Advanced Sciences and Technology Journal

ASTJ vol. 2, no. 2 (2025), P 1018
10.21608/astj.2025.343418.1018
https://astj.journals.ekb.eg/

Advancing Creativity: A Comprehensive Review of AI-Driven Text-to-Image Generation and Its Applications

Noha Hussen*, Ahmed Samir, Aliaa Adel, Abdelrahman Gaber, Mommen Attaia, Ahmed Mohamed

Software Engineering Department, Faculty of Engineering and Technology, Egyptian Chinese University, Cairo, Egypt
**noha.hussen@ecu.edu.eg*

ARTICLEINFO

Article history:
Received 10 December 2024
Revised 26 December 2024
Accepted 9 January 2025
Available online 2 April 2025

Handling Editor:
Prof. Dr. Mohamed Talaat Moustafa

Keywords:
Book Illustration
Context Transition
Generative AI
Narrative Coherence
Prompt Engineering
Text-to-Image Synthesis
Text-to-Image Illustration
User Experience

ABSTRACT

The field of AI-driven text-to-image generation has emerged as a transformative intersection of technology and creativity, enabling the automatic synthesis of visuals from textual descriptions. This capability has profound implications for diverse applications, from storytelling and education to digital art and design. By automating the translation of textual content into visually rich representations, text-to-image generation bridges the gap between linguistic and visual modalities, fostering novel opportunities for innovation and exploration. This review explores the state-of-the-art advancements in text-to-image synthesis, emphasizing the technological evolution from Generative Adversarial Networks (GANs) to diffusion models and transformer-based architectures. It highlights how these models, including tools like DALL-E-2, Midjourney, and Stable Diffusion, have advanced in generating semantically aligned, visually coherent, and aesthetically appealing images. Despite notable progress, significant challenges remain. These include maintaining contextual coherence across sequences, adhering to artistic and compositional principles, and addressing the dependency on detailed textual prompts. Moreover, the limitations of existing evaluation metrics, such as the Inception Score (IS) and Fréchet Inception Distance (FID), are critically analyzed, underscoring the need for metrics that account for semantic fidelity, emotional resonance, and user-centric perspectives. The review synthesizes insights from recent studies to identify key areas for innovation, such as enhanced context management, integration of 3D modeling capabilities, and real-time user interaction mechanisms. Finally, the paper outlines future directions to address current limitations, promote interdisciplinary collaboration, and establish ethical guidelines for responsible AI deployment. By doing so, this work aims to provide a comprehensive foundation for advancing generative AI and its applications across creative industries

1. Introduction

Mark Twain's observation, "The secret of getting ahead is getting started," resonates strongly with the transformative power of Artificial Intelligence (AI) in visual content creation. In recent years, AI has emerged as a game-changing tool capable of converting textual descriptions into vivid, contextually aligned visuals via text-to-image synthesis. AI has revolutionized creative and professional industries by automating the generation of high-quality, dynamic visuals using technologies such as Generative Adversarial Networks (GANs) and transformer-based models[1, 2].

Traditionally, visual content creation relied on skilled artists, which, while effective, was often time-consuming and costly. AI-powered technologies have disrupted this paradigm, enabling visuals to be generated directly from textual descriptions[3]. This approach has significant advantages for various industries by making visual content creation more accessible, engaging, and scalable. Text-to-image synthesis, which aligns visuals closely with contextual input, opens up new possibilities for generating visuals efficiently, reducing reliance on manual artistry.

The use of AI in visual content creation, namely text-to-image synthesis, is the main topic of this review. It looks at the technological developments that support this field, such as transformer-based models, attention mechanisms, and GANs. It also examines challenges like ensuring coherence, maintaining stylistic consistency, and converting complex textual inputs with multiple elements into cohesive visual outputs. The analysis further identifies shortcomings in current evaluation metrics and proposes new methods for assessing the effectiveness of AI-generated visuals.

The review is divided into multiple sections in order to guarantee depth and clarity. An outline of AI's revolutionary potential in text-to-image generation and its applicability in creative sectors is given in the Introduction. The section on the Foundations of AI-Driven Text-to-Image Generation highlights the major technologies that make text-to-image synthesis possible while charting the evolution of generative AI models, including GANs, diffusion models, and transformer architectures. The State-of-the-Art Techniques section explores state-of-the-art developments, with particular emphasis on models such as DALL-E, MidJourney, and Stable Diffusion, as well as advancements in assessment frameworks and attention mechanisms for improving semantic alignment and image quality.

The Literature Review synthesizes existing research, exploring advancements, practical applications, and studies on the realism, aesthetic appeal, and contextual accuracy of AI-generated images. Key Findings emphasizes improvements in AI capabilities, challenges in semantic alignment, and the influence of prompt engineering. The Limitations section discusses ongoing issues, such as contextual inconsistencies, reliance on textual prompts, inadequate evaluation metrics, and scalability challenges for long-form narratives. Future Directions outlines opportunities for innovation, including enhanced context management, integration of 3D modeling, improved evaluation metrics, and real-time user interaction mechanisms. Finally, the Conclusion summarizes key insights, highlights research gaps, and reiterates the transformative potential of AI-driven text-to-image synthesis in reshaping creativity and innovation.

2. Foundations of AI-Driven Text-to-Image Generation

The quick growth of text-to-image generating technologies is supported by the creation of generative AI models. GANs, which use a dual-model system—a generator producing images and a discriminator assessing them—to produce realistic graphics, were at the forefront of early developments. More accurate text-to-image alignment was made possible by the introduction of multi-stage architectures and attention methods by StackGAN and AttnGAN[4].

A strong substitute was provided by diffusion models, such as Stable Diffusion, which use probabilistic techniques to repeatedly de-noise data and provide high-fidelity image synthesis. Transformer-based architecture has further revolutionized the area by effectively bridging textual and visual data through the use of large datasets and multimodal embeddings, as demonstrated by models such as DALL-E-2. In order to ensure coherence and contextual relevance, these models incorporate frameworks like CLIP (Contrastive Language–visual Pre-training), which aligns textual and visual representations [5].

The launch of ChatGPT on November 30, 2022[6] ,catalyzed an unprecedented rise in the public's recognition and adoption of Generative Artificial Intelligence (GAI). This milestone traces its roots back to the seminal 1956 Dartmouth College summer project, led by McCarthy, which marked the birth of AI[7]. The project's primary objective was to develop machines capable of performing tasks traditionally requiring human intelligence[8-11], including fields such as computer vision, natural language processing (NLP)[12],robotics, and more. Since then, substantial progress has been made in enabling machines to mimic human capabilities such as communication, locomotion, reasoning, and decision-making[13].

Prior to 2014, existing deep learning models were predominantly descriptive, focusing on identifying, summarizing, and interpreting data patterns and relationships. These models were geared toward understanding data trends and making predictions based on available information. However, in 2014, Goodfellow et al. authers in[14] revolutionized the field with the introduction of the GAN, which marked a pivotal shift toward realizing GAI. Unlike descriptive models, generative models such as GANs are designed to learn the data's underlying probability distribution, allowing them to produce new data samples that closely mimic the patterns found in training datasets [15-17].

The advent of GANs represented a significant evolution from conventional deep learning methods, unlocking vast potential for Generative Artificial Intelligence. GAI has since gained widespread acclaim for its transformative capabilities across diverse fields. It offers innovative solutions to complex challenges[18], enabling the generation of synthetic data, artistic creations, and highly realistic simulations.

The rapid growth of GAI has spurred extensive research and inquiry, emphasizing the importance of comprehensive exploration into this transformative technology. Despite many recent studies addressing the proliferation of GAI[15], which delve into challenges, applications, and models, there remains a notable gap in examining the theoretical and mathematical underpinnings of contemporary GAI models. This includes analyzing tools evolving at an exponential pace, as schematically illustrated in Figure 1.

2.1. Key Technologies and Algorithms

1. Generative Adversarial Networks (GANs): The backbone of early text-to-image systems, GANs generate visually realistic outputs while improving semantic alignment with techniques like stacked architectures and attention-based refinements.
2. Diffusion Models: These models iteratively refine outputs by reducing noise, producing photorealistic and stylistically diverse images. Stable Diffusion, for instance, exemplifies the model's ability to produce detailed visuals across a range of prompts[19].
3. Transformer Architectures: Models like DALL-E and GPT integrate attention mechanisms to focus on specific textual elements, enhancing image quality and alignment[20].
4. Prompt Engineering: Effective text-to-image systems rely on well-crafted textual prompts. Innovations in prompt engineering enable nuanced and contextually appropriate image generation, particularly for narrative-driven applications[21].

2.2. Significance and Impact

The fundamental technologies that enable text-to-image generation redefine creativity and accessibility. They enable a wide range of applications by automating the process of converting text into graphics, such as storytelling, education, and digital art. The seamless integration of AI systems into these fields improves workflows, lowers manual effort, and increases access to high-quality visual content[22].

3. State-of-the-Art Techniques

DALL-E and DALL-E-2: These models demonstrate advanced capabilities in generating images with semantic fidelity and high resolution. DALL-E-2, specifically, utilizes improved attention layers and CLIP-based alignment, excelling in realistic and diverse outputs[23].

Midjourney: Popular for its aesthetic flexibility, Midjourney allows users to create visually striking images. Its iterative feedback mechanism ensures continuous refinement of generated visuals, making it a preferred tool in creative industries[24].

Stable Diffusion: As an open-source platform, Stable Diffusion provides unparalleled adaptability. Its ability to process intricate prompts and deliver stylistic coherence makes it invaluable for professional and artistic use.

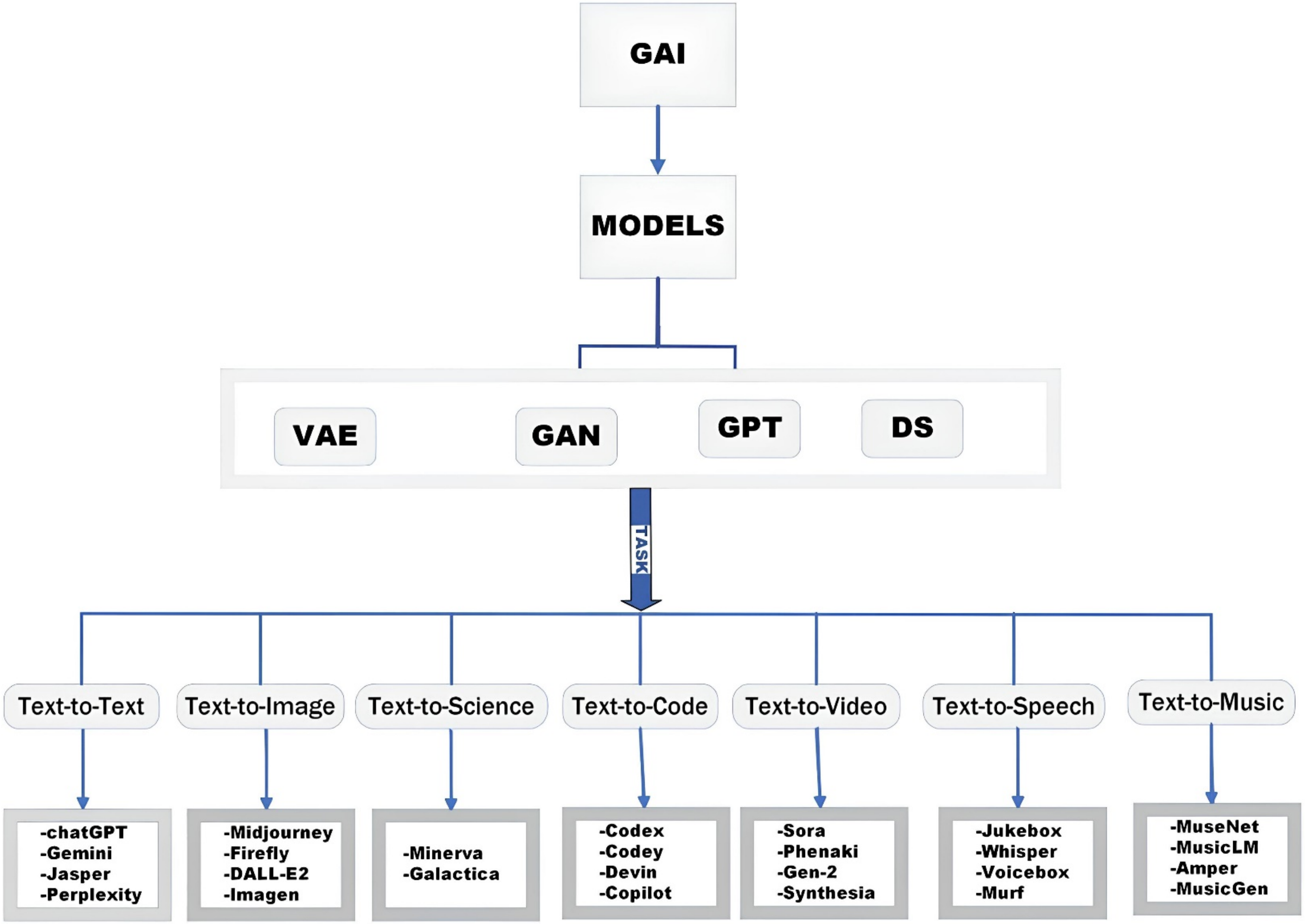


Fig 1. GAI Models & Use Cases.

3.1. Technological Innovations

- Attention Mechanisms: By focusing on critical textual components, attention mechanisms improve the contextual accuracy and visual quality of generated images.
- Stacked Architectures: Multi-layered approaches, as seen in Attn GAN, enable the gradual refinement of images, resulting in more detailed and contextually aligned outputs.
- Integration of Evaluation Frameworks: Metrics such as Inception Score (IS) and Fréchet Inception Distance (FID) are employed to assess visual appeal, while subjective evaluations capture user-centric aspects like realism and creativity.

3.2. Challenges Addressed by State-of-the-Art Techniques

- Semantic Coherence: Advanced alignment techniques address inconsistencies between text and generated images.
- Diversity in Outputs: Innovations such as diffusion models enhance the range and quality of visual outputs.
- User Accessibility: Simplified interfaces, like those in Midjourney, reduce barriers for non-technical users.

3.3. Applications and Implications

State-of-the-art techniques empower industries such as publishing, where they facilitate the creation of compelling book illustrations, and education, where they enhance visual learning. By generating visually appealing and semantically accurate images, these models streamline creative workflows, allowing professionals to focus on refinement and storytelling.

4. Literature Review

In this review, we focus on exploring the diverse applications of Artificial Intelligence (AI) across various real-life domains. To provide a comprehensive analysis, we have identified key themes to guide our discussion: Research Purpose and Objectives, Methodology, Key Findings, Strengths and Contributions, Limitations, Practical Results and Implications and Implications and Future Directions. Each theme serves as a lens through which we examine how AI is being utilized to address specific challenges, enhance workflows, and introduce innovative solutions in different fields. For each application, a dedicated research paper is analyzed, offering unique insights and evidence to support the discussion within these themes. By structuring the review around these consistent themes, we aim to highlight the versatility of AI technologies, uncover trends in their adoption, and identify areas for further development and refinement to use in our project.

4.1. Analysing the Authenticity and Visual Appeal of AI-Generated Images

Authors in this study evaluate the aesthetic quality and realism of images generated by state-of-the-art AI systems, such as DALLE-2 and Midjourney, which are highly relevant for digital applications that require convincing visual appeal. The research specifically assesses how these images are perceived by users, examining qualities such as realism, detail, and aesthetic satisfaction [25].

AI-generated images have garnered significant attention recently, driven by advancements in artificial intelligence technology. Tools such as DALL-E-2, Midjourney, and Craiyon empower users to generate a wide range of visuals, from photorealistic to artistic or even humorous images, using simple text prompts[26]. Despite their impressive capabilities, the quality and realism of these generated images vary depending on the specific tool used and the text input provided. This study explores the extent to which AI-generated visuals achieve realism and visual appeal, comparing their outputs to real photographs for a comprehensive evaluationn text-to-image generators, including Stable Diffusion and Glide, utilize sophisticated methods such as diffusion models and Generative Adversarial Networks (GANs) to produce visuals from textual descriptions. These models work by iteratively de-noising input data, effectively transforming abstract textual concepts into coherent images. Some tools integrate frameworks like CLIP to improve semantic alignment, ensuring that the generated visuals closely match the input text. For example, Craiyon leverages BART encoding and VQGAN decoders to deliver computationally efficient image generation while maintaining semantic relevance and visual coherence[27, 28].

The authors in [29] evaluated the performance of various AI-generated image models by curating a dataset of 146 images, including real photographs and AI-generated visuals created from 27 text prompts sourced from benchmark collections like Draw Bench. This dataset emphasized realism and visual appeal, incorporating carefully crafted captions and high-quality image generation across diverse inputs. A subjective evaluation was conducted with 22 participants using an online testing platform to rate visual appeal, realism, and text alignment on a scale of 1 to 5. Results showed that models like Midjourney and DALL-E-2 outperformed others, such as Glide, in achieving higher visual appeal and realism. However, the study also highlighted that existing no-reference image quality models, such as NIMA and NIQE, are inadequate for evaluating AI-generated visuals. These models, trained on real-world photographs, fail to account for the unique distortions and aesthetic differences inherent in AI-generated content, pointing to the need for more tailored evaluation frameworks[30].

The subjective evaluations revealed that participants could generally distinguish between AI-generated visuals and real photographs. Among the models tested, DALL-E-2 produced the most realistic images, though none of the AI tools consistently matched the realism of the actual photographs in the dataset. To address low-resolution outputs from models like Glide and Craiyon, upscaling tools such as Real-ESRGAN and Topaz Gigapixel were employed, enhancing image clarity but occasionally introducing artifacts that diminished aesthetic appeal. The accuracy of AI tools in interpreting and aligning generated images with text prompts varied significantly; DALL-E-2 excelled in this aspect, creating visuals closely aligned with descriptions, while Glide frequently deviated from the prompts. Furthermore, the authors assessed how well AI-generated images adhered to photographic principles, such as the rule of thirds and simplicity, finding that most models, including DALL-E-2, struggled to consistently follow these guidelines. The study concludes that while AI generators hold significant promise for producing visually appealing and realistic images, their outputs remain distinguishable from real photographs. Future efforts should prioritize the development of more sophisticated evaluation models for AI-generated images and the improvement of datasets to better address the nuances of aesthetics and realism [31, 32].

4.2. Investigating Advancements in Text-to-Image Generation

Generative AI technologies have shown remarkable advancements in synthesizing images from textual descriptions, transforming diverse fields such as design, education, and digital art. The emergence of models like DALL-E-2, Midjourney, and Craiyon has enabled the creation of visually striking images with semantic alignment to text inputs. However, maintaining contextual coherence across sequences remains a challenge. Tools like GPT-4o excel at generating high-quality individual images but often struggle with sequence consistency, as evidenced by lower ROUGE-N recall scores and minimal inter-image similarity. This study reviews the evolution of text-to-image synthesis methods, highlighting key strengths, persistent challenges, and opportunities for refining coherence and resolution in AI-generated visuals. These insights underline the transformative potential of generative models while identifying critical areas for further innovation and application [33].

4.3. Artificial intelligence for content generation

Generative AI has emerged as a transformative technology capable of producing novel content[34], such as text, images, and audio, by learning patterns from extensive datasets [35]. Its applications range from artistic content creation to intelligent question-answering systems, revolutionizing industries reliant on creativity and knowledge processing. The technology is underpinned by generative modelling, which differs from traditional discriminative approaches by focusing on inferring data distributions to generate synthetic samples[36].

4.4. Handling Diverse Contexts in Text-to-Image Generative AI

The authors of this study aim to present a solution to the "context transition problem" in text-to-image generative artificial intelligence. This problem emerges when AI struggles to depict transitions between different scenes or subjects across a sequence of images generated from multiple input sentences [37].

4.5. Applying Generative AI Midjourney to Foster Divergent and Convergent Thinking Skills

The paper explores how Generative AI (GenAI) tools, particularly Midjourney, can support architects in enhancing their creative process[25, 38]. Creativity in design is often described through two complementary processes: divergent thinking, which involves generating a wide array of ideas, and convergent thinking, where ideas are narrowed down and refined into practical solutions. This duality forms the core of the paper's research questions.

5. Methodology

In this section, we explain the approach used to explore how Generative AI is shaping creative industries. The goal was to ensure a thorough and well-rounded review of the most relevant studies in this evolving field. The studies reviewed in this paper employed a variety of methodologies to explore the capabilities of AI-driven text-to-image generation tools. Below is a comprehensive summary, combining all methodological approaches highlighted in those studies.

5.1 Systematic Literature Review

A systematic review methodology was adopted in multiple studies to assess GAN-based text-to-image models published between 1997 and 2024[39]. These studies utilized datasets such as Oxford-102 Flowers, CUB-200 Birds, and COCO as shown in Fig. 2 to evaluate diverse methodologies and techniques. Evaluation metrics, including the Inception Score (IS) and Fréchet Inception Distance (FID), were used to analyze model performance, emphasizing semantic fidelity and visual appeal as shown in figure 2.

5.2 Experimental Design

Several papers utilized experimental designs with curated datasets, including real photographs and AI-generated visuals, to assess performance under controlled conditions. For example, one study created a dataset with 135 images as shown in figure 3 based on 27 carefully crafted text prompts. Participant evaluations were conducted to rate attributes like realism, visual appeal, and semantic alignment on a scale of 1 to 5 as shown in figure 3.

5.3 Tool-Specific Experiments

Studies often focused on particular AI tools like Midjourney, DALL-E-2, and Craiyon. For instance, Midjourney was explored for its ability to foster divergent and convergent thinking through iterative visual ideation. This study included two phases Generating visual outputs to refine user personas using detailed prompts like “Melbourne city” and “coffee culture”. And exploring abstract spatial designs prompted by keywords such as “ornithological habitat” and “symmetrical geometry.” Outputs were iteratively refined to develop five distinct spatial concepts.

6. Key Findings

Table 1 provides a concise overview of recent assessments of AI-driven text-to-image generation tools. It emphasizes the exceptional performance of DALL-E-2 in producing realistic and visually appealing images, the importance of detailed and specific text prompts, and the ability of tools like Midjourney to reduce cognitive effort during the creative process. Furthermore, it explores advancements in semantic alignment achieved through innovations such as attention mechanisms and GANs, addresses the challenges of maintaining contextual coherence in visual storytelling, and highlights the shortcomings of current evaluation metrics. The table also illustrates the potential of AI-generated imagery in applications like book illustration, where it can enhance reader engagement and creativity while offering cost-effective solutions.

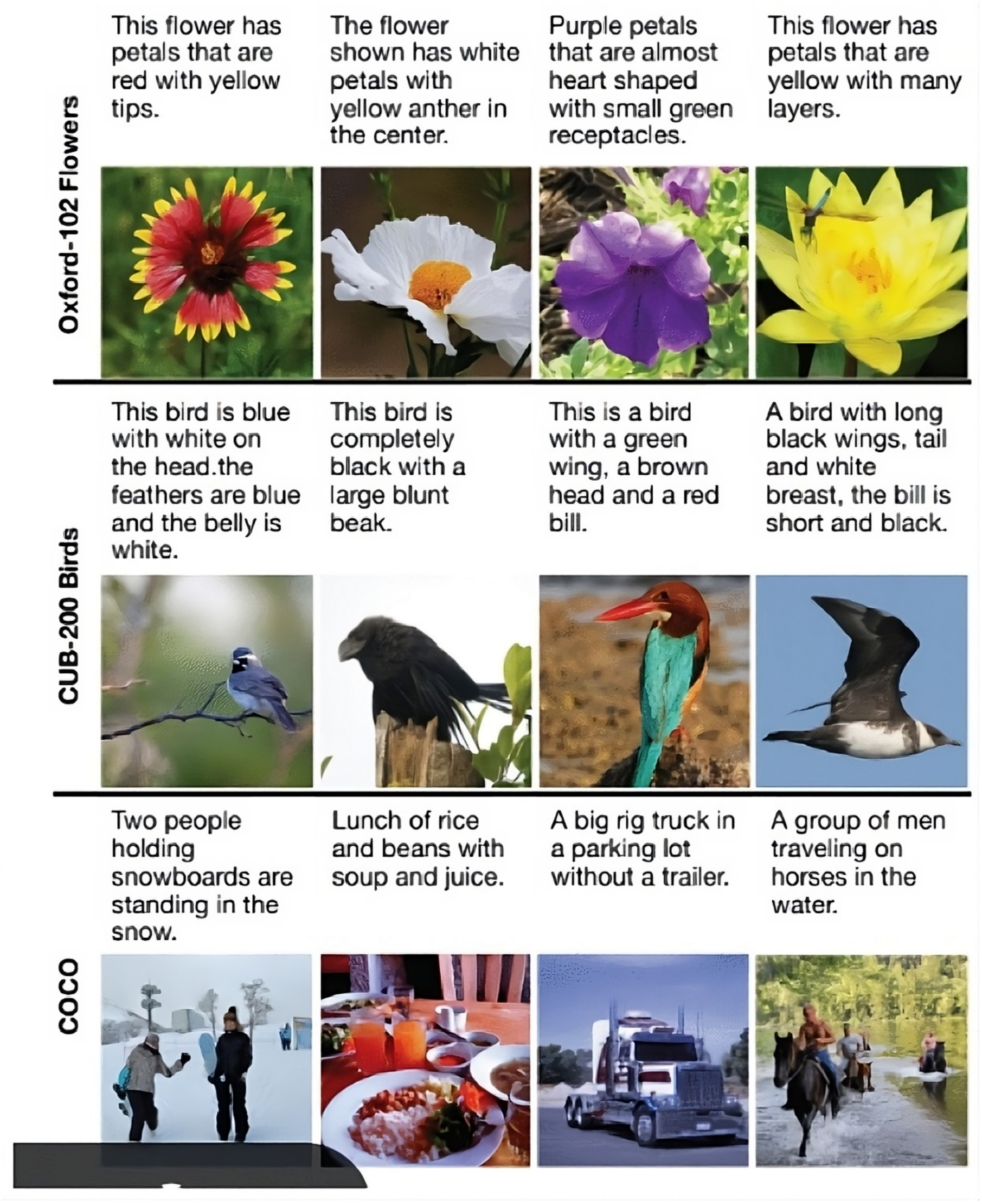
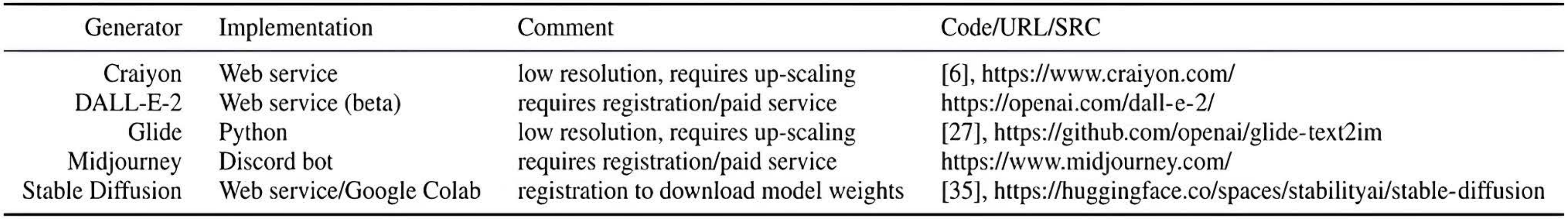

Oxford-102 Flowers
This flower has petals that are red with yellow tips.
The flower shown has white petals with yellow anther in the center.
Purple petals that are almost heart shaped with small green receptacles.
This flower has petals that are yellow with many layers.
CUB-200 Birds
This bird is blue with white on the head.the feathers are blue and the belly is white.
This bird is completely black with a large blunt beak.
This is a bird with a green wing, a brown head and a red bill.
A bird with long black wings, tail and white breast, the bill is short and black.
COCO
Two people holding snowboards are standing in the snow.
Lunch of rice and beans with soup and juice.
A big rig truck in a parking lot without a trailer.
A group of men traveling on horses in the water.

Fig 2. T2I Datasets Examples.



Generator	Implementation	Comment	Code/URL/SRC
Craiyon	Web service	low resolution, requires up-scaling	[6], https://www.craiyon.com/
DALL-E-2	Web service (beta)	requires registration/paid service	https://openai.com/dall-e-2/
Glide	Python	low resolution, requires up-scaling	[27], https://github.com/openai/glide-text2im
Midjourney	Discord bot	requires registration/paid service	https://www.midjourney.com/
Stable Diffusion	Web service/Google Colab	registration to download model weights	[35], https://huggingface.co/spaces/stabilityai/stable-diffusion

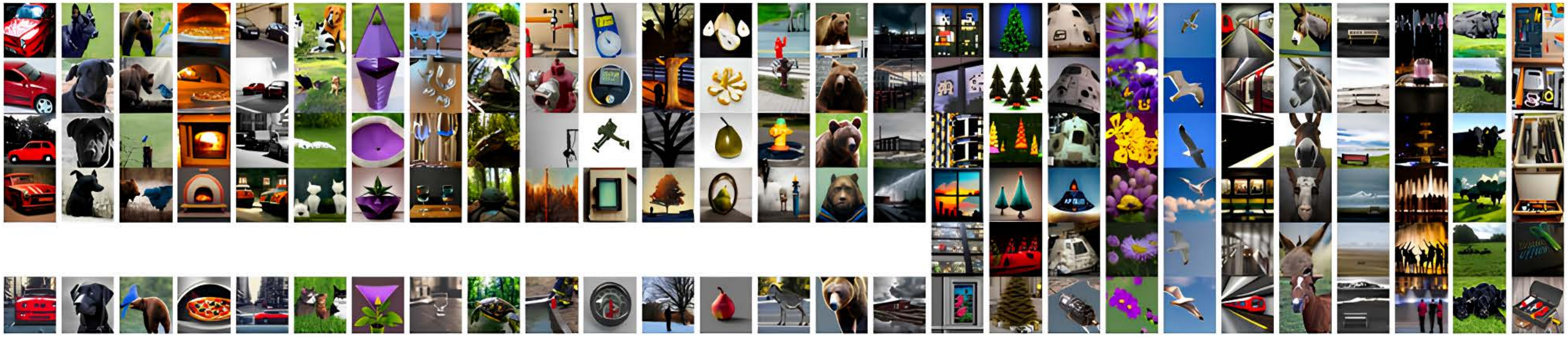


Fig 3. Image Dataset Comparison.

Table 1. Insights *on Text-to-Image Generative AI Performance and Applications*

Key findings	Details
Generator Performance[40]	Among the text-to-image generators evaluated, DALL-E-2 consistently emerged as the top performer, excelling in realism and user appeal. Midjourney followed closely, demonstrating high aesthetic quality and versatility. In contrast, Glide struggled to produce lifelike images, reflecting its limitations in achieving photorealistic outputs. These results highlight the importance of selecting the right generator for applications like book illustration, where high-quality visuals enhance reader immersion
Impact of Text Prompts[41]	The quality and appeal of AI-generated images were found to heavily depend on the specificity and detail of text prompts. Context-rich prompts led to more engaging and realistic visuals, underscoring the significance of effective prompt engineering. This is particularly relevant for narrative-based applications, such as book illustrations, where precise descriptions can help generate images that align with the storyline
Cognitive Load Reduction[42]	Generative AI tools like Midjourney significantly reduced cognitive load during the creative process as shown in figure 4. By automating visual ideation, these tools alleviated the manual effort required for sketching or modeling, allowing designers to focus on refining AI-generated visuals. This streamlined workflow was particularly advantageous for exploring complex design concepts and enhancing creativity.
Semantic Alignment and Attention Mechanisms[43]	Advancements in architectural innovations, such as attention mechanisms and stacked GANs, have improved semantic alignment between text inputs and generated visuals. These innovations enhanced image diversity and quality, addressing a key challenge in text-to-image synthesis. However, further improvements are required to achieve consistent semantic fidelity across diverse applications.
Evaluation Constraints[44]	Existing metrics, such as the Inception Score (IS) and Fréchet Inception Distance (FID), often fail to capture subjective aspects like aesthetic appeal and contextual accuracy. This limitation underscores the need for human-centric evaluation methods tailored to AI-generated content, particularly in creative domains where user perception plays a critical role
Consistency Challenges[45]	AI generators face challenges in maintaining contextual coherence across sequences, which is critical for applications like storytelling. Tools like GPT-4o, while adept at generating individual high-quality images, often struggled to create cohesive visual narratives. This was evident in the disjointed transitions between images and lower ROUGE-N recall scores.
Potential for Book Illustration[46]	High-quality, realistic images generated by tools like DALL-E-2 and Midjourney hold great potential for book illustration. These tools can help readers visualize intricate scenes, characters, or abstract concepts, making the reading experience more immersive. They also offer opportunities for authors and publishers to create compelling visual narratives at a fraction of traditional costs.

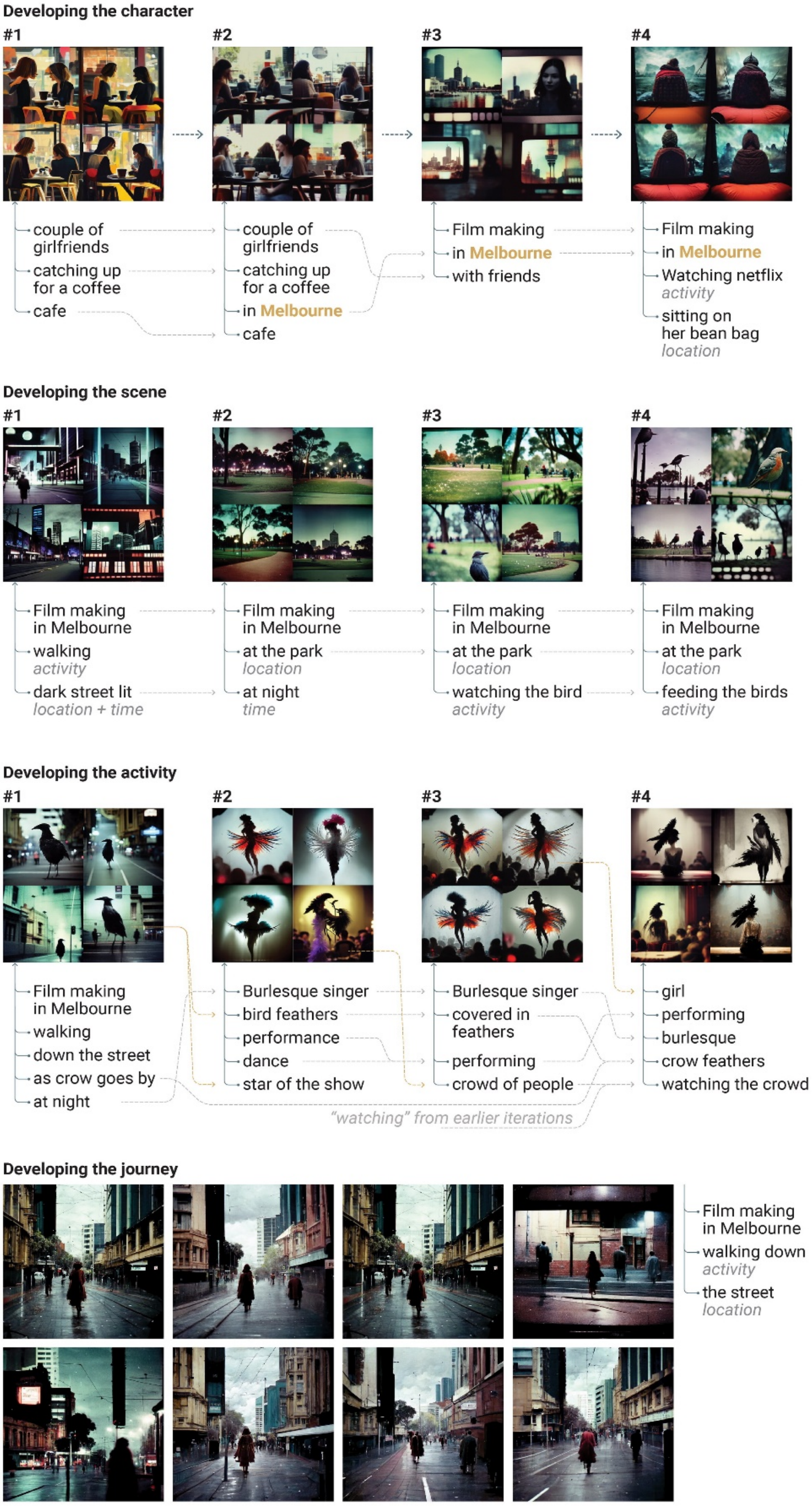


Fig 4. Midjourney Interaction Process

7. Limitations

The reviewed studies revealed several limitations inherent in AI-driven text-to-image generation tools[47-49]. These limitations span technical, methodological, and application-focused challenges as shown in table 2

Table 2 Limitations of AI-Driven Text-to-Image Generative Models

Limitations	Details
Evaluation Constraints	Many studies highlight the inadequacy of standard image quality metrics, such as Inception Score (IS) and Fréchet Inception Distance (FID), for evaluating AI-generated visuals. These metrics are often tailored for natural images and fail to capture semantic nuances and aesthetic qualities specific to generative content. User-centric assessment frameworks, like subjective evaluations, offer better insights but lack standardization, making cross-study comparisons difficult.
Contextual Coherence	Maintaining narrative consistency across sequences of images remains a significant challenge. Tools like GPT-4o often produce high-quality individual images but struggle with transitions and overall coherence, resulting in disjointed storylines. Existing models are better suited for generating single images, with limited ability to capture thematic or temporal relationships in multi-image narratives.
Dependence on Textual Prompts	The output quality is heavily reliant on the specificity and clarity of user-provided prompts. Designers or authors unfamiliar with prompt engineering may face difficulties achieving desired results.
Artistic and Compositional Limitations	Adherence to established artistic principles, such as the rule of thirds and simplicity, is inconsistent. Even advanced models like DALL-E-2 struggle to produce images that consistently meet these standards. The inability to maintain a consistent art style across a sequence of images poses challenges for applications requiring visual coherence, such as webtoons and illustrated books.
2D and Spatial Representation Constraints	Current models are predominantly limited to generating 2D outputs, restricting their use in fields requiring three-dimensional spatial design, such as architecture and virtual environments.
Software Evolution Challenges	Continuous updates to AI tools, such as Midjourney, introduce changes in style and algorithmic behavior. This disrupts long-term projects and undermines visual consistency in outputs.
Dataset Diversity and Generalizability	Limited dataset diversity affects the generalizability of models, making them less effective when generating visuals for underrepresented cultural or stylistic contexts.
Scalability for Long-Form Narratives	Most studies focused on short-form textual inputs. Handling longer narratives, such as novels or scripts, presents significant challenges, including maintaining character consistency and thematic progression over extended sequences.

8. Future Directions

Future research should concentrate on improving AI's contextual understanding in multi-image narratives, particularly for use in storytelling and design. Techniques such as hierarchical context management, which captures relationships between sentences and larger textual structures, could help models interpret and generate more complex narratives effectively. Creating tailored evaluation metrics for AI-generated content is also critical, as current measures like IS and FID frequently overlook subtle qualities like semantic alignment and emotional resonance. To address this, researchers could create metrics that include human-in-the-loop assessments of semantic consistency, sentiment analysis tools for assessing emotional resonance, and graph-based techniques for analyzing relationships between narrative elements. Furthermore, using multi-modal embeddings to assess alignment between text and images may provide a more comprehensive evaluation framework.

Incorporating human-centric assessments, such as reader engagement or the emotional impact of visuals, would provide a more comprehensive evaluation framework. Additionally, integrating 3D modeling capabilities into generative tools could expand their applicability to fields like architecture, product design, and immersive media, bridging the gap between 2D visualization and spatial representation.

Real-time interaction mechanisms, where users guide AI outputs dynamically during the generation process, could enhance usability and precision, allowing for outputs that meet specific creative requirements.

Finally, addressing ethical and legal challenges is essential as these tools become more prominent in creative industries. This includes establishing guidelines to ensure content inclusivity, transparency, and fair usage policies, particularly in resolving copyright concerns and promoting responsible AI deployment. By advancing these areas, the field can overcome current limitations, paving the way for more robust and versatile applications of generative AI.

9. Conclusion

This review paper provides a comprehensive exploration of the capabilities, challenges, and future directions of AI-driven text-to-image generation tools. The analysis spans various models, including DALL-E-2, Midjourney, Stable Diffusion, and Craiyon, highlighting their transformative potential in creative domains such as storytelling, design, and digital art. These tools have demonstrated remarkable advancements in generating visually appealing and semantically aligned images, democratizing creative processes and enabling non-experts to produce professional-grade visuals. By facilitating rapid ideation and iterative refinement, they bridge the gap between human creativity and technological innovation, offering unprecedented opportunities for fields that rely on visual storytelling and design thinking.

Despite these advancements, significant challenges remain. The reviewed studies emphasize limitations such as maintaining contextual coherence across sequential outputs, adherence to established artistic principles, and reliance on textual prompts that often yield variable results. Furthermore, the inadequacy of current evaluation metrics to assess the nuanced quality of AI-generated content underscores the need for more sophisticated assessment frameworks. These gaps highlight that while AI tools excel in generating individual visuals, achieving holistic and contextually consistent outputs remains a critical area for improvement.

Looking forward, the field must address these challenges through targeted research and technological refinement. Advancements in contextual understanding, integration of 3D modeling capabilities, and the development of tailored evaluation metrics are necessary to expand the applicability and utility of these tools. Furthermore, fostering interdisciplinary collaboration between technologists, designers, and ethicists can help ensure the responsible deployment of AI in creative industries, addressing ethical concerns and promoting inclusivity and transparency.

In conclusion, this review underscores the transformative impact of AI in redefining creativity and visual storytelling. As these technologies continue to evolve, they are poised to revolutionize how narratives are crafted, how designs are conceptualized, and how art is experienced. By overcoming current limitations and exploring emerging opportunities, generative AI has the potential to become a cornerstone of modern creativity, fostering innovation across disciplines and industries.

References

[1] Y. Hou, W. Zhang, Z. Zhu, and H. Yu, "Language-vision matching for text-to-image synthesis with context-aware GAN," Expert Systems with Applications, vol. 255, p. 124615, 2024.

[2] N. Hussen, M. Salem, A. I. El-Desouky, and S. M. Elghamrawy, "A Machine Vision Approach to Real-Time Drone Detection Using Gray Level Co-occurrence Matrix and Interactive GUI," in 2024 International Telecommunications Conference (ITC-Egypt), 2024: IEEE, pp. 554-560.

[3] H. Vartiainen and M. Tedre, "How Text-to-Image Generative AI Is Transforming Mediated Action," IEEE Computer Graphics and Applications, 2024.

[4] R. Gopalakrishnan, N. Sambagni, and P. Sudeep, "An Improved AttnGAN Model for Text-to-Image Synthesis," in International Conference on Computer Vision and Image Processing, 2023: Springer, pp. 139-151.

[5] Z. Zhao et al., "Masking-Based Cross-Modal Remote Sensing Image-Text Retrieval via Dynamic Contrastive Learning," IEEE Transactions on Geoscience and Remote Sensing, 2024.

[6] C. Leiter et al., "Chatgpt: A meta-analysis after 2.5 months," Machine Learning with Applications, vol. 16, p. 100541, 2024.

[7] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955," AI magazine, vol. 27, no. 4, pp. 12-12, 2006.

[8] C. Zhang and Y. Lu, "Study on artificial intelligence: The state of the art and future prospects," Journal of Industrial Information Integration, vol. 23, p. 100224, 2021.

[9] N. J. Nilsson, The quest for artificial intelligence. Cambridge University Press, 2009.

[10] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," metabolism, vol. 69, pp. S36-S40, 2017.

[11] R. S. Michalski, Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

[12] P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, "Natural language processing: History, evolution, application, and future work," in Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020, 2021: Springer, pp. 365-375.

[13] R. Olusegun, T. Oladunni, H. Audu, Y. Houkpati, and S. Bengesi, "Text mining and emotion classification on monkeypox Twitter dataset: A deep learning-natural language processing (NLP) approach," IEEE Access, vol. 11, pp. 49882-49894, 2023.

[14] I. Goodfellow et al., "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[15] E. Brophy, Z. Wang, Q. She, and T. Ward, "Generative adversarial networks in time series: A systematic literature review," ACM Computing Surveys, vol. 55, no. 10, pp. 1-31, 2023.

[16] G. Zhou et al., "Emerging synergies in causality and deep generative models: A survey," arXiv preprint arXiv:2301.12351, 2023.

[17] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, "The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges," Future Internet, vol. 15, no. 8, p. 260, 2023.

[18] N. R. Mannuru et al., "Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development," Information Development, p. 02666669231200628, 2023.

[19] R. Po et al., "State of the art on diffusion models for visual computing," in Computer Graphics Forum, 2024, vol. 43, no. 2: Wiley Online Library, p. e15063.

[20] S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers," IEEE Access, 2024.

[21] K. Jasmine, "Unlocking the power of prompt engineering: diverse applications and case studies," in Transforming Education With Generative AI: Prompt Engineering and Synthetic Content Creation: IGI Global, 2024, pp. 411-432.

[22] J. Hutson, J. Lively, B. Robertson, P. Cotroneo, and M. Lang, "Creative Convergence."

[23] X. Liu et al., "Toward the unification of generative and discriminative visual foundation model: a survey," The Visual Computer, pp. 1-42, 2024.

[24] T. Naik, H. Gostu, and R. Sharma, "Navigating Ethics of AI-Powered Creativity in Midjourney," in 2024 3rd International Conference for Innovation in Technology (INOCON), 2024: IEEE, pp. 1-6.

[25] L. Tan and M. Luhrs, "Using Generative AI Midjourney to Enhance Divergent and Convergent Thinking in an Architect's Creative Design Process," The Design Journal, pp. 1-23, 2024.

[26] M. Zaralli, Virtual Reality and Artificial Intelligence: Risks and Opportunities for Your Business. CRC Press, 2024.

[27] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!," in 2011 third international workshop on quality of multimedia experience, 2011: IEEE, pp. 131-136.

[28] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in 2012 IEEE conference on computer vision and pattern recognition, 2012: IEEE, pp. 2408-2415.

[29] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a “completely blind” image quality analyzer," IEEE Signal processing letters, vol. 20, no. 3, pp. 209-212, 2012.

[30] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1905-1914.

[31] S. Göring and A. Raake, "Rule of thirds and simplicity for image aesthetics using deep neural networks," in 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), 2021: IEEE, pp. 1-6.

[32] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative ai," Business & Information Systems Engineering, vol. 66, no. 1, pp. 111-126, 2024.

[33] S. Amershi et al., "Guidelines for human-AI interaction," in Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1-13.

[34] T. Dzieduszyński, "Machine learning and complex compositional principles in architecture: Application of convolutional neural networks for generation of context-dependent spatial compositions," International Journal of Architectural Computing, vol. 20, no. 2, pp. 196-215, 2022.

[35] A. Kuzior, M. Sira, and P. Brożek, "Effect of Artificial Intelligence on the Economy," Zeszyty Naukowe. Organizacja i Zarządzanie/Politechnika Śląska, no. 176 Contemporary management= Współczesne zarządzanie, pp. 319-331, 2023.

[36] H. Cao et al., "A survey on generative diffusion models," IEEE Transactions on Knowledge and Data Engineering, 2024.

[37] M. Florian, "Can Artificial Intelligence Systems like DALL-E or Midjourney Perform Creative Tasks?," ArchDaily, August, vol. 15, 2022.

[38] N. D’souza and M. R. Dastmalchi, "Architectural Creativity Stranded at Mid Journey? Evaluating Creative Potential of Prompts and Images in Generative AI," in International Conference on-Design Computing and Cognition, 2024: Springer, pp. 224-240.

[39] S. K. Alhabeeb and A. A. Al-Shargabi, "Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction," IEEE Access, 2024.

[40] O. Högblom and R. Andersson, "Analysis of thermoelectric generator performance by use of simulations and experiments," Journal of Electronic Materials, vol. 43, pp. 2247-2254, 2014.

[41] R. T. Miller, T. D. Mitchell, and S. Pessoa, "Impact of source texts and prompts on students’ genre uptake," Journal of Second Language Writing, vol. 31, pp. 11-24, 2016.

[42] L. Guo, "The Effects of the Format and frequency of prompts on source evaluation and multiple-text comprehension," Reading Psychology, vol. 44, no. 4, pp. 358-387, 2023.

[43] S. Cheng, L. Wang, A. Du, and Y. Li, "Bidirectional focused semantic alignment attention network for cross-modal retrieval," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: IEEE, pp. 4340-4344.

[44] M. Bamberger, J. Rugh, M. Church, and L. Fort, "Shoestring evaluation: Designing impact evaluations under budget, time and data constraints," The American Journal of Evaluation, vol. 25, no. 1, pp. 5-37, 2004.

[45] N. Ali, S. Baker, R. O’Crowley, S. Herold, and J. Buckley, "Architecture consistency: State of the practice, challenges and requirements," Empirical Software Engineering, vol. 23, pp. 224-258, 2018.

[46] P. Nodelman, "Picture books and illustration," in Intl Comp Ency Child Lit E2 V1: Routledge, 2018, pp. 154-165.

[47] R. Dhand et al., "Creating Realities: An In-Depth Study of AI-Driven Image Generation with Generative Adversarial Networks," in 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2024: IEEE, pp. 1-6.

[48] J. Agnese, J. Herrera, H. Tao, and X. Zhu, "A survey and taxonomy of adversarial neural networks for text-to-image synthesis," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 4, p. e1345, 2020.

[49] M. A. Habib et al., "Exploring Progress in Text-to-Image Synthesis: An In-Depth Survey on the Evolution of Generative Adversarial Networks," IEEE Access, 2024.