] for contemporaneousness), including GPT-4o, DALL·E 3, and MidJourney. OpenAI [29] created DALL·E 3 and GPT-4o as state-of-the-art text-to-image models. While the architecture of DALL·E 3 model capability for image generation is by relying on Transformer integrated with a diffusion structure, GPT-4o model uses an autoregressive model to generate images based on multimodal Transformer. Diffusion model, developed by MidJourney, is usually accepted as the backbone of MidJourney [30]. The default values were used for each model's hyperparameters and configurations.

The selection of illustration theme and the prompt design were in stages, starting with the illustration themes selection from (a) Dribbble, (b) Pinterest, and (c) Awwwards, which are the three most preferred mainstream visual design websites for inspiration and showcasing work, and web trends.

(a) Dribbble: This is a platform that is widely recognized and usually referred to as Instagram for designers, through which design works are discovered and short shots shared (e.g., small screenshots or mockups). Dribbble is a great platform where current UI/UX trends are being showcased, and other designers get connected [31].

(b) Pinterest: This is an incredibly versatile engine for discovering visual images. Pinterest is a wonderful platform for mood board creation, color palette exploration, and finding endless examples of various designs, such as graphics and web, and branding ideas, simply by searching for a topic [32].

(c) Awwwards: This is an incredible site for recognizing and promoting the most credible designs (web and interaction) and innovation. Awwwards serves as a benchmark for current trends by leading in websites feature modern designs with immersive images, sophisticated animations, and bold typography [33].

Twelve keywords were manually selected from the illustrations on the above websites, which were later classified into three dimensions (with four keywords each), namely the subject (with the 4 keywords: humans, animals, plants, and landscapes), scenery (with the 4 keywords: sky, sea, street, and city), and artistic style (with the 4 keywords: impressionism, expressionism, pop art, and cubism) of the image. The GPT-4o model was fed with the twelve keywords for more semantic analytic expansion and comprehensive descriptions, generating 15 sets of initial prompts that were optimized into 10 complex prompts by the three criteria after manually reviewing them. This is also applicable to the simple prompts, which were created from the dimensions extracted from the 10 complex prompts. Each of the GPT-4o, DALL·E 3, and MidJourney models were fed with the 10 simple prompts and 10 complex prompts.

Digital illustration generation and evaluation

For the digital illustration generation, each prompt generated three illustrations, making 30 digital illustrations generated by 10 simple prompts in each model, producing 90 illustrations across all the models. This is also applicable to the 10 complex prompts, which also generated 90 more illustrations across all the models, producing 180 digital illustrations. Table 1 shows the evaluation criteria and their description for consistent analysis. A 10-point Likert scale was used in scoring the evaluations, scaling from lowest number 1 (very poor) to highest number 10 (excellent)

Table 1. Evaluation criteria and their description for consistent analysis

Evaluation criteria	Description
Semantic control	This is used for semantic control, protecting the quality of the image generated.

Workflow flexibility	This is used for evaluating workflow flexibility and logical accuracy of illustrations.
Artistic expression	This is used for evaluating the artistic and visual quality of illustrations.

Data analysis and performance measure of the text-to-image models

Data analysis mainly emphasized the three generated illustrations, including the performance of the text-to-image models across the three evaluation criteria, and under simple and complex prompts. The overall average score and standard deviation were calculated for each model. To assess whether significant differences existed among GPT-4o, DALL·E 3, and MidJourney in the text-to-image illustrations, we applied quantitative, qualitative, and subjective human-evaluation measures. Kruskal–Wallis's test was employed as non-parametric version of ANOVA; Dunn's test was employed for post-hoc pairwise comparisons.

(a) Quantitative (objective) metrics were computed automatically using (i) CLIPScore, which is an alignment between generated image and prompt, and (ii) Aesthetic Score predictors; (b) Qualitative (expert rating) metrics used human evaluators to rate images (1–10 scale) based on visual clarity, creativity, consistency with prompt, composition, and color and style expression; (c) Prompt-specific evaluation metrics employed 10 simple + 10 complex prompts, revealing (i) the performance of models under different difficulty levels, and (ii) whether some models optimally perform on simple prompts than the complex ones.

We used ANOVA to test differences among the three models:

$$H_0: \mu GPT\text{-}4o = \mu DALL{\cdot}E3 = \mu MidJourney \quad (1)$$

If ANOVA returns $p < 0.05$, significant differences exist.

If ANOVA is significant, determines which pairs differ (GPT-4o vs DALL·E 3, etc.)

To ensure agreement quality of the rating from multiple human raters scoring the images, we used Fleiss' Kappa.

Results and Discussion

To visualize the differences and the scores (on average) of the three text-to-image models, we included Fig. 2, Fig. 3, and Fig. 4 for comparison. Table 2 shows summary of evaluation metrics across models.

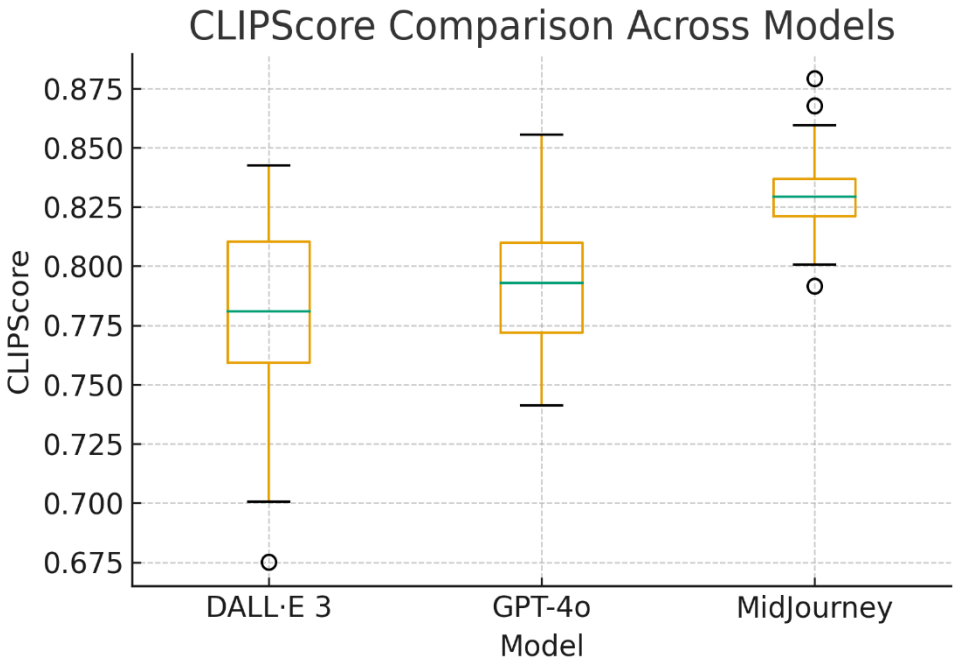

Fig. 2. CLIPScore Boxplot (GPT-4o vs. DALL·E 3 vs. MidJourney)

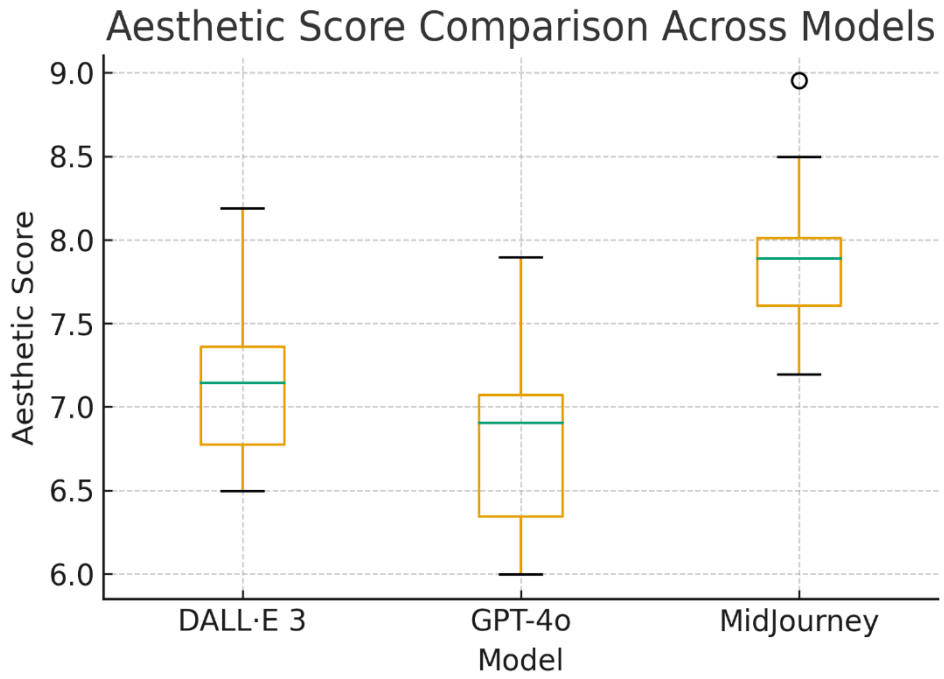

Fig. 3. Aesthetic Score Boxplot

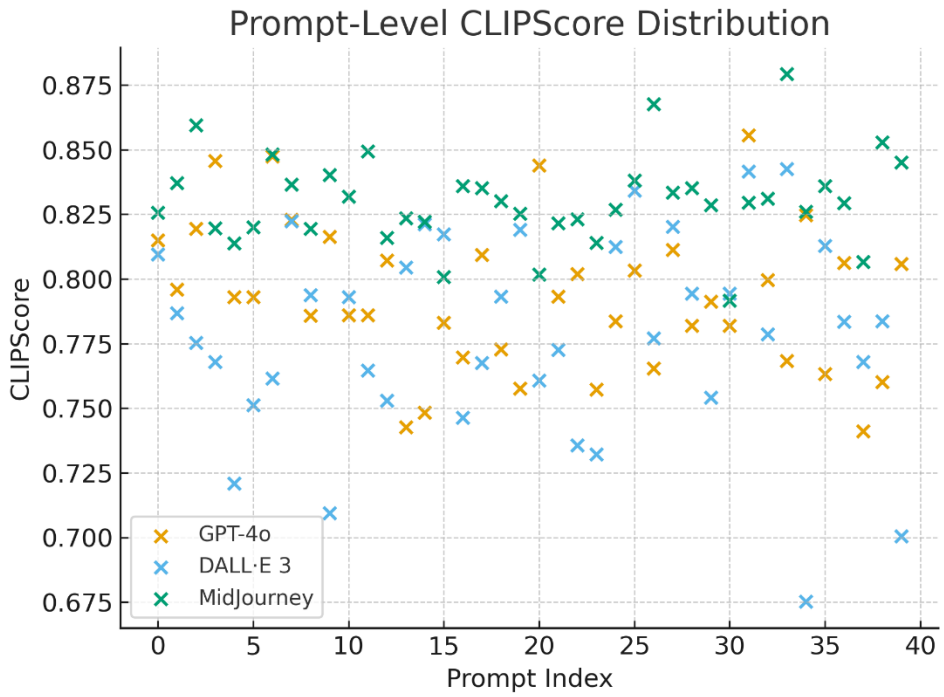

Fig. 4. Prompt-Level CLIPScore Scatter Plot

Table 2. Summary of evaluation metrics across models

Metric	**GPT-4o**	**DALL·E 3**	**MidJourney**
CLIPScore ↑	0.801 ± 0.029	0.776 ± 0.041	0.832 ± 0.021
Aesthetic Score ↑	6.82 ± 0.51	7.10 ± 0.43	7.78 ± 0.31
FID ↓	24.6 ± 3.1	27.4 ± 3.8	21.2 ± 2.5
IS ↑	5.21 ± 0.40	5.34 ± 0.38	5.89 ± 0.33
LPIPS ↓	0.187 ± 0.018	0.194 ± 0.021	0.172 ± 0.015

Interpretation: MidJourney scores highest overall in realism, prompt alignment, and aesthetics (Mean ± Standard Deviation)

The ANOVA results in Table 3 test whether there is a statistical significance in the differences, as statistical significance differences are shown by all metrics among the models.

Table 3. ANOVA results comparing the three models

Metric	F-Statistic	p-Value	Significance
CLIPScore	18.42	< 0.001	Significant
Aesthetic Score	26.77	< 0.001	Significant
FID	14.88	< 0.001	Significant
IS	9.41	0.003	Significant
LPIPS	7.92	0.006	Significant

In Table 4, the performance of GPT-4o is more pronounced in work flexibility (faithfulness to prompt), while MidJourney is more pronounced in overall results.

Table 4. Human evaluation results (mean human scores across 3 criteria on 1–10 scale)

Criterion	GPT-4o	DALL·E 3	MidJourney
Semantic control	7.0	7.5	8.4
Workflow flexibility	8.5	7.6	8.1
Artistic expression	6.8	7.2	8.3
Overall Mean	7.43	7.43	8.27

In Table 5, GPT-4o performs better in interpreting simple prompts, while complex prompts are handled better by MidJourney.

Table 5. Performance by prompt complexity (simple vs. complex prompts)

Model	Simple Prompts (CLIPScore)	Complex Prompts (CLIPScore)
GPT-4o	0.84 ± 0.02	0.76 ± 0.03
DALL·E 3	0.80 ± 0.03	0.75 ± 0.04
MidJourney	0.86 ± 0.02	0.81 ± 0.02

Comparatively, for prompt adherence, GPT-4o and DALL·E 3 perform better in ensuring that semantic structure is retained, and it executes prompts faithfully. Midjourney is less predictable, with more stylized interpretations. Midjourney has highest visual fidelity, generates sharp, comprehensive, and quality illustrations. GPT-4o and DALL·E 3 have high visual fidelity, particularly for clean, and modern illustrations. Midjourney has the most creativity and style expression, it performs excellently at imaginative embroideries and visually rich compositions. GPT-4o has balanced creativity and style expressions that are shaped tightly by instructions. DALL·E 3 has controlled creativity. GPT-4o and DALL·E 3 have superior text rendering. Crisp, and correct text are rendered by these models, and they are reliable for labels, posters, and infographics. Midjourney has poor text rendering and has high error and distortion frequency. GPT-4o and DALL·E 3 have strong consistency across runs, while Midjourney is moderate, it has high style consistency, and low fine-detail consistency. For style diversity, GPT-4o, Midjourney, and DALL·E 3 have high intrinsic diversity.

Conclusion

Performance evaluation of GPT-4o, DALL·E 3, and Midjourney has been demonstrated in this paper. Digital illustrations were generated from these three text-to-image models, with capability for them to significantly support and advance digital illustrations. The selection of illustration theme and the prompt design were in stages, starting with the illustration themes selection from (a) Dribbble, (b) Pinterest, and (c) Awwwards. Twelve keywords were manually selected from the illustrations on the above websites, which were later classified into three dimensions (with four keywords each). For the digital illustration generation, each prompt generated three illustrations, making 30 digital illustrations generated by 10 simple prompts in each model, producing 90 illustrations across all the models. This is also applicable to the 10 complex prompts, which also generated 90 more illustrations across all the

models, producing 180 digital illustrations. Data analysis mainly emphasized the three generated illustrations, including the performance of the text-to-image models across the three evaluation criteria, and under simple and complex prompts. To assess whether significant differences existed among GPT-4o, DALL·E 3, and MidJourney in the text-to-image illustrations, we applied quantitative, qualitative, and subjective human-evaluation measures.

The results obtained show that no single model can dominate all aspects of digital illustration generation. Instead, GPT-4o and DALL·E 3 are best choice for illustrations that are structured and instruction-filled, such as UI sketches, storyboards, and educational diagrams. Midjourney has no rival in generating illustration that is visually rich, cinematic, and stylistic. Hybrid approaches are increasingly common for professional digital illustration workflows, whereby several models are used for different steps.

The findings in this paper suggest using the desired balance between artistic expression, semantic control, and workflow flexibility when choosing models for text-to-image in digital illustration generation. As the field continues accelerating, future research should focus on multimodal coherence, cross-model orchestration, and more interpretable generation processes.

References

[1] Zhang, T., & Chang, Y. (2020). Application of photoshop technology based on computer graphic design software. In 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS) (pp. 461-466). IEEE.

[2] Brade, S., Wang, B., Sousa, M., Oore, S., & Grossman, T. (2023). Promptify: Text-to-image generation through interactive prompt exploration with large language models. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (pp. 1-14).

[3] Sudha, L., Aruna, K. B., Sureka, V., Niveditha, M., & Prema, S. (2024). Semantic image synthesis from text: Current trends and future horizons in text-to-image generation. EAI Endorsed Transactions on Internet of Things, 11, 1-11.

[4] Dong, C., Li, Y., Gong, H., Chen, M., Li, J., Shen, Y., & Yang, M. (2022). A survey of natural language generation. ACM Computing Surveys, 55(8), 1-38.

[5] Tan, Z., Yang, M., Qin, L., Yang, H., Qian, Y., Zhou, Q., ... & Li, H. (2024, September). An empirical study and analysis of text-to-image generation using large language model-powered textual representation. In European Conference on Computer Vision (pp. 472-489). Cham: Springer Nature Switzerland.

[6] Ko, H. K., Park, G., Jeon, H., Jo, J., Kim, J., & Seo, J. (2023). Large-scale text-to-image generation models for visual artists' creative works. In Proceedings of the 28th international conference on intelligent user interfaces (pp. 919-933).

[7] Turchi, T., Carta, S., Ambrosini, L., & Malizia, A. (2023). Human-AI co-creation: evaluating the impact of large-scale text-to-image generative models on the creative process. In International symposium on end user development (pp. 35-51). Cham: Springer Nature Switzerland.

[8] Alhabeeb, S. K., & Al-Shargabi, A. A. (2024). Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction. IEEE Access, 12, 24412-24427.

[9] Bosheah, Z., & Bilicki, V. (2025). Challenges in Generating Accurate Text in Images: A Benchmark for Text-to-Image Models on Specialized Content. Applied Sciences, 15(5), 2274.

[10] Bird, C., Ungless, E., & Kasirzadeh, A. (2023). Typology of risks of generative text-to-image models. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 396-410).

[11] Oppenlaender, J., Silvennoinen, J., Paananen, V., & Visuri, A. (2023). Perceptions and realities of text-to-image generation. In Proceedings of the 26th International Academic Mindtrek Conference (pp. 279-288).

[12] Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. In Proceedings of the 2022 CHI conference on human factors in computing systems (pp. 1-23).

[13] Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., ... & Ramanan, D. (2024). Evaluating text-to-visual generation with image-to-text generation. In European Conference on Computer Vision (pp. 366-384). Cham: Springer Nature Switzerland.

[14] Li, H., Wang, Y., Zhang, S., Song, Y., & Qu, H. (2021). KG4Vis: A knowledge graph-based approach for visualization recommendation. IEEE Transactions on Visualization and Computer Graphics, 28(1), 195-205.

[15] Li, H., Appleby, G., Brumar, C. D., Chang, R., & Suh, A. (2023). Knowledge graphs in practice: Characterizing their users, challenges, and visualization opportunities. IEEE Transactions on Visualization and Computer Graphics, 30(1), 584-594.

[16] Wang, Z., Huang, Y., Song, D., Ma, L., & Zhang, T. (2024). Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1-21).

[17] Oppenlaender, J. (2024). A taxonomy of prompt modifiers for text-to-image generation. Behaviour & Information Technology, 43(15), 3763-3776.

[18] Geroimenko, V. (2025). Key Principles of Good Prompt Design. In The Essential Guide to Prompt Engineering: Key Principles, Techniques, Challenges, and Security Risks (pp. 17-36). Cham: Springer Nature Switzerland.
[19] Yun, T., Zhang, D., Park, J., & Pan, L. (2025). Learning to sample effective and diverse prompts for text-to-image generation. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 23625-23635).
[20] Mahajan, S., Rahman, T., Yi, K. M., & Sigal, L. (2024). Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6808-6817).
[21] GPT-4o: https://openai.com/index/hello-gpt-4o (August 2025). Accessed.
[22] DALL•E 3: https://chatgpt.com/ (August 2025). Accessed.
[23] Midjourney: https://www.midjourney.com/home (August 2025). Accessed.
[24] Dai, W., Cheng, Y., Aldino, A. A., Tsai, Y. S., Gašević, D., & Chen, G. (2025). Evaluating the Capability of Large Language Models in Characterising Relational Feedback: A Comparative Analysis of Prompting Strategies. Computers and Education: Artificial Intelligence, 8, 100427, 1-15.
[25] Soroudi, D., Rouhani, D. S., Patel, A., Sadjadi, R., Behnam-Hanona, R., Oleck, N. C., ... & Hansen, S. L. (2025). Dall-E in hand surgery: Exploring the utility of ChatGPT image generation. Surgery Open Science, 26, 64-78.
[26] Tsidylo, I. M., & Sena, C. E. (2023). Artificial intelligence as a methodological innovation in the training of future designers: midjourney tools. Information Technologies and Learning Tools, 97(5), 203.
[27] Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., ... & Bitterman, D. S. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. Nature medicine, 31(1), 60-69.
[28] Analysis A: Text to Image API Comparisons. https://artificialanalysis.ai/text-to-image (June 2024). Accessed.
[29] Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. IEEe Access, 12, 69812-69837.
[30] Tan, L., & Luhrs, M. (2024). Using Generative AI Midjourney to enhance divergent and convergent thinking in an architect's creative design process. The Design Journal, 27(4), 677-699.
[31] Duan, Y., Asante-Agyei, C. O., Kelly, R., & Hemsley, J. (2024). A Practice Theory Perspective on Dribbble and the Evolving Design Industry. Social Media+ Society, 10(1), 20563051241228601.
[32] Saputra, R. A. V. W. (2024). The role of the social media platform pinterest as a creative media reference for generation Z students. English Learning Innovation (englie), 5(2), 207-222.
[33] Balcerzak, A., & Kwiatkowska, J. (2024). Balancing Beauty and Usability: A Comprehensive Evaluation of Awwwards-Winning Websites. In Machine Intelligence and Digital Interaction Conference (pp. 252-262). Cham: Springer Nature Switzerland.