Revolutionizing Visuals: The Role of Generative AI in Modern Image Generation

GAURANG BANSAL, National University of Singapore, Singapore, Singapore
ADITYA NAWAL, Microsoft, India
VINAY CHAMOLA, BITS-Pilani, Pilani, India
NORBERT HERENCSAR, Brno University of Technology, Brno, Czech Republic

Traditional multimedia experiences are undergoing a transformation as generative AI integration fosters enhanced creative workflows, streamlines content creation processes, and unlocks the potential for entirely new forms of multimedia storytelling. It has potential to generate captivating visuals to accompany a documentary based solely on historical text descriptions, or creating personalized and interactive multimedia experiences tailored to individual user preferences. From the high-resolution cameras in our smartphones to the immersive experiences offered by the latest technologies, the impact of generative imaging undeniable. This study delves into the burgeoning field of generative AI, with a focus on its revolutionary impact on image generation. It explores the background of traditional imaging in consumer electronics and the motivations for integrating AI, leading to enhanced capabilities in various applications. The research critically examines current advancements in state-of-the-art technologies like DALL-E 2, Craiyon, Stable Diffusion, Imagen, Jasper, NightCafe, and Deep AI, assessing their performance on parameters such as image quality, diversity, and efficiency. It also addresses the limitations and ethical challenges posed by this integration, balancing creative autonomy with AI automation. The novelty of this work lies in its comprehensive analysis and comparison of these AI systems, providing insightful results that highlight both their strengths and areas for improvement. The conclusion underscores the transformative potential of generative AI in image generation, paving the way for future research and development to further enhance and refine these technologies. This article serves as a critical guide for understanding the current landscape and future prospects of AI-driven image creation, offering a glimpse into the evolving synergy between human creativity and artificial intelligence.

CCS Concepts: • **Computing methodologies** → **Image representations**;

Additional Key Words and Phrases: Generative AI, LLMs, Image Generation, Computing, Multimedia Computing

ACM Reference format:
Gaurang Bansal, Aditya Nawal, Vinay Chamola, and Norbert Herencsar. 2024. Revolutionizing Visuals: The Role of Generative AI in Modern Image Generation. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 11, Article 356 (November 2024), 22 pages.
https://doi.org/10.1145/3689641

Authors' Contact Information: Gaurang Bansal, National University of Singapore, Singapore, Singapore; e-mail: gaurang@u.nus.edu; Aditya Nawal, Microsoft, Hyderabad, India; e-mail: aditya1nawal@gmail.com; Vinay Chamola, BITS-Pilani, Pilani, India; e-mail: vinay.chamola@pilani.bits-pilani.ac.in; Norbert Herencsar (corresponding author), Brno University of Technology, Brno, Czech Republic; e-mail: herencsn@vut.cz.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2024 Copyright held by the owner/author(s).
ACM 1551-6865/2024/11-ART356
https://doi.org/10.1145/3689641

1 Introduction

The ever-evolving field of multimedia encompasses various communication elements that combine text, audio, images, and video to deliver information and experiences. Multimedia plays a vital role in our daily lives, influencing how we learn, create, and interact with the world around us. From educational resources and interactive entertainment to captivating marketing campaigns and social media platforms, multimedia content shapes our understanding and engagement with information. The field of **artificial intelligence (AI)** has made remarkable strides in recent years, and one of its most intriguing and impactful developments is **generative AI (GAI)** in imaging [15, 16] in electronic devices. GAI refers to a class of machine learning models that have the ability to create new data instances that resemble, or even transcend, human-generated content. In other words GAI represents a transformative branch of AI that focuses on the development of algorithms and models capable of autonomously generating data, such as images, text, audio, and more, that closely resembles content created by humans. The foundation of GAI lies in deep learning methodologies, particularly neural networks [3].

In the context of imaging, this technology has opened up a world of possibilities for creative professionals, medical practitioners, and researchers alike [1, 12]. This technology is revolutionizing imaging by empowering creators, personalizing **user experiences (UXs)**, and increasing accessibility. GAI automates tasks, generates variations of content, and creates entirely new visuals, freeing creators to focus on storytelling and design. It personalizes images by tailoring content to user preferences and fosters inclusivity by generating captions, translating languages, and creating image descriptions. These advancements mark a significant step forward in the way we create and experience image content. GAI in imaging has left an indelible mark on various facets of our lives, ushering in a new era of visual content creation and manipulation. Its impact is felt across diverse domains, from art and entertainment to healthcare and beyond. In the realm of art and design, artists now collaborate with AI to push the boundaries of creativity, resulting in mesmerizing artworks and innovative designs that were once inconceivable. Photographers and historians benefit from AI's ability to restore and colorize old, cherished images, bringing history to life in vivid detail [13]. In healthcare, GAI assists medical professionals by generating synthetic medical images for training diagnostic algorithms, ultimately improving patient care [14, 16].

This article is comprehensive study of impact of GAI models in imaging. The contributions of the article are highlighted as follows:

(1) *Comprehensive Model Analysis for Multimedia Applications:* This article conducts an extensive comparative analysis of various **generative adversarial networks (GANs)**, **variational autoencoders (VAEs)**, and other GAI models. It evaluates their architectural nuances, training methodologies, and capabilities in generating high-quality multimedia content, aiding researchers in selecting the most appropriate model for specific multimedia tasks.
(2) *Integration of Advanced Multimedia Techniques:* The article explores the integration of cutting-edge multimedia techniques, such as medical imaging, video synthesis, style transfer, and GAI models. This integration demonstrates how GAI can enhance data acquisition, denoising, and multimedia reconstruction, improving the quality and accuracy of medical, scientific, and artistic visual content.
(3) *Assessment of State-of-the-Art Multimedia Technologies:* It offers an exhaustive survey and critical assessment of state-of-the-art GAI technologies within the multimedia domain. This includes an in-depth examination of the latest advancements such as Imagen, DALLE-2, and Stable Diffusion, helping researchers stay current with the most innovative techniques for multimedia creation.

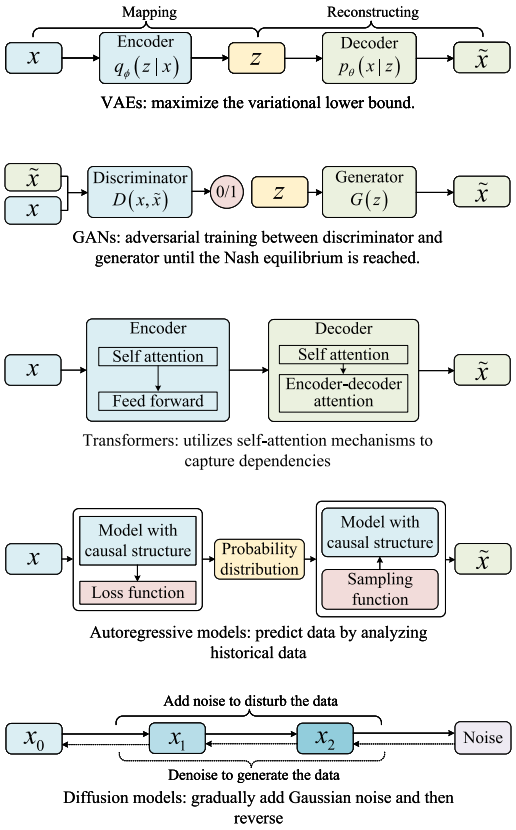


Fig. 1. Explores diverse architectures of GAI models, including VAE, GANs, autoregressive, diffusion, and transformers. It illustrates the components of input, transformation process, and resultant output within these models.

(4) *Addressing Multimedia Challenges and Limitations:* In addition to highlighting the strengths of GAI in multimedia, the article elucidates the associated challenges and limitations. It delves into issues such as mode collapse, training instability, and ethical concerns related to deepfakes in multimedia. Moreover, it proposes mitigation strategies and novel research directions to address these issues, advancing the robustness and ethical use of GAI in multimedia contexts.

2 GAI Background

Generative models, a subset of machine learning models, play a pivotal role in creating data that mimic real-world examples. They enable us to generate new, synthetic data instances that resemble the distribution of the training data. As depicted in Figure 1, these generative models can be broadly categorized into five major model categories, each with its unique approach and characteristics:

(1) *VAEs:* VAEs are generative models that acquire a compact representation, referred to as a latent space, of the input data. Comprising an encoder network for mapping input data to latent space and a decoder network for data reconstruction from latent representation, VAEs are commonly applied to tasks such as image generation and data synthesis.

(2) *GANs:* GANs encompass a generator network and a discriminator network engaged in a minimax game. The generator network produces synthetic data instances, while the discriminator network undertakes the task of discerning real from generated data. GANs are extensively utilized for the creation of realistic images, videos, and diverse data forms.
(3) *Transformers:* Transformers, grounded in attention-based mechanisms, excel in capturing intricate dependencies across extensive sequences. Their utility extends to various generative tasks encompassing **natural language processing (NLP)**, language translation, and image synthesis. Self-attention mechanisms within Transformers facilitate the processing of input sequences and the generation of high-quality outputs.
(4) *Autoregressive Models:* Autoregressive Models unfurl data sequentially, where each sequential element conditions on its preceding constituents. Employed extensively in text, image, and other sequential data generation, Autoregressive Models hold significance in the generation of data that upholds temporal coherence.
(5) *Diffusion Models:* Diffusion models center their focus on comprehending the denoising procedure by directly modeling the noisy data. Rooted in principles stemming from non-equilibrium thermodynamics, these models emulate source datasets by adhering to predefined Markov chains of diffusion iterations.

Today GAI has profoundly transformed the field of imaging. It leverages advanced machine learning techniques to create, enhance, and manipulate images in ways that were once considered the realm of science fiction. This transformative technology is centered around the development of algorithms and models that can autonomously generate images, modify existing ones, or even fill in missing information within images.

3 Exploring the Multimedia Domains

Multimedia encompasses a diverse spectrum of nine distinct categories, each contributing uniquely to our understanding and utilization of visual information. These categories can be delineated as follows based on [2, 6]:

(1) *Image Synthesis:* Generating high-quality image of objects, scenes, or even entirely new concepts. This has applications in art, design, and entertainment. For instance, producing lifelike depictions of fantastical landscapes for video games or crafting realistic-looking creatures that exist only in digital realms.
(2) *Style Transfer:* Altering the style of an image while preserving its content. For instance, making a photograph look like it was painted by a famous artist. By mimicking the techniques of famous artists, it transforms ordinary pictures into artistic masterpieces. An example is turning a regular photograph into a painting that captures the distinct brushwork and color palette of Vincent van Gogh.
(3) *Data Augmentation:* Generating new training data to improve the performance of machine learning models by introducing diversity in the training set. For instance, replicating an image of a car under various lighting conditions to help a self-driving car's recognition system adapt to different environments. It enhances model robustness by introducing variations like lighting changes, rotations, or object overlays.
(4) *Image-to-Image Translation:* Converting image from one domain to another, such as turning satellite image into maps or black-and-white photos into color. This can be transforming aerial satellite image into user-friendly maps or converting black-and-white historical photos into vibrant color representations, bridging the gap between past and present.

(5) *Super-Resolution:* Enhancing the resolution of images, making them clearer and more detailed. This is significant for tasks like improving the fidelity of surveillance camera footage or restoring old photographs to reveal finer details that were previously obscured.
(6) *Anomaly Detection:* Creating models that can generate "normal" image and then identifying anomalies by detecting deviations from the normal patterns. By training on standard data, the model can spot irregularities, like identifying defective products on an assembly line through deviations from the standard appearance.
(7) *Face Generation:* Creating realistic human faces, which has applications in video games, movies, and even generating avatars. It enables video games to feature lifelike characters, movies to use digital actors, and individuals to create unique avatars for online identities.
(8) *Medical Imaging:* Generating synthetic medical images for research, training, and augmentation of datasets. It aids in developing and fine-tuning diagnostic tools by generating diverse images of medical conditions or rare cases that might not be abundant in real-world datasets.
(9) *Data Compression:* Creating compressed representations of images, which can be useful for efficient storage and transmission. By reducing redundant information while preserving critical details, it enables quicker data transfer and efficient usage of storage space, critical for transmitting high-resolution images over limited bandwidth or storing vast images libraries.

4 Amalgamation of GAI in Metaverse

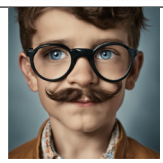
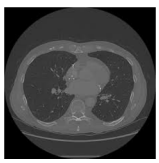
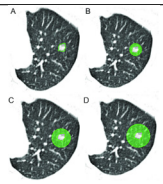
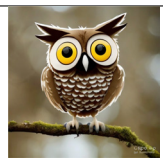
In this section, we elucidate the utilization of various GAI models to generate diverse domains of imaging, as outlined in Table 1. The amalgamation of these five robust GAI models—VAE, Diffusion Models, GANs, Transformers, and Autoregressive Models—yields a versatile and dynamic approach with the potential to foster innovation and advancements across a broad spectrum of imaging domains. This synthesis not only unifies distinct methodologies but also unveils promising avenues for exploration, experimentation, and groundbreaking developments in the continually evolving landscape of imaging technology and its applications.

In image synthesis, VAEs play a pivotal role by creating a structured latent space representation, enabling the generation of diverse image concepts. GANs complement this process by fine-tuning these generated images to ensure high levels of realism. Transformers provide contextual coherence to the synthesized scenes, allowing for the understanding of spatial relationships and contextual cues. Diffusion models come into play to refine image details and enhance the overall visual quality. Autoregressive Models, with their sequential prediction capabilities, further refine generated images by capturing fine-grained features, such as intricate textures and patterns. For instance, in the domain of generating realistic faces, VAEs might capture the structural elements of the face, GANs add facial features, Transformers ensure realistic facial expressions and poses, and Autoregressive Models capture nuanced textures like skin pores and hair strands.

Style transfer, another crucial domain, leverages VAEs to extract the content information from the source image, preserving its structural elements. GANs are adept at capturing the stylistic elements from a target artist or image, applying techniques such as texture and color mapping to transform the image's style. Autoregressive Models excel in fine-tuning the stylized image by preserving essential details and harmonizing the content with the newly applied style. This collaborative approach empowers the transformation of an ordinary photograph into a visually striking masterpiece, imitating the brushwork of artists like Vincent van Gogh or the surrealistic tendencies of Salvador Dali.

For data augmentation, GANs generate diverse variations of images, introducing novel elements, environmental changes, or variations in lighting conditions. This diversification enhances the robustness of machine learning models by exposing them to a broader spectrum of training data.

Table 1. Delineates the Varied Roles of Distinct GAI Models across Imaging Domains

Category	Type	Input 1	Type	Input 2	Model	Output
Image Synthesis	Text	"A Bee flying out of a glass jar in a green and red leafy basket"	-	-	Diffusion	
Style Transfer	Image		Image		GANs, VAEs	
Data Augmentation	Image		Text	"Change type to black and white"	Image manipulation methods	
Image-to-Image Translation	Image		Text	"Transform doodle to complete image"	GANs	
Super-Resolution	Image		Text	"Enhance the resolution of the given image while preserving its content and maintaining natural details."	Diffusion	
Anomaly Detection	Image		Text	Detect the anomaly in the given image	VAEs	
Face Generation	Text	"A handsome boy with moustache having brown hair and blue eyes wearing rounded googles"	-	-	GANs	
Medical Imaging	Image		Text	Automatic tumour detection	GANs, VAEs	
Data Compression	Image		Text	"Compress the image to reduce its file size while retaining essential visual information and minimizing loss in image quality."	GANs	

It specifies input types (Type), the primary input (Input 1), a secondary input (Input 2), and the resulting image (Output). Each model's utilization is categorized by the Imaging domain.

VAEs ensure that the generated data maintain semantic integrity and are meaningful within the given context.

In image-to-image translation, Transformers excel at handling complex domain mappings, such as translating satellite images into detailed maps or converting monochromatic historical photos into vibrant color representations. GANs are instrumental in maintaining visual realism during this translation process, ensuring that the transformed images are visually convincing. Autoregressive Models come into play to fine-tune details and textures in the translated output, further enhancing the quality of the final result.

Super-resolution tasks benefit from the collaborative efforts of these models, with VAEs generating a high-level structural representation of the image, GANs adding finer details and textures, and Diffusion models reducing noise and artifacts, resulting in clearer and more detailed images.

In the domain of anomaly detection, VAEs are adept at learning representations of normal images. GANs then generate diverse "normal" samples, allowing for a comprehensive understanding of what constitutes typical data. Detection models can identify anomalies by comparing real data with generated samples, effectively spotting deviations and outliers.

Face generation, a domain with wide applications in entertainment and personalization, benefits from GANs creating realistic facial features and expressions. VAEs ensure diversity in generated faces, preventing repetition, and Transformers handle facial landmarks and pose, enabling the generation of expressive and lifelike digital avatars.

In medical imaging, VAEs generate synthetic representations of medical conditions, aiding research, training, and dataset augmentation. GANs add realism and variability to these synthetic medical images, ensuring that they closely resemble real-world clinical cases. Autoregressive Models refine anatomical details, enhancing the diagnostic utility of the generated images.

Finally, data compression becomes more efficient with VAEs creating compact latent representations of images, which encapsulate essential information while reducing redundancy. Transformers introduce context-aware compression techniques, optimizing the encoding process. Diffusion models further enhance compression by reducing noise and minimizing information loss, facilitating efficient storage and transmission of high-resolution images over bandwidth-constrained networks.

This collaborative integration of GAI models, with their specialized capabilities, ushers in a new era of imaging, characterized by unprecedented levels of quality, diversity, and adaptability across a spectrum of domains, from healthcare and entertainment to data management and beyond.

5 State-of-Art Technologies

After providing an overview of the fundamental GAI models and their significant influence, our exploration takes a more detailed and focused approach. We will delve into the specific technologies that are used in practical applications within the domain of image synthesis from text descriptions. This deeper dive will illuminate the unique contributions and methodologies associated.

The basis for model selection included several key criteria such as:

- *State-of-the-Art:* We prioritized models like DALL-E 2, Imagen, and Jasper because they represent the cutting edge of text-to-image generation. These models are constantly evolving and pushing the boundaries of what's possible.
- *Technical Diversity:* We aimed to showcase a range of technical approaches. DALL-E 2 uses a combination of transformers and VAEs, while Deep AI is more grounded in GANs. This allows for a more comprehensive understanding of the different techniques driving this field.
- *Accessibility:* The selection includes both open-access options like NightCafe and Stable Diffusion, alongside research-focused models like DALL-E 2. This provides insights into both cutting-edge advancements and user-friendly tools.

—*Strengths and Applications:* We considered models with distinct strengths. For instance, Midjourney is known for its artistic capabilities, while Imagen prioritizes semantic accuracy. This showcases the diverse applications of text-to-image generation.

By considering these criteria, we aimed to present a well-rounded comparison that highlights the innovation, technical approaches, and practical applications within the realm of text-to-image synthesis.

Thus, the list of diverse applications of these cutting-edge GAI models in various multimedia domains considered are

—*DALL-E 2:* DALL-E 2 is a state-of-the-art text-to-image generative model, building upon its predecessor, DALL-E. Leveraging a synthesis of transformer and VAE techniques, this model excels at transmuting textual descriptions into intricate visual representations. DALL-E 2 demonstrates remarkable prowess in comprehending and reproducing complex scenes, objects, and abstract concepts from textual inputs. It stands out for its capacity to produce high-resolution images with exceptional detail, making it a formidable tool for applications demanding creative visual synthesis [24]. Within the multimedia landscape, DALL-E 2's capabilities shine in several areas. It can craft high-resolution fantasy landscapes for immersive video games, seamlessly transform photos into artistic masterpieces using style transfer, and breathe life into history by converting black and white photos into vibrant color representations. Additionally, DALL-E 2 can generate realistic soundscapes to complement visuals in movies, video games, or **virtual reality (VR)** experiences. Its versatility and attention to detail make it a valuable tool for multimedia creators seeking to enhance their projects with high-quality visuals and sound.

—*Deep AI:* Deep AI is a class of generative models, typically grounded in the GAN architecture. While specifics may vary across implementations, Deep AI's core objective is to convert textual descriptions into corresponding images by learning from extensive datasets containing text–image pairs. Commonly, this involves the utilization of convolutional and recurrent neural networks to process textual input and generate corresponding visual content. Technical implementations may diverge, reflecting variations in architectural choices and training methodologies [26].

—*Imagen:* Imagen is a model that harnesses NLP techniques and deep neural networks. It commences by processing textual descriptions through NLP models to extract semantic information. Subsequently, a generative network, often grounded in GANs or VAEs, employs this information to craft visually coherent images. Imagen places substantial emphasis on capturing the essence of textual descriptions, emphasizing realism and contextual relevance to produce semantically meaningful visual outputs [23]. Within the multimedia sector, Imagen excels at image-to-image translation, converting multimedia content from one domain to another. This includes transforming complex medical scans into user-friendly visualizations or historical sketches into **three-dimensional (3D)** models. Moreover, Imagen is adept at face generation, creating realistic human faces for use in video games, movies, or avatar creation, which opens new possibilities for character design and personalization. It also excels in video generation, producing short video clips based on textual descriptions or existing multimedia content, thus enabling innovative storytelling and content creation. Imagen's versatility makes it a powerful tool for multimedia creators looking to push the boundaries of their work.

—*Jasper:* Jasper model is predicated on deep neural networks and attention mechanisms. It prominently employs transformer-based architectures, endowing it with the capability to assimilate global contextual information from textual input and generate corresponding images

with fidelity. Jasper excels in comprehending intricate textual descriptions and translating them into visually coherent images, rendering it suitable for a broad spectrum of applications, including artistic content generation [22]. Jasper's strengths lie in data augmentation, generating new training data with diverse variations to enhance the performance of machine learning models in various multimedia tasks beyond image recognition. Additionally, Jasper excels in anomaly detection, identifying deviations from normal patterns within multimedia content, which aids in quality control processes and anomaly detection systems. Jasper's focus on data manipulation empowers multimedia developers to create robust and efficient AI systems, enhancing the quality and reliability of their projects.

— *NightCafe:* NightCafe is a text-to-image generative model that adopts relatively simpler neural network architectures for the task. Consequently, its capacity to map textual descriptions to images is less sophisticated compared to certain counterparts. NightCafe's performance is often characterized by a propensity to generate images of lower realism and semantic coherence, yielding results that may fall below the desired standards. Such limitations may be attributed to less intricate training strategies or network architectures [21]. Within the multimedia domain, NightCafe finds applications in several key areas. It excels in face generation, creating realistic human faces for video games and movies, which offers new possibilities for character design and personalization. NightCafe also excels in image-to-image translation, converting multimedia from one domain to another, such as transforming aerial photographs into 3D terrain maps or historical sketches into full-color animations. Additionally, NightCafe is proficient in music generation, creating new musical pieces based on genre, mood, or specific instructions. This capability allows for generating background music for a historical documentary that reflects the era or composing a soundtrack that perfectly complements the visuals of a video game. NightCafe's emphasis on realism makes it a valuable tool for creating immersive and believable multimedia experiences.

— *Midjourney:* Midjourney relies on neural networks for the task. However, it tends to prioritize computational efficiency over the generation of highly detailed and realistic images. As a consequence, Midjourney often simplifies complex scenes and may exhibit limitations in capturing fine-grained visual details. Its technical approach reflects a tradeoff between speed and computational resources versus the quality and intricacy of generated images [9].

— *Stable Diffusion:* Stable Diffusion constitutes a generative model engineered to strike a balance between the fidelity of image generation and computational stability. Often rooted in architectures like GANs, Stable Diffusion may incorporate diffusion models into its framework to synthesize images. While it is proficient in generating realistic images with commendable detail, occasional distortions may arise during the generation process. The technical underpinnings of Stable Diffusion are meticulously designed to deliver dependable performance across a diverse array of text-to-image generation tasks [7]. Stable Diffusion prioritizes quality and robustness in multimedia generation. Its applications within the multimedia sector are diverse and impactful. It excels in style transfer, altering the artistic style of images, audio, or video while preserving their content. This capability allows for creative transformations, such as turning a historical film into a modern action movie aesthetic or applying a classic rock style to a contemporary song. Additionally, Stable Diffusion is proficient in super-resolution, enhancing the resolution and detail of multimedia content, thereby improving the clarity of low-quality audio recordings or grainy historical videos. Its emphasis on quality control ensures that the generated multimedia is both aesthetically pleasing and rigorously scrutinized.

— *Craiyon:* Craiyon offers state-of-the-art technology for content creation and image editing, particularly within the multimedia sector. It excels at data augmentation by generating diverse variations of existing multimedia content—images, audio, and video—to improve machine

learning models in tasks such as speech recognition and music composition. Furthermore, Craiyon is adept at anomaly detection, identifying deviations from normal patterns in multimedia, such as spotting unusual activity in surveillance footage or detecting inconsistencies in audio recordings. This focus on data manipulation empowers multimedia developers to train more robust and efficient AI systems, enhancing the quality and reliability of their projects.

6 Comparison of Technologies for Image Synthesis

After having discussed the overview of the major technologies, we proceed to offer a comprehensive comparison of these technologies based on the following parameters presented in Table 2. Our objectives encompass two main aspects: first, to provide an exhaustive overview covering aspects such as performance and image synthesis quality, and second, to enable an informed comparison of these technologies in the context of UX, customization options, and technical details. Through this comparative analysis, our goal is to illuminate the strengths, limitations, and ideal use cases associated with each approach. This information is intended to empower professionals, researchers, and enthusiasts, aiding them in making well-informed decisions when selecting the most suitable image synthesis technology tailored to their specific requirements.

(1) *Performance and Robustness:*
- —*Image Realism:* This criterion assesses the degree to which generated images closely resemble real-world visuals. DALL-E 2 stands out with the highest level of image realism, producing outputs that are virtually indistinguishable from actual images. For instance, in a VR game set in a historical era, DALL-E 2 could generate period-accurate visuals, enhancing the authenticity of the metaverse experience. MidJourney also achieves remarkable realism, enhancing its potential for immersive experiences. This high level of realism elevates user engagement and immersion.
- —*Quality of Image Generation:* Evaluating the quality of generated images is vital. DALL-E 2 and Deep AI provide good to high-quality outputs, suitable for a wide range of applications. For example, Deep AI's high-quality outputs can be leveraged for medical imaging in the metaverse, aiding in diagnostics and surgical planning. In contrast, Imagen focuses on producing photorealistic images, and Stable Diffusion ensures both stability and photorealism in its results. This level of quality contributes to more accurate virtual environments and enhances user presence in the metaverse.
- —*Scalability:* The scalability of generative models plays a crucial role in accommodating various scales of usage. NightCafe and Jasper excel in this aspect, showcasing commendable scalability that allows them to generate content efficiently for large datasets and user bases. This scalability is particularly valuable in scenarios like virtual events, where a large number of users simultaneously interact with generated content. However, MidJourney and Imagen exhibit certain limitations when it comes to scalability. This limitation might impact their usability in applications that require rapid content generation for a large audience.
- —*Robustness:* Ensuring robustness against variations is essential for consistent performance. Deep AI and Imagen exhibit high robustness, making them resilient to diverse input conditions and variations. This robustness is crucial in metaverse applications where the generated content needs to remain consistent across different user interactions. NightCafe and Stable Diffusion also possess robustness, but they come at the expense of compromising user control. On the other hand, Jasper faces challenges in achieving robustness while maintaining user control. This tradeoff might affect its suitability for applications where both robustness and user-directed content are critical.

—*Semantic Coherence:* The coherence of generated content in maintaining contextual meaning and relevance is crucial. Both DALL-E 2 and MidJourney shine in maintaining semantic coherence, ensuring that the generated content aligns contextually and makes sense within the given input. For instance, DALL-E 2 can generate coherent historical narratives based on user prompts, enhancing the storytelling potential of the metaverse. Additionally, Deep AI, Imagen, and Stable Diffusion also perform well in maintaining semantic coherence. This semantic alignment ensures that user interactions remain meaningful and contextually relevant.

(2) *Customization and Control:*

—*Diversity and Creativity:* Diverse and creative outputs are highly desirable in generative models. DALL-E 2, Deep AI, and Imagen excel in providing a wide range of diverse and imaginative outputs, catering to a variety of creative needs. For example, DALL-E 2's diverse outputs can be used in art creation, generating novel visual concepts that spark creativity and innovation.

—*Control and Manipulation:* The extent of user control and manipulation capabilities offered by generative models significantly impacts their usability. DALL-E 2 empowers users with control over various aspects of output, such as style, color, and composition. This level of control enables users to customize generated images for specific artistic expressions. Deep AI, Imagen, and Stable Diffusion enable customizable style parameters, allowing users to tailor the generated content to their preferences. However, Jasper lacks user-controlled effects, potentially limiting creative freedom [17]. This limitation might impact users who seek highly personalized content.

—*Learning from Imperfect Data:* The ability of models to learn from imperfect or noisy data is crucial for adapting to real-world scenarios. Most models demonstrate effective learning from imperfect data, enabling them to generate quality results even when faced with incomplete or noisy input information. This adaptability is particularly valuable for scenarios where input data may have variations, such as user-generated content [8].

—*User Control:* The degree to which users can influence and guide the generative process impacts user satisfaction. DALL-E 2 provides users with high levels of control, allowing them to shape the output according to their preferences. Users can adjust parameters to match their creative vision, resulting in highly personalized outputs. In contrast, models like Deep AI and Imagen offer limited user control. NightCafe and Stable Diffusion further limit user control, and Jasper completely lacks user control options. The availability of user control can impact user engagement, especially for users who desire hands-on customization.

—*Fine-Tuning Efficiency:* Fine-tuning models for specific tasks or styles is an essential consideration. DALL-E 2 exhibits high fine-tuning efficiency, making it well-suited for customizing its outputs for specific needs. For instance, a user creating content for a specific visual theme can fine-tune DALL-E 2 to produce images that align with that theme. Deep AI and Imagen offer moderate fine-tuning efficiency. In contrast, Jasper's fine-tuning capability is not applicable. Fine-tuning efficiency affects the ease and speed of adapting the model to user requirements.

(3) *Ethical and Accessibility:*

—*Bias and Fairness:* Ensuring fairness and mitigating biases in generative models is a critical ethical consideration. While specific evaluation is challenging and not specified in the table, models like Imagen and Stable Diffusion emphasize the importance of addressing biases. This impact is significant in applications where biased content could lead to misinformation or exclusion of certain groups.

Table 2. Comparison of GAI Technologies in the Imaging

Evaluation Criteria	Parameter	DALL-E 2	Deep AI	IMAGEN	Stable Diffusion	Jasper	NightCafe	MidJourney
Performance and Robustness	Image realism	Highest	High	High	High	Low	Medium	Highest
	Quality of image generation	Good quality	High-quality, diverse	Photorealistic	Photorealistic, stable	Recognizable	Photorealistic, accessible	Good quality
	Scalability	Good scalability	Scalable to large datasets, users	Large datasets, high resolutions	Scalable to large	Scalable to large datasets	Fair scalability	Good scalability
	Robustness	Medium	High	Not evaluated	High	Low	Low	High
	Semantic coherence	Highest	High	High	High	Low	Medium	Highest
Customization and Control	Diversity and creativity	Diverse, creative	Customizable	Diverse, creative	Diverse	Diverse, creative	Customizable	-
	Control and manipulation	User control	Customizable	Customizable style, color, composition	User-controlled effects	User-controlled effects	Customizable	-
	Learning from imperfect data	Yes	Yes	Yes	No	Yes	Yes	Yes
	User control	High	Limited	Limited	Limited	No user control	Medium	Highest
	Fine-tuning efficiency	High	Medium	Medium	Medium	Not applicable	Low	High
Ethical and Accessibility	Bias and fairness	-	-	Potential biases, mitigation	Potential biases, mitigation	-	-	Potential biases
	Robustness to textual noise	High	Medium	Low	Medium	Low	Medium	Highest
	Open-source availability	Not open source	Open source	Potential future open source	Potential future open source	Not open source	Open source	-
	Accessibility	Medium	High	Low	Medium	High	High	Low
UX and Handling	UX	User-friendly	User-friendly	User-friendly, accessible	User-friendly, accessible	User-friendly, accessible	User-friendly, accessible	-
	Long text handling	Fair	Poor	Excellent	Good	Poor	Poor	Good
Technical Aspects	Model information	Developed by OpenAI	Deep AI	Google research	Developed by AI	Jasper	NightCafe	-
	Training process	Diffusion modeling	Generative pre-training	Generative pre-training	Diffusion modeling	Transformers	Deep learning	Diffusion modeling
	Training data size	1.5B images	1.5B images	400M images	1.5B images	400M images	400M images	100M images
	Model architecture	Transformer-based diffusion model	Transformer-based diffusion model	Transformer-based diffusion model	Diffusion model with a Wasserstein GAN discriminator	Transformer-based language model	Combination of AI algorithms	Transformer-based diffusion model
	Multilingual capabilities	Limited	Limited	Limited	Limited	No multilingual capabilities	Limited	Not available
	Visual storytelling	Supports visual storytelling	Supports visual storytelling	Supports visual storytelling	Supports visual storytelling	Supports visual storytelling	Supports visual storytelling	-
	Zero-shot generation	Yes	Yes	Yes	Yes	No	Yes	Yes

Technologies like DALL-E 2, Deep AI, Imagen, and others are evaluated for image realism, customization, ethics, UX, and technical aspects. Insights into their strengths and limitations offer guidance for their application in the evolving Imaging landscape.

—*Robustness to Textual Noise:* The ability of models to handle textual noise affects their practicality in real-world scenarios. DALL-E 2 demonstrates high robustness against textual noise, making it suitable for applications that involve noisy or imperfect input text. In contrast, models like Deep AI and Imagen have a moderate level of robustness, while Jasper and Stable Diffusion struggle more in this aspect. This robustness impacts the model's reliability when processing user-generated or imperfect text inputs.

— *Open Source Availability:* The availability of models as open source resources contributes to transparency and collaboration. Deep AI and NightCafe are open source, encouraging the community to contribute and enhance the technology. While Imagen and Stable Diffusion indicate potential future openness, DALL-E 2, Jasper, and MidJourney are not open source. Open source availability fosters innovation and empowers developers to create diverse metaverse applications.
— *Accessibility:* The accessibility of generative models influences their adoption by a wider user base. Deep AI, NightCafe, and Imagen offer high accessibility, ensuring that users with varying technical backgrounds can utilize the technology effectively. However, MidJourney has lower accessibility due to its complexity, potentially limiting its usage to users with advanced technical expertise. This impacts the inclusivity of the metaverse, as less technically skilled users may find it challenging to engage with MidJourney's capabilities.

(4) *UX and Handling:*
— *UX:* The UX of interacting with generative models greatly influences their usability. Models like DALL-E 2, Deep AI, Imagen, NightCafe, and MidJourney prioritize user-friendliness, ensuring that users can interact with the technology intuitively. A seamless and user-friendly experience contributes to higher user satisfaction and engagement.
— *Long Text Handling:* The ability to handle long texts is essential for applications that involve detailed or extensive input. Deep AI struggles with handling long texts, offering only poor performance. In contrast, Imagen excels in handling long texts, making it suitable for applications that require detailed input descriptions. NightCafe also faces challenges in this aspect, potentially limiting its effectiveness in scenarios involving lengthy prompts.

(5) *Technical Aspects:*
— *Model Information:* The source and origin of generative models provide insights into their development and credibility. DALL-E 2 is developed by OpenAI, indicating a reputable source and established research foundation. Deep AI and NightCafe's open source nature enhances their transparency and community collaboration. Imagen and Stable Diffusion show potential future open source availability, hinting at possible future developments.
— *Training Process:* Understanding the training process sheds light on the model's underlying mechanisms. DALL-E 2 employs diffusion modeling, Deep AI utilizes generative pre-training, and Imagen uses a similar generative pre-training approach. Jasper's approach combines diffusion modeling with a Wasserstein GAN discriminator. These insights provide technical depth and guide practitioners in selecting models aligned with their preferences.
— *Training Data Size:* The amount of training data impacts the model's capacity and performance. DALL-E 2 and MidJourney are trained on 1.5B images, enabling them to capture a diverse range of visual concepts. Deep AI and Imagen leverage 400M images for their training, still providing substantial data to learn from. NightCafe also uses 400M images, while Jasper's training size is 100M images. Larger training data sizes can enhance the model's ability to generalize to various inputs [4].
— *Model Architecture:* The architectural foundation of models influences their capabilities. Most models rely on Transformer-based diffusion models, known for their effectiveness in capturing complex patterns. Stable Diffusion, however, incorporates diffusion modeling with a Wasserstein GAN discriminator. This architecture combination enhances Stable Diffusion's generation process by integrating GAN-based adversarial learning.
— *Multilingual Capabilities:* Models' ability to operate across multiple languages impacts their versatility. DALL-E 2 and MidJourney have limited multilingual capabilities, which might

limit their usefulness in diverse linguistic contexts. On the other hand, NightCafe does not support multilingualism, potentially excluding users from non-supported languages.
— *Visual Storytelling:* Supporting visual storytelling enriches narrative-driven experiences in the metaverse. DALL-E 2, Deep AI, Imagen, NightCafe, and MidJourney offer capabilities for visual storytelling. This enhances immersive scenarios, allowing users to create interactive narratives within virtual environments.
— *Zero-Shot Generation:* The ability to generate content without specific training data is valuable for rapid content creation. DALL-E 2, Deep AI, Imagen, Stable Diffusion, and MidJourney are capable of zero-shot generation. This empowers users to quickly create content without the need for extensive training, accelerating the creative process in the metaverse.

7 Visualization Analysis and Inference

Following the theoretical comparison, this section delves into the practical comparison of visual images generated by various technologies. We conduct a comprehensive comparative analysis for four specific models: MidJourney, Stable Diffusion, NightCafe, and Craiyon [5].

Our objective is to scrutinize and evaluate the performance and artistic capabilities of each model by subjecting them to three distinct and challenging case scenarios. These scenarios have been thoughtfully selected to represent a wide spectrum of visual content, ensuring a well-rounded assessment of the models' capabilities. To evaluate their image generation abilities, we consider metrics such as image realism, quality, scalability, robustness, and semantic coherence. Additionally, we examine aspects of bias, fairness, and open source availability to ensure ethical transparency. This comprehensive approach provides insights into the strengths and weaknesses of each model in generating realistic images across diverse scenarios.

The three distinct case scenarios (Human Faces, Nature, Technology) were chosen for the analysis because they represent a broad spectrum of visual content.

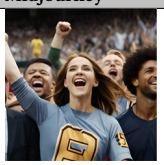
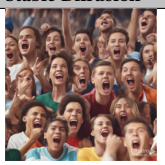
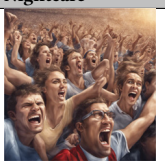
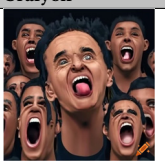
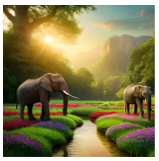
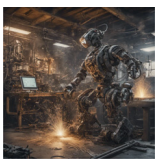
— The Human Faces scenario challenges models to generate realistic and nuanced depictions of people.
— The Nature scenario focuses on the model's ability to create diverse and intricate landscapes encompassing flora and fauna.
— The Technology scenario involves various man-made objects, including machinery and metallic elements, with precision.

These scenarios were selected to evaluate the models' capabilities in capturing different aspects of visual content, from human expressions to natural landscapes and technological environments.

The prompts guiding our evaluation in each case scenario are as follows:

(1) *Human Faces:* In this scenario, we challenge each model to generate a realistic image featuring 4-5 individual faces of people in a crowded stadium. The goal is to capture the varied expressions of individuals as they cheer and shout with enthusiasm during an event. This task demands a keen understanding of facial features, emotions, and crowd dynamics to produce compelling and authentic portraits.
(2) *Nature:* The natural world serves as an abundant source of inspiration, and our evaluation in this scenario revolves around the creation of a vivid and lifelike image of a lush jungle landscape. The image should depict a thriving ecosystem filled with joyful elephants and a rich variety of vibrant, colorful animals. This jungle is set amidst a dense forest adorned with flowering trees. The prompts will test the models' ability to replicate the intricate details and vibrant diversity found in the natural environment.

Table 3. This Presentation Table Offers a Vivid Comparative Analysis of Cutting-Edge GAI Technologies, Specifically Midjourney, Stable Diffusion, Nightcafe, and Craiyon

Theme	Prompt	Midjourney	Stable Diffusion	Nightcafe	Craiyon
Human faces	"Generate a realistic image featuring 4-5 individual faces of people in a crowded stadium, showcasing their expressions while cheering and shouting with enthusiasm."				
Nature	"Generate a vivid and realistic image of a lush jungle landscape teeming with joyful elephants and vibrant, colorful animals, set amidst a dense forest filled with flowering trees."				
Tech	"A robot inside a dimly lit, cluttered workshop, sparks flying from a welding operation, computer screens displaying schematics, the robot's body covered in welding residue."				

The comparison is situated within the context of three distinct image categories: human faces, nature, and technology. Each row presents a unique challenge to these AI systems, pushing them to craft highly realistic and creatively imaginative images. The resulting artwork, thoughtfully juxtaposed with their respective prompts, offers profound insights into the capabilities and artistic ingenuity of each AI model. It sheds light on their proficiency in interpreting and generating visual content across these diverse thematic domains. This analytical exploration serves as an invaluable resource for evaluating and comprehending the creative potential and performance of these cutting-edge GAI technologies.

(3) *Technology:* The ever-evolving landscape of technology presents a captivating subject for image synthesis. In this scenario, our prompts call upon the models to craft an image featuring an industrial robot inside a dimly lit, cluttered workshop. The scene should depict sparks flying from a welding operation, computer screens displaying intricate schematics, and the robot's metallic body covered in welding residue. This image should be a striking representation of the intersection between art and technology, showcasing the use of steel and aluminum in a sculptural context.

Through this meticulous comparative analysis, we aim to shed light not only on the strengths and weaknesses of each model but also their suitability for a diverse range of artistic and practical applications. By evaluating their performance in these challenging scenarios, we empower artists, developers, and researchers to make well-informed decisions when selecting an image synthesis technology that aligns with their specific creative or practical objectives.

The results are presented in Table 3. Based on analysis of Table 3, we have following inferences:

—*Human Faces:*
  —*Midjourney:* Midjourney consistently produced realistic images featuring 4-5 individuals in a crowded stadium. The expressions of the people in the images were vivid and conveyed enthusiasm.
  —*Stable Diffusion:* Stable Diffusion generated images with decent facial details, although some faces appeared slightly distorted. The overall quality was acceptable.
  —*Nightcafe:* Nightcafe yielded images where facial expressions were captured well, but there were occasional issues with the rendering of hands.

— *Craiyon:* Craiyon exhibited the poorest performance, failing to produce realistic faces or backgrounds.
— *Nature:*
— *Midjourney:* Midjourney generated images that, while realistic, lacked creativity and intricacy in depicting the lush jungle landscape and vibrant animals.
— *Stable Diffusion:* Stable Diffusion excelled in adding intricate details to the jungle scene, creating a vivid and realistic depiction with vibrant colors.
— *Nightcafe:* Nightcafe particularly excelled in rendering skin tones and overall image composition, creating a visually appealing result.
— *Craiyon:* Craiyon produced images with a more cartoonish and less realistic appearance, especially in its portrayal of elephants.
— *Technology:*
— *Midjourney:* Midjourney simplified the complex elements of the technology-themed prompt, resulting in images with less intricate features.
— *Stable Diffusion:* Stable Diffusion added adequate details, but the overall meaning and coherence of the images may have been lacking.
— *Nightcafe:* Nightcafe effectively covered all aspects of the prompt, but the images might not have conveyed a clear message or purpose.
— *Craiyon:* Craiyon lagged behind its competitors, failing to generate images that met the standards of the prompt's complexity and meaning.

8 Challenges to Integrating GAI in Imaging

The challenges associated with including GAI in imaging are multifaceted and require a comprehensive approach. Table 4 outlines these challenges (highlighted in other works such as [10, 11, 19]), potential solution approaches, and areas for future work.

— *Ethical Concerns:*
— *Misuse:* The generation of deceptive deepfakes using GAI is a major ethical concern. To combat this, solutions like developing deep neural network-based content authenticity classifiers have been proposed. These classifiers aim to detect deepfakes, ensuring that manipulated images and videos can be identified. However, future work suggests advancing deepfake detection further by using GAN-based counter-GANs. This would involve creating more robust and effective mechanisms for identifying and countering deepfake content.
— *Bias and Fairness:* Addressing bias and fairness in GAI models is crucial. Adversarial training is one solution to reduce dataset bias during model training. It helps ensure that AI-generated content is fair and unbiased. Future work involves exploring differentiable fairness constraints for generative models. This approach aims to make the process of mitigating bias more integral to the training process.
— *Privacy and Consent:*
— *Privacy Violations:* GAI can inadvertently lead to privacy violations. Federated learning approaches have been suggested to mitigate this concern. These approaches keep data decentralized and private, ensuring that sensitive information is not exposed. Future work focuses on developing homomorphic encryption methods for even more secure data sharing, enhancing data privacy in imaging applications.
— *Consent Challenges:* Ensuring consent for data usage is essential. Cryptographic image watermarking is a solution to verify the origin of content, providing evidence that consent was obtained. Additionally, ongoing research into decentralized identity and blockchain-based consent systems is suggested to enhance consent management.

Table 4. Provides a Comprehensive Overview of the Multifaceted Challenges Encountered in the Field of GAI for Imaging

Category	Subcategory	Solution Approach	Future Work
Ethical Concerns	Misuse	Develop deep neural network-based content authenticity classifiers to detect deepfakes.	Advance deepfake detection using GAN-based counter-GANs.
	Bias and Fairness	Implement adversarial training to reduce dataset bias during model training.	Explore differentiable fairness constraints for generative models.
Privacy and Consent	Privacy Violations	Implement federated learning approaches to keep data decentralized and private.	Develop homomorphic encryption methods for secure data sharing.
	Consent Challenges	Create cryptographic image watermarking for content origin verification.	Research decentralized identity and blockchain-based consent systems.
Security Threats	Authentication Issues	Implement anti-spoofing techniques like liveness detection in facial recognition systems.	Explore AI-resistant biometric authentication methods.
	Cybersecurity	Employ differential privacy mechanisms for AI model training to protect against model inversion attacks.	Investigate AI model security in federated learning settings.
Legal and Regulatory Challenges	Intellectual Property	Develop AI model licensing frameworks and smart contracts for ownership verification.	Create decentralized copyright registries using blockchain.
	Regulation	Establish regulatory sandboxes for AI research, development, and deployment.	Collaborate internationally to standardize AI governance protocols.
Data Privacy and Security	Data Leakage	Utilize federated learning and secure aggregation techniques to prevent data leakage.	Enhance SMPC for data privacy.
	Security Risks	Implement model encryption and secure model serving frameworks to safeguard AI models.	Research AI model watermarking techniques for model validation.
Technological Challenges	Resource Intensity	Optimize model architectures with hardware accelerators like TPUs and GPUs.	Investigate low-power and energy-efficient model architectures.
	Scalability	Design distributed GAI frameworks to efficiently scale for real-time deployment.	Explore edge AI deployment for on-device image generation.
Verification and Trust	Deepfakes Detection	Develop multi-modal deepfake detection systems utilizing audio and video analysis.	Investigate explainable AI techniques for deepfake detection.
	Authentication Difficulty	Research image hashing and blockchain-based image authenticity verification systems.	Explore zero-knowledge proofs for image origin verification.
Potential for Harm in Healthcare	Misdiagnosis	Develop AI-guided diagnosis validation tools with rigorous clinical testing.	Investigate federated learning for distributed healthcare diagnostics.
Cultural and Social Impacts	Cultural Sensitivity	Implement AI content moderation tools that recognize cultural context and nuances.	Develop AI systems with cultural and ethical reasoning capabilities.
	Social Isolation	Promote collaborative AI-human creative endeavors and interdisciplinary research.	Investigate AI-human symbiotic systems for content creation.

It not only identifies these challenges but also offers practical solutions aimed at addressing them. Furthermore, it outlines promising areas for future research and advancement within the field. The table serves as a valuable resource for researchers, developers, and policymakers seeking to navigate the complex landscape of ethical, privacy, security, and technological considerations associated with GAI in the context of image generation and manipulation. GPUs, graphics processing units; SMPC, secure multi-party computation; TPUs, tensor processing units.

- *Security Threats:*
  - *Authentication Issues:* GAI can face authentication challenges, especially in facial recognition systems. Anti-spoofing techniques like liveness detection are employed to enhance security. Future work explores AI-resistant biometric authentication methods, aiming to make authentication more robust and secure against spoofing attacks.
  - *Cybersecurity:* Cybersecurity is a paramount concern in the deployment of GAI models. Differential privacy mechanisms are employed to protect against model inversion attacks during AI model training. Future research delves into AI model security in federated learning settings, ensuring that models remain secure and uncompromised.
- *Legal and Regulatory Challenges:*
  - *Intellectual Property:* Intellectual property concerns can be mitigated through AI model licensing frameworks and smart contracts. These mechanisms help establish ownership and usage rights. Future work suggests creating decentralized copyright registries using blockchain, further strengthening intellectual property protection.
  - *Regulation:* Regulatory sandboxes are established for AI research, development, and deployment. These sandboxes allow controlled testing and innovation. Collaborative international efforts are recommended to standardize AI governance protocols, ensuring consistency in regulations across borders.
- *Data Privacy and Security:*
  - *Data Leakage:* Data leakage is a significant concern in data-driven applications. Solutions like federated learning and secure aggregation techniques are used to prevent data leakage. Enhancing secure multi-party computationis a future area of research to further safeguard data privacy [20, 25].
  - *Security Risks:* Ensuring the security of AI models is vital. Model encryption and secure model serving frameworks are implemented to protect AI models. Ongoing research into AI model watermarking techniques is proposed, enhancing model validation and security.
- *Technological Challenges:*
  - *Resource Intensity:* Optimizing model architectures with hardware accelerators like **tensor processing units (TPUs)** and **graphics processing units (GPUs)** helps address resource intensity challenges. Exploring low-power and energy-efficient model architectures is recommended to reduce resource consumption.
  - *Scalability:* Scalability is essential for real-time deployment. Designing distributed GAI frameworks enables efficient scaling. Additionally, exploring edge AI deployment for on-device image generation can further enhance scalability.
- *Verification and Trust:*
  - *Deepfakes Detection:* Multi-modal deepfake detection systems are developed to analyze audio and video in addition to images. This comprehensive approach helps in detecting more sophisticated deepfakes. Future work explores explainable AI techniques to provide transparency in deepfake detection.
  - *Authentication Difficulty:* Image hashing and blockchain-based image authenticity verification systems are utilized for authentication. Investigating zero-knowledge proofs is suggested to enhance image origin verification and trust in the authenticity of images.
- *Potential for Harm in Healthcare:*
  - *Misdiagnosis:* Rigorous clinical testing and AI-guided diagnosis validation tools are developed to mitigate the potential for harm in healthcare applications. Future research explores federated learning to enhance patient safety by ensuring distributed and secure healthcare diagnostics.

— *Cultural and Social Impacts:*
  — *Cultural Sensitivity:* AI content moderation tools that recognize cultural context and nuances are implemented to address cultural and social impacts. Additionally, future work focuses on developing AI systems with cultural and ethical reasoning capabilities to ensure responsible content creation that respects cultural diversity.
  — *Social Isolation:* Promoting collaborative AI-human creative endeavors and interdisciplinary research is recommended to combat social isolation. Investigating AI-human symbiotic systems for content creation aims to foster collaboration and reduce the isolation potential associated with AI-generated content.

In summary, including GAI in imaging poses numerous challenges across ethical, privacy, security, legal, and technological domains. Addressing these challenges requires a multi-pronged approach involving technical solutions, ongoing research, and collaboration across disciplines and borders. The future work outlined in the table highlights the need for continuous innovation and development to ensure the responsible and effective use of GAI in imaging [4, 18].

9 Future Work

The challenges and risks surrounding GAI in multimedia pave the way for a rich landscape of future work and research. As technology evolves, it is essential to remain proactive in addressing these issues and driving innovation in the field. Future improvements for GAI in multimedia include:

— *Understanding User Preferences:* Developing GAI models that adapt to user preferences in real-time could revolutionize how multimedia content is delivered. Personalized and tailored content ensures that every user receives exactly what they want, optimizing their engagement and satisfaction.
— *Multilingual and Multicultural Content Generation:* Future projects should focus on developing generative models that consider cultural peculiarities when generating content. This approach would allow content creators to address a global audience with due regard for cultural differences, ensuring positive reception worldwide.
— *AI-Assisted Storytelling:* GAI tools can help multimedia developers weave their stories by creating multiple concept art sketches based on high-level story descriptions or proposing visually striking multimedia that match the emotional tone of a narrative. This fosters agility and creativity in the work process.
— *Human-in-the-Loop Content Curation:* Incorporating ethical considerations into GAI models allows human creators to refine AI-generated content, ensuring it is responsible and unbiased. This man–machine collaboration can produce value-based, non-biased outputs.
— *Explainable AI for Multimedia Content:* Developing explainable AI techniques specifically designed for the multimedia domain would allow users to understand the rationale behind AI-generated content, fostering trust and transparency.
— *Combating Deepfakes and Misinformation:* Research insights can inform the development of robust deepfake detection methods tailored to the multimedia landscape. This empowers users to critically evaluate multimedia content and mitigate the spread of misinformation.
— *Combating Social Isolation:* Promoting collaborative AI-human creative endeavors and interdisciplinary research can help address social isolation.
— *Technological Advancements:* Future work should focus on improving multimedia realism, quality, scalability, robustness, and semantic coherence. Exploring the amalgamation of different GAI models (VAEs, GANs, Transformers, Autoregressive Models) can drive innovation. Advancing data compression techniques will enable efficient storage and transmission of multimedia content.

10 Future Improvements for GAI in Imaging Technology

Image synthesis technology can revolutionize many fields. Here are some key domains it can impact

- —*Medical Imaging:* Image synthesis can enhance early disease detection by analyzing medical scans, such as mammograms, to identify subtle anomalies, leading to improved patient outcomes. It also allows doctors to simulate treatment effects on a patient's condition, enabling personalized treatment plans and improving surgical results through AI-generated 3D models of organs and tissues.
- —*Materials Science:* AI can enable virtual testing and optimization of new material properties before physical creation, accelerating innovation. It also plays a crucial role in simulating material degradation over time, facilitating non-destructive testing, preventative maintenance, and infrastructure safety improvements.
- —*Entertainment:* In movies and video games, AI can generate fantastical creatures and replicate historical settings, enhancing realism and immersion.
- —*Fashion Design:* AI can create virtual clothing prototypes, allowing designers to experiment with different styles and iterate quickly, streamlining the design process.
- —*Space Exploration:* AI can improve planetary imaging by analyzing telescope data to remove noise and enhance images of distant planets, revealing crucial atmospheric details and the potential for life. Additionally, it can generate simulations of Martian landscapes, aiding in mission planning and astronaut training.

11 Conclusion

Historically, researchers have always been intrigued by the intersection of technology and arts in the multimedia domain. However, GAI opens a new chapter of the narrative that is fundamentally changing how people produce multimedia, and how human beings co-create together. The emergence of GAI has revolutionized image generation, pushing the boundaries of creativity and collaboration. This technology has not only amplified human creativity but also streamlined content creation processes, fostering new synergies between AI systems and human creators. The article meticulously explores state-of-the-art GAI models, including DALL-E 2, Craiyon, Stable Diffusion, Imagen, Jasper, NightCafe, and Deep AI, assessing their prowess in image quality, diversity, interpretability, and computational efficiency. The fusion of AI and image generation ventures into the realm of text-to-image synthesis, where AI transforms textual input into captivating visual representations. However, this journey also unveils a tapestry of ethical complexities and challenges, particularly concerning the equilibrium between creative autonomy and AI-driven automation. The article adeptly navigates these intricacies, providing profound insights into the evolving landscape of image creation. This research not only serves as an enlightening exploration of GAI's potential in image synthesis but also as a guiding light, illuminating the path to uncharted territories in image generation. It underscores the dynamic interplay between human ingenuity and AI, propelling the future of image creation towards boundless horizons. As GAI continues to evolve, it promises to enrich the creative landscape further, pushing the boundaries of visual content generation and shaping the future of artistic expression and automation.

References

[1] Ioannis D. Apostolopoulos, Nikolaos D. Papathanasiou, Dimitris J. Apostolopoulos, and George S. Panayiotakis. 2022. Applications of generative adversarial networks (GANs) in positron emission tomography (PET) imaging: A review. *European Journal of Nuclear Medicine and Molecular Imaging* 49, 11 (2022), 3717–3739.

[2] Jayme Garcia Arnal Barbedo. 2013. Digital image processing techniques for detecting, quantifying and classifying plant diseases. *SpringerPlus* 2, 1 (2013), 1–12.

[3] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society,* 396–410.
[4] Eva Cetinic and James She. 2022. Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–22.
[5] Yick Hin Edwin Chan and A. Benjamin Spaeth. 2020. Architectural visualisation with conditional generative adversarial networks (cGAN). In *Proceedings of the 38th eCAADe Conference,* 299–308.
[6] Abbas Cheddad, Joan Condell, Kevin Curran, and Paul Mc Kevitt. 2010. Digital image steganography: Survey and analysis of current methods. *Signal Processing* 90, 3 (2010), 727–752.
[7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.
[8] Zhineng Chen, Shanshan Ai, and Caiyan Jia. 2019. Structure-aware deep learning for product image classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1s (2019), 1–20.
[9] Giannis Daras and Alexandros G. Dimakis. 2022. Discovering the hidden vocabulary of DALLE-2. Retrieved from https://doi.org/10.48550/arXiv.2206.00169
[10] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R. Frank, Matthew Groh, Laura Herman, Neil Leach, Robert Mahari, Alex Sandy Pentland, Olga Russakovsky, Hope Schroeder, and Amy Smith. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.
[11] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research* 25, 3 (2023), 277–304.
[12] Athanasios Karapantelakis, Pegah Alizadeh, Abdulrahman Alabassi, Kaushik Dey, and Alexandros Nikou. 2024. Generative AI in mobile networks: A survey. *Annals of Telecommunications* 79 (2024), 15–33.
[13] Mohamad Koohi-Moghadam and Kyongtae Ty Bae. 2023. Generative AI in medical imaging: Applications, challenges, and ethics. *Journal of Medical Systems* 47, 1 (2023), 94.
[14] Kazuhiro Koshino, Rudolf A. Werner, Martin G. Pomper, Ralph A. Bundschuh, Fujio Toriumi, Takahiro Higuchi, and Steven P. Rowe. 2021. Narrative review of generative adversarial networks in medical and molecular imaging. *Annals of Translational Medicine* 9, 9 (2021), 1–15.
[15] Maria Elena Laino, Pierandrea Cancian, Letterio Salvatore Politi, Matteo Giovanni Della Porta, Luca Saba, and Victor Savevski. 2022. Generative adversarial networks in brain imaging: A narrative review. *Journal of Imaging* 8, 4 (2022), 83.
[16] Xiang Li, Yuchen Jiang, Juan J. Rodriguez-Andina, Hao Luo, Shen Yin, and Okyay Kaynak. 2021. When medical images meet generative adversarial network: Recent development and research opportunities. *Discover Artificial Intelligence* 1 (2021), 1–20.
[17] Bahar Mahmud, Guan Hong, and Bernard Fong. 2023. A study of human–AI symbiosis for creative work: Recent developments and future directions in deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2 (2023), 1–21.
[18] Weizhi Nie, Weijie Wang, Anan Liu, Jie Nie, and Yuting Su. 2019. HGAN: Holistic generative adversarial networks for two-dimensional image-based three-dimensional object retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 4 (2019), 1–24.
[19] M. R. Pavan Kumar and Prabhu Jayagopal. 2021. Generative adversarial networks: A survey on applications and challenges. *International Journal of Multimedia Information Retrieval* 10, 1 (2021), 1–24.
[20] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1 (2019), 1–24.
[21] Kaiqi Qiu, Feiru Wang, and Yingxi Tang. 2022. Machine learning approach on AI painter: Chinese traditional painting classification and creation. In *Proceedings of the International Conference on Cultural Heritage and New Technologies*, 1–6.
[22] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.
[23] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18359–18369.
[24] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 22428–22437.

[25] Yi Yu, Abhishek Srivastava, and Simon Canales. 2021. Conditional LSTM-GAN for melody generation from Lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 1 (2021), 1–20.

[26] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 5907–5915.

Received 15 February 2024; revised 11 June 2024; accepted 13 August 2024