Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/95/2024BJ0056

Evolution and future directions of Artificial Intelligence Generated Content (AIGC): A comprehensive review

Yihan Xu

School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Huhan, China

yhxu022@hust.edu.cn

Abstract. Artificial Intelligence Generated Content (AIGC) has rapidly evolved, revolutionizing the creation of text, images, audio, and video content. Despite these advancements, research on the development process of AIGC technology remains scarce, necessitating a systematic discussion of its current state and future directions. So this paper delves into the significant advancements and foundational technologies driving AIGC, emphasizing the contributions of state-of-the-art models such as DALL-E 3 [1] and Sora [2]. We analyze the evolution of generative models from single-modal approaches to the current multimodal generative models. The paper further explores the application prospects of AIGC across various domains such as office work, art, education, and film, while addressing the existing limitations and challenges in the field. We propose potential improvement directions, including more efficient model architectures and enhanced multimodal capabilities. Emphasis is placed on the environmental impact of AIGC technologies and the need for sustainable practices. Our comprehensive review aims to provide researchers and professionals with a deeper understanding of AIGC, inspiring further exploration and innovation in this transformative domain.

Keywords: Artificial Intelligence Generated Content, Neural Networks, Multimodal Generative Models, Large AI Models.

1. Introduction

Artificial Intelligence Generated Content refers to the use of generative AI technologies to automatically create text, images, audio, or video content. In recent years, AIGC has advanced rapidly, with numerous exceptional models emerging, such as Llama 3 [3] and Sora. OpenAI's DALL-E 3 can produce high-quality images from simple text descriptions, while the Sora model, released in February 2024, can generate high-quality videos based on user text inputs, significantly streamlining the video creation process. Sora represents a significant breakthrough in generative AI models, poised to disrupt related industries and profoundly impact academia and society.

The rapid development of AIGC is largely due to the advancement of algorithms for large models, making AIGC products promising tools that enhance our lives. In recent years, multimodal technology has seen substantial progress, enabling multimodal generative models to process and produce information across various data types, achieving cross-modal content generation. Examples include DALL-E 3, Parti [4], and Vidu [5]. The Vidu model, released on April 27 2024, can generate high-

DOI: 10.54254/2755-2721/95/2024BJ0056

definition videos up to 16 seconds long with a resolution of 1080P. The iteration of AIGC products and models is extremely rapid, with their capabilities and generalization continuously improving.

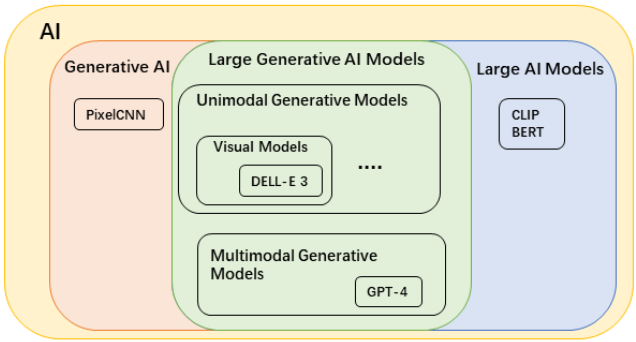

Figure 1: Relation between existing representative large AI models and AIGC.

Technological advancements and foundational hardware upgrades in recent years have provided significant momentum for the AIGC field. Progress in large model [6,7] has allowed for the scaling up of models, enhancing both parameters and performance limits. Larger datasets [8–11] have highlighted the importance of data engineering, significantly improving data quality and quantity. Hardware advancements have enabled academia to build more powerful data and computing centers, supporting the training of large models.

The main contributions of this paper are as follows:

- This paper provides a detailed overview of the foundational technologies of image generation models in generative AI, including the principles of underlying model architectures and the integration of modal generation. It offers a comprehensive analysis of visual AIGC tasks and models from both single-modal and multi-modal perspectives.
- This paper reviews and analyzes the recent advancements and current development status of AIGC technology, discussing its application prospects.
- This paper focuses on the existing limitations of AIGC technology and corresponding improvement directions. We discuss the challenges faced by the AIGC field, the potential social impacts, and propose possible solutions and feasible directions.

The remainder of this paper is structured as follows: The first section introduces the history and foundational models of single-modal generation models. The second section summarizes the characteristics and development of multimodal generative models. The third section presents the development status and applications of AIGC technology. The fourth section highlights the current limitations, challenges, and potential solutions and feasible directions of AIGC technology. Finally, we conclude with a summary of our research findings.

2. Vision Generative Models

In this section, we focus on visual unimodal generative models, which accept single-modal inputs such as image-level noise [12] and specific noise sequences to generate image data.

The history of generative models dates back to the Gaussian Mixture Models (GMMs) [13] of the 20th century. GMMs generate specific data distributions by combining multiple Gaussian distributions. Traditionally, there was limited overlap between different fields of AIGC. In the field of NLP, N-gram language models [14] were used to learn word distributions, while in the CV field, algorithms like texture synthesis [15] and image editing [16] were commonly used, which limited diversity in generation.

With the advent of neural network algorithms [17], Recurrent Neural Networks (RNN) [18] and subsequent improvements like Long Short-Term Memory (LSTM) [19] and Gated Recurrent Units (GRU) [20] were introduced into language modeling tasks with notable success. Variational Autoencoders (VAE) [21] provided finer control for image generation in the CV field. The Generative

DOI: 10.54254/2755-2721/95/2024BJ0056

Adversarial Network (GAN) [22], proposed in 2014, marked a significant milestone in image generation, as GANs can generate high-quality, photorealistic images through adversarial training between the

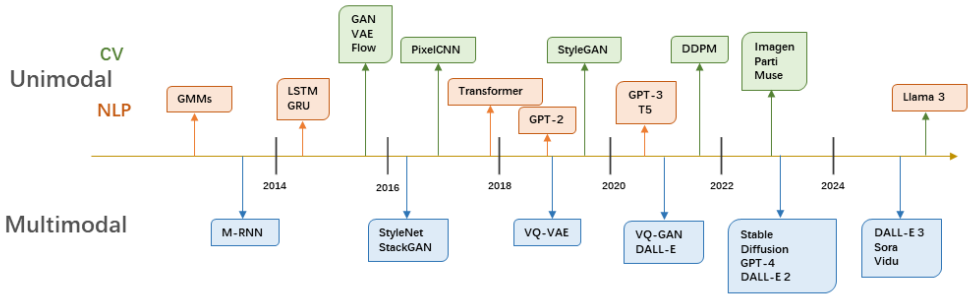

Figure 2: The history of Generative AI in unimodal and multimodal.

generator and discriminator. Diffusion models [23], introduced in 2020, add noise to real images gradually using Markov chains, and then learn Markov transition kernels to achieve reverse diffusion. These models have become a mainstream paradigm in AIGC. This section will mainly cover four types of single-modal generative models: VAE, GAN, Autoregressive models, and Diffusion models.

VAE. VAE, a generative network structure based on variational Bayesian inference, was proposed in 2014. The traditional VAE consists of an encoder and a decoder. After training, the encoder transforms the input X into a latent vector Z, while the decoder samples points z from the latent variable distribution to reconstruct the original input.

The core idea of VAE is to treat the encoder's output as a probability distribution, i.e., the latent vector. The decoder reconstructs the original input from this probability distribution. The latent vector represents the probability distribution of the latent features of the image, and the decoder samples points from the latent vector space using reparameterization:

$$z = \mu_\phi(x) + \sigma_\phi(x) \cdot \epsilon \tag{1}$$

where $\mu_\phi(x)$, $\sigma_\phi(x)$represent the mean and standard deviation of the input data, respectively, while ϵ denotes noise from a standard normal distribution.

To constrain the latent vector distribution, VAE's optimization objective comprises reconstruction loss and KL divergence [24]:

$$\mathscr{L}(\theta, \phi; x) = -E_{q(z|x)}[log\, p\,(x|z)] + KL\big(q(z|x)\,|\,p(z)\big) \tag{2}$$

where $\mathscr{L}$ is the loss function, ϕ, θ are the parameters of the encoder and decoder. $p(x|z)$represents the likelihood of the data given the latent variable z, $q(z|x)$ denotes the conditional distribution of the latent variable z, and $p(z)$ is the prior distribution.

VAE provides a strong theoretical foundation and model basis for subsequent image generation work, leading to numerous improvements. ELBO improves the variational bounds, balancing reconstruction error and latent space distribution differences, enabling the model to generate new samples similar to real data. Wasserstein VAE [25] uses the Wasserstein distance instead of KL divergence to enhance training stability. Beta-VAE [26] introduces an adjustable hyperparameter β to control the weight of the KL divergence term, improving the model's disentangling ability and interpretability. VQVAE [27] discretizes the latent variables, generating a set of discrete codebooks, and combines these codebooks to create a prior, resulting in the final image. VQVAE2 [28] builds on this by layering the encoder and decoder, modeling different levels of features separately, producing highly impressive results.

GAN. Since its introduction in 2014, GANs have achieved remarkable success in image generation. The core concept of GANs involves generating realistic images through the adversarial interaction of two neural networks, the generator and the discriminator. The generator learns to produce data that

DOI: 10.54254/2755-2721/95/2024BJ0056

mimics the real data distribution, while the discriminator learns to distinguish whether the input data is real or generated.

The vanilla GAN's optimization objective is defined as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[log\, D(x)] + E_{z\sim p_z(z)}\left[log\left(1 - D\big(G(z)\big)\right)\right] \quad (3)$$

where G represents the generator, D represents the discriminator, p_{data} is the real data distribution, and $p_z(z)$ is the input distribution.

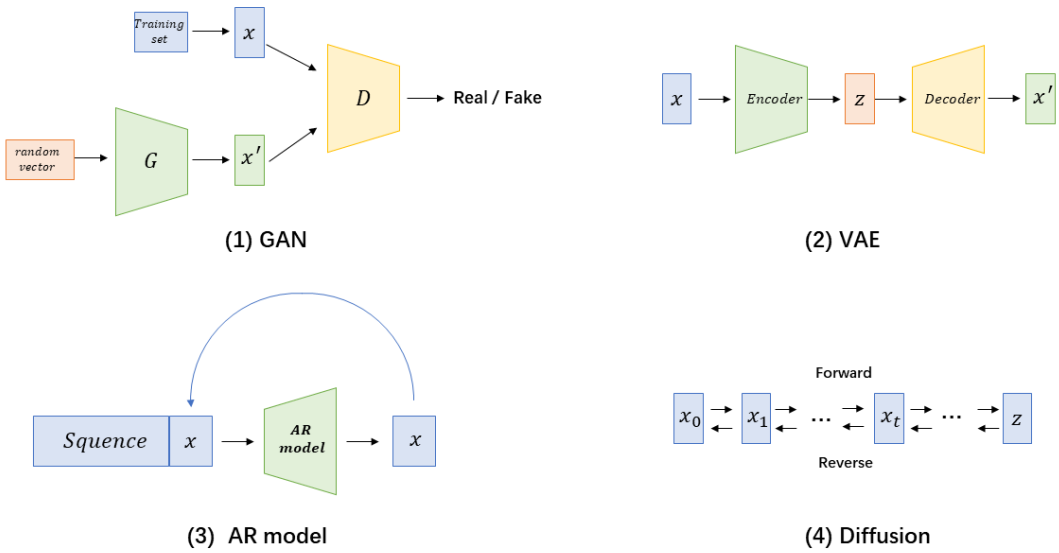

Figure 3: A brief overview of vision generative models.

In fact, this optimization creates a zero-sum game (minimax game) between the two networks. Theoretically, when the generator and discriminator reach a Nash equilibrium, the generated data will become indistinguishable from real data by the discriminator, meaning the generated data distribution closely approximates the real data distribution, significantly enhancing the realism of the generated images.

With deeper research into GANs, several training challenges were identified, such as mode collapse and training instability. To address these, Mirza and Osindero introduced Conditional GAN [29], which add conditional information into the generator and discriminator inputs, enabling the generation of data with specific attributes. Radford et al. introduced Deep Convolutional GAN(DCGAN) [30], which employ a fully convolutional architecture to improve the quality of generated images. Arjovsky et al. proposed the Wasserstein GAN [31], using the Wasserstein distance as the optimization objective and weight clipping to maintain the Lipschitz constraint, thereby improving training stability. Zhu et al. developed CycleGAN [32], which incorporates cycle consistency to allow the input image to be reconstructed back to the original without paired training samples, making it highly effective for style transfer tasks. Zhang et al. introduced the Self-Attention GAN (SAGAN) [33], incorporating a self-attention mechanism to improve the model's ability to capture long-range dependencies, enhancing both the overall coherence and fine details of generated images.

AR models. Autoregressive models are widely used in time series analysis [34]. They predict the next element in a sequence by modeling the interdependencies among the elements of the sequence. In the field of image generation, a notable work using autoregressive models is PixelRNN [35], proposed in 2016. PixelRNN treats pixels as a sequential series, where each pixel depends on the preceding ones, allowing for pixel-by-pixel prediction using a recursive method. The model employs the chain rule to decompose the likelihood of a data sample x into a product of one-dimensional distributions, and then uses an LSTM-based model to learn and predict these distributions. PixelCNN [36] uses masked

DOI: 10.54254/2755-2721/95/2024BJ0056

convolution, efficiently learning pixel distributions through the sliding of convolution kernels and the application of masks, thus generating images. GatedPixelCNN incorporates LSTM gate mechanisms into the convolution process and decomposes masked convolutions to address the receptive field blind spot problem. VQGAN [37] exemplifies a combination of autoregressive models, GAN, and VAE. It generates discrete sequences by dividing images into patches and encoding them into a codebook. The decoder then reconstructs the codebook sequences. Simultaneously, VQGAN trains an autoregressive sequence prediction model to predict the codebook sequences autoregressively during inference, ultimately generating a complete image.

Diffusion. Diffusion models have recently achieved breakthrough progress in the field of image generation. The core idea of these models is to gradually add noise to data and then learn how to reverse this process to generate data, particularly images. Diffusion models have become a mainstream approach in image generation. These models are based on a progressive process where data samples are gradually transformed by adding noise until they become pure noise. This process can be viewed as a Markov chain that starts from the data distribution and transitions to a pure noise distribution.

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{4}$$

where β_t denotes the noise level, $\mathcal{N}$ represents the Gaussian distribution.

During training, the generative model learns to reconstruct data from noise, effectively reversing the Markov chain process:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I\right) \tag{5}$$

Diffusion models learn across multiple time steps how to reconstruct data from a noisy state, contrasting with traditional GANs or VAEs. A key advantage of diffusion models is their ability to maintain high diversity and consistency in generated images, thereby avoiding the mode collapse issue commonly seen in GANs.

DDIM [38] represents a significant improvement by reducing sampling steps to accelerate generation while ensuring high-quality content. This advancement facilitated the creation of the notable AIGC open-source project, Stable Diffusion [39]. Ho et al. [40] enhanced the quality and detail of generated images by adjusting parameters directly during training and sampling, improving performance under unsupervised conditions. Rombach et al. [39] make process occur in a low-dimensional latent space, and a decoder converts these representations back into high-resolution images, significantly improving image resolution and visual quality. Song et al. [41] employ stochastic differential equations to simulate a continuous diffusion process, offering a theoretically rigorous and flexible method to handle complex data distributions, further enhancing the generative capability and application range of the model.

These advancements not only enhance the functionality and efficiency of diffusion models but also expand their potential applications in high-quality image generation, demonstrating their broad applicability and superior performance in various image and data generation tasks.

Table 1. Multimodal models and their supported modalities.

Modality	Input				Output			
	text	audio	image	video	text	audio	image	video
GPT-3, llama	✓				✓			
wav2vec-U		✓			✓			
Imagen, Dall-E	✓						✓	
Flamingo	✓		✓	✓	✓		✓	
Sora	✓	✓	✓	✓	✓	✓	✓	✓
GPT-4	✓	✓	✓	✓	✓		✓	

DOI: 10.54254/2755-2721/95/2024BJ0056

3. Multimodal Generative Models

Multimodal generation represents a crucial component and development trend within the AIGC field. The powerful foundational models and advancements previously discussed offer a solid basis for further multimodal research. The introduction of attention mechanisms [42] and Transformers [43] has created new possibilities for handling complex multimodal data and integrating various modalities. In computer vision, multimodal generative models primarily rely on GAN and Diffusion models, integrating new modalities and establishing connections between them to achieve multimodal generation. Recent works can generally be categorized into three types: GAN-based, VAE-AR-based, and Diffusion-based models.

GAN-based Models: Early generative models predominantly utilized GAN architectures, resulting in significant progress in text-to-image generation. StackGAN [44] employs a two-stage GAN structure: the first stage generates the primary structure and shape of objects from text, while the second stage refines this into high-resolution images. In 2018, Tao Xu et al. [45] uses an attention generative network

and a deep attention multimodal similarity model (DAMSM) to synthesize images from text descriptions through multiple stages. GigaGAN [46] explores the performance of large-scale GANs on extensive datasets, highlighting GANs' advantages in generation speed and high-resolution image production.

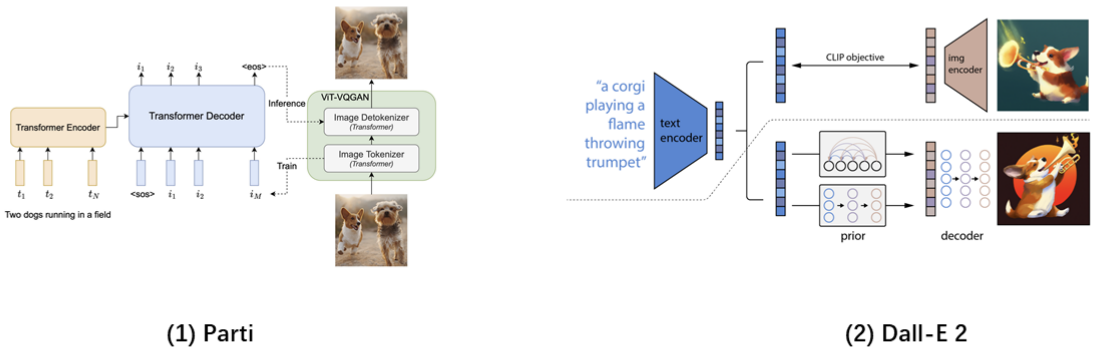

Figure 3: The structure of Parti (**GAN-based**) and Dall-E 2 (**Diffusion-based**).

VAE-AR-based Models: ViT-VQGAN combines the ViT [47] encoder with the VQGAN structure, resulting in an improved codebook and quantization process. Parti [4] represents a large-scale implementation of ViT-VQGAN, incorporating a large NLP model text encoder for processing text inputs. Muse [48], another VQGAN variant, uses a fully CNN-based VQencoder-VQdecoder structure for image input, accommodating variable resolution inputs and employing an external super-resolution Transformer to produce high-resolution images. Both models exemplify text-to-image generation with notable performance and generalization capabilities. DALL-E [49] integrates dVAE [27], Transformer, and CLIP [50] models, each trained independently, resulting in a logically coherent model capable of understanding semantic information clearly.

Diffusion-based Models: Diffusion models have rapidly gained prominence in image generation, with numerous efforts to incorporate text modalities. The GLIDE [51] model introduces an ablated diffusion model for the generation process, leveraging larger models and paired datasets for training. Imagen [52] utilizes the T5 [53] model as a text encoder and employs two diffusion models for generating low and high-resolution images, effectively handling text generation and super-resolution tasks. DALL-E 2, built on a diffusion-based framework derived from GLIDE, maps image descriptions from the CLIP text encoder to the representation space, followed by diffusion priors mapping to the CLIP image encoding, and finally, the GLIDE generative model uses reverse diffusion to produce images. DALL-E 3 enhances this process by integrating GPT [54–57], improving the precision of prompt-to-image conversion.

DOI: 10.54254/2755-2721/95/2024BJ0056

4. Applications

AIGC has made significant technical advancements and continues to progress. At the same time, AIGC technology demonstrates extensive application potential in various fields such as office work, art, film and television, and education. Despite some controversies over its use, the immense convenience provided by AIGC makes it hard to resist. The deep integration of AIGC into everyday life is imminent.

Office Work: The development and application of AIGC in office environments are enhancing productivity and efficiency in numerous ways. Microsoft has introduced AI-driven tools like Copilot [58] in Windows and Office applications, leveraging GPT-4 to assist with data analysis, content generation, and even data predictions using web and internal information [59]. Companies like Synthesia are developing tools for creating deepfake avatars for training and marketing purposes, simplifying the production of high-quality video content [60].

Art: AIGC's application in art is widespread, enabling creators to produce complex artworks at unprecedented speeds. DALLE-2 [61] has helped numerous users create visual stories, illustrations, animations, and digital content, making it an essential tool for visualizing concepts, designing unique works, and even creating entire art collections. HitPaw's 4AiPaw allows users to experiment with different styles and prompts, facilitating various artistic expressions through a single text input. DreamStudio, developed by Stability.ai, uses diffusion models to generate high-resolution images from text prompts, becoming a popular choice among users. Palette [62] incorporates AI coloring technology, effectively replacing traditional manual coloring processes and significantly enhancing drawing and artistic creation efficiency.

Education: In education, AIGC can enhance learning experiences and materials. Teachers can use the DALL-E series models [1,49,61] to generate visual representations of historical events or complex scientific processes, aiding students in visual learning. These models can also generate custom images for textbooks and presentations, thereby concretizing abstract concepts and improving students' understanding and memory of classroom content. In the field of mathematical tasks, Google Research has introduced Minerva [63], capable of solving academic quantitative problems in algebra, probability, and physics.

Film and Television: In the short drama industry, some content creators are beginning to use AIGC tools to expand their market, including face-swapping technology to change actors' faces to those familiar to target market audiences. Heygen [64], for instance, uses AIGC to instantly generate digital human videos resembling users and provides a rich library of backgrounds and templates. Platforms like QuickVid integrate text generation from GPT-4 and text-to-image models like DALL-E 2 to create complete video clips from brief text inputs. OpenAI's Sora model can generate complex scenes with multiple characters and detailed backgrounds from text prompts, enabling creators to quickly visualize and produce high-quality video content with minimal effort.

5. Limitations and Future directions

The rapid development of artificial intelligence has made the future of AIGC brim with potential. However, it is evident that current technologies and models still have many problems and shortcomings. We can foresee numerous innovations and advancements in the coming years. This section will discuss the current flaws and issues of AIGC and tentatively propose solutions and possible future directions for development.

More Efficient Model Architectures: While current AIGC models are highly powerful, since the advent of Transformers, large-parameter Transformer-like structures have dominated the field, including models like DALL-E, Parti, Sora, and Llama. In Diffusion-based models, attention mechanisms have become crucial for multimodal connections, as seen in Stable Diffusion, DALL-E 2, and Imagen. However, the computational complexity of attention mechanisms grows quadratically with the input sequence length, leading to substantial computational and resource costs. Consequently, it is challenging for most individual users to deploy and train large AIGC models. Future model architectures need to be more efficient, maintaining or enhancing the generative capabilities of existing models while consuming fewer computational resources, thus reducing dependence on computing power. For instance,

DOI: 10.54254/2755-2721/95/2024BJ0056

Rewon Child et al. [65] proposed the Sparse Transformer to reduce the complexity of the transformer's attention mechanism. Optimizing model architecture can adjust computational resources and enhance computation speed, thereby broadening the application scenarios of AIGC technology.

Enhanced Multimodal Capabilities: Current text-to-image models can generate high-quality images from text descriptions, but these images still have some defects and biases [66,67]. For instance, the generated images are often rich in color and complexity, but they struggle with generating simple and abstract images. Additionally, there are biases in images generated from vague text descriptions, such as DALL-E 3's admitted tendency to produce images depicting primarily white, young, and female characters when describing people. In this context, future enhancements in multimodal capabilities must address these issues. One potential solution is machine unlearning [68], which aims to remove biased data from model training. Future advancements should involve developing sophisticated algorithms to handle abstract concept representations and correct biases in the generative process, thereby endowing models with deeper contextual understanding to better meet user needs in creation, design, and other fields.

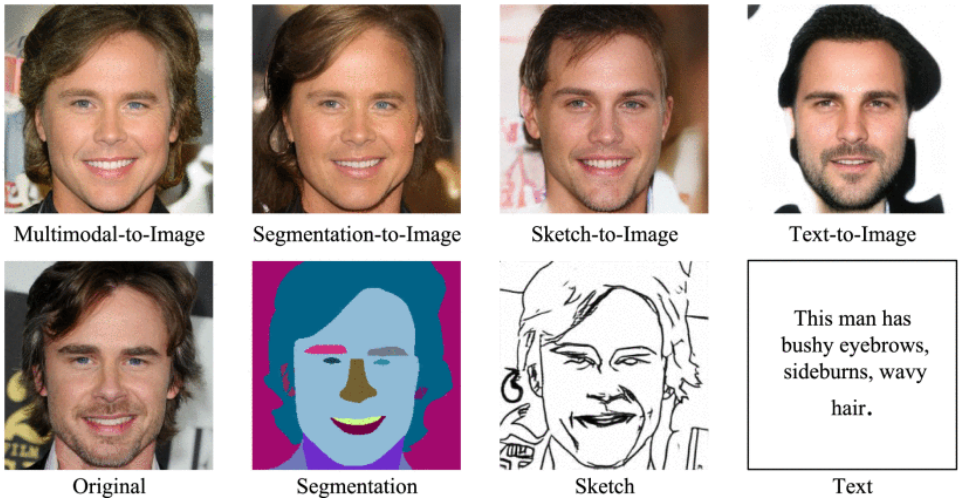

Figure 4: Illustration of Modality Bias in face generation [69], "Multimodal-to-Image" is highly consistent with "Segmentation-to-Image," while the information in text like "sideburns" is not fully utilized, which implies the generation is biased towards the modality of segmentation.

Deeper Content Personalization and Broader Modal Integration: Research on AIGC technology in more modalities remains relatively immature. By deeply learning user preferences and behaviors, and combining this with advanced recommendation systems, AIGC models can provide more customized content generation services. Current models also show insufficient capabilities in modal integration. For example, Su et al. [69] have found that models have strong modal biases when generating faces, indicating a strong correlation between multimodal generated face images and 'Segmentation-to-Image'. Therefore, a significant research direction for the future is enhancing the ability to process more modal data and fully utilizing the differential features of multimodal inputs to generate image data that better meets user needs.

Collaborative Global Research Framework: The current AIGC technology environment still faces issues such as inconsistent standards, unclear frameworks, technology misuse, and security challenges. Standardized and regulated technical research frameworks are the cornerstone of ensuring the reliability and compatibility of AIGC technology globally. In the future, we foresee international organizations and research institutions working together to issue a series of unified technical standards. These standards will cover various aspects such as model training, content generation, and user privacy protection, aiming to ensure the fair use of technology, prevent misuse, and promote the accessibility of technology across cultures and languages. By adhering to these common standards, the development

and deployment of AIGC technology will become more transparent, fostering global trust and acceptance of these technologies.

Energy and Carbon Emissions: Larger models lead to greater resource demands and increased carbon emissions. ChatGPT, for instance, responds to approximately 200 million requests daily, consuming over 500,000 kWh of electricity. The GPT-3 model, which serves as the foundation for ChatGPT, emitted 502 metric tons of carbon during its training [70]. Llama 3's model pre-training required 7.7 million GPU hours [71], with total carbon emissions equivalent to 2,290 tons of CO2. While AI enhances efficiency, it also poses potential energy crises and environmental pollution. Future development directions must include sustainability plans to optimize and offset the significant energy consumption and carbon emissions generated by model training.

6. Conclusion

This paper thoroughly investigates the historical development, foundational technologies, and underlying principles of image generation models. It provides a comprehensive analysis of generative models from both single-modal and multimodal perspectives. Additionally, the paper explores the latest applications and current advancements in AIGC technology, examining in detail the societal transformations and impacts brought about by AIGC. Based on the current technical shortcomings of AIGC, the paper proposes potential improvements and feasible directions, and discusses the future development of AIGC models, emphasizing the integration of scientific and humanistic aspects. This paper aims to help researchers and related professionals gain a deeper understanding of this field through a detailed investigation of AIGC image generation models, and to inspire further exploration and research in the AIGC domain.

References

[1] Betker Goh Jing Brooks Wang Li et al. *Improving image generation with better captions*. Computer Science https://cdn openai com/papers/dall-e-3 pdf. 2023;2(3):8.

[2] Liu Zhang Li Yan Gao Chen et al. *Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models* [Internet]. arXiv; 2024 [cited 2024 Jun 21]. Available from: http://arxiv.org/abs/2402.17177

[3] *Introducing Meta Llama 3: The most capable openly available LLM to dateProc.Symp.Int.Conf.2nd* [Internet]. Meta AI. [cited 2024 Jun 21]. Available from: https://ai.meta.com/blog/meta-llama-3/

[4] Yu Xu Koh Luong Baid Wang et al. *Scaling Autoregressive Models for Content-Rich Text-to-Image Generation* [Internet]. arXiv; 2022 [cited 2024 Jun 21]. Available from: http://arxiv.org/abs/2206.10789

[5] Bao Xiang Yue He Zhu Zheng et al. *Vidu: a Highly Consistent, Dynamic and Skilled Text-to-Video Generator with Diffusion Models* [Internet]. arXiv; 2024 [cited 2024 Jun 21]. Available from: http://arxiv.org/abs/2405.04233

[6] Wang Fu He Hao Wu. *A survey on large-scale machine learning*. IEEE Transactions on Knowledge and Data Engineering. 2020;34(6):2574–94.

[7] Li Gan Yang Yang Li Wang et al. *Multimodal foundation models: From specialists to general-purpose assistants*. Foundations and Trends® in Computer Graphics and Vision. 2024;16(1–2):1–214.

[8] Mahajan Girshick Ramanathan He Paluri Li et al. *Exploring the Limits of Weakly Supervised Pretraining*. In 2018 [cited 2024 Jun 21]. p. 181–96. Available from: https://openaccess.thecvf.com/content_ECCV_2018/html/Dhruv_Mahajan_Exploring_the_Limits_ECCV_2018_paper.html

[9] Deng Dong Socher Li Li Fei-Fei. *Imagenet: A large-scale hierarchical image database*. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248–55.

[10] Schuhmann Vencu Beaumont Kaczmarczyk Mullis Katta et al. *Laion-400m: Open dataset of clip-filtered 400 million image-text pairs*. arXiv preprint arXiv:211102114. 2021;

[11] Sun Shrivastava Singh Gupta. *Revisiting unreasonable effectiveness of data in deep learning era*. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 843–52.
[12] Zhan Yu Wu Zhang Lu Liu et al. *Multimodal image synthesis and editing: The generative AI era*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023;45(12):15098–119.
[13] Stauffer Grimson. *Adaptive background mixture models for real-time tracking*. In: Proceedings 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat No PR00149) [Internet]. 1999 [cited 2024 Jun 22]. p. 246-252 Vol. 2. Available from: https://ieeexplore.ieee.org/abstract/document/784637
[14] Bengio Ducharme Vincent. *A Neural Probabilistic Language Model*. In: Advances in Neural Information Processing Systems [Internet]. MIT Press; 2000 [cited 2024 Jun 21]. Available from: https://proceedings.neurips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html
[15] Efros Leung. *Texture synthesis by non-parametric sampling*. In: Proceedings of the Seventh IEEE International Conference on Computer Vision [Internet]. 1999 [cited 2024 Jun 22]. p. 1033–8 vol.2. Available from: https://ieeexplore.ieee.org/abstract/document/790383
[16] Pérez Gangnet Blake. *Poisson Image Editing*. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2 [Internet]. 1st ed. New York, NY, USA: Association for Computing Machinery; 2023 [cited 2024 Jun 22]. p. 577–82. Available from: https://doi.org/10.1145/3596711.3596772
[17] Rumelhart Hinton Williams. Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, ed. DE Rumelhart and J. McClelland. Vol. 1. 1986. Biometrika. 1986;71:599–607.
[18] Elman. Finding Structure in Time.
[19] Hochreiter Schmidhuber. *Long short-term memory*. Neural computation. 1997;9(8):1735–80.
[20] Cho Van Merriënboer Gulcehre Bahdanau Bougares Schwenk et al. *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. arXiv preprint arXiv:14061078. 2014;
[21] Kingma Welling. *Auto-encoding variational bayes*. arXiv preprint arXiv:13126114. 2013;
[22] Goodfellow Pouget-Abadie Mirza Xu Warde-Farley Ozair et al. *Generative Adversarial Nets*. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2014 [cited 2024 Jun 21]. Available from: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html
[23] Ho Jain Abbeel. *Denoising Diffusion Probabilistic Models*. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2020 [cited 2024 Jun 21]. p. 6840–51. Available from: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html
[24] Kullback Leibler. *On information and sufficiency*. The annals of mathematical statistics. 1951;22(1):79–86.
[25] Tolstikhin Bousquet Gelly Schoelkopf. *Wasserstein Auto-Encoders* [Internet]. arXiv; 2019 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/1711.01558
[26] Higgins Matthey Pal Burgess Glorot Botvinick et al. *beta-vae: Learning basic visual concepts with a constrained variational framework.* ICLR (Poster). 2017;3.
[27] van den Oord Vinyals kavukcuoglu. *Neural Discrete Representation Learning*. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2024 Jun 22]. Available from: https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html
[28] Razavi van den Oord Vinyals. *Generating Diverse High-Fidelity Images with VQ-VAE-2*. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2019 [cited 2024 Jun 22]. Available from: https://proceedings.neurips.cc/paper/2019/hash/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Abstract.html

DOI: 10.54254/2755-2721/95/2024BJ0056

[29] Mirza Osindero. *Conditional generative adversarial nets*. arXiv preprint arXiv:14111784. 2014;
[30] Radford Metz Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434. 2015;
[31] Arjovsky Chintala Bottou. *Wasserstein GAN* [Internet]. arXiv; 2017 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/1701.07875
[32] Zhu Park Isola Efros. *Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks*. In 2017 [cited 2024 Jun 22]. p. 2223–32. Available from: https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html
[33] Zhang Goodfellow Metaxas Odena. *Self-attention generative adversarial networks*. In: International conference on machine learning. PMLR; 2019. p. 7354–63.
[34] Ho. Autoregressive Models in Deep Learning—A Brief Survey. George Ho. 2019;
[35] Oord Kalchbrenner Kavukcuoglu. *Pixel Recurrent Neural Networks*. In: Proceedings of The 33rd International Conference on Machine Learning [Internet]. PMLR; 2016 [cited 2024 Jun 22]. p. 1747–56. Available from: https://proceedings.mlr.press/v48/oord16.html
[36] van den Oord Kalchbrenner Espeholt kavukcuoglu Vinyals Graves. *Conditional Image Generation with PixelCNN Decoders*. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2016 [cited 2024 Jun 22]. Available from: https://proceedings.neurips.cc/paper_files/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html
[37] Esser Rombach Ommer. *Taming Transformers for High-Resolution Image Synthesis*. In 2021 [cited 2024 Jun 22]. p. 12873–83. Available from: https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html?ref=
[38] Song Meng Ermon. *Denoising Diffusion Implicit Models* [Internet]. arXiv; 2022 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/2010.02502
[39] Rombach Blattmann Lorenz Esser Ommer. *High-Resolution Image Synthesis With Latent Diffusion Models*. In 2022 [cited 2024 Jun 22]. p. 10684–95. Available from: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html
[40] Ho Salimans. *Classifier-Free Diffusion Guidance* [Internet]. arXiv; 2022 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/2207.12598
[41] Song Sohl-Dickstein Kingma Kumar Ermon Poole. *Score-Based Generative Modeling through Stochastic Differential Equations* [Internet]. arXiv; 2021 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/2011.13456
[42] Bahdanau Cho Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate* [Internet]. arXiv; 2016 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/1409.0473
[43] Vaswani Shazeer Parmar Uszkoreit Jones Gomez et al. *Attention is All you Need*. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2017 [cited 2024 Jun 22]. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
[44] Zhang Xu Li Zhang Wang Huang et al. *StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks*. In 2017 [cited 2024 Jun 22]. p. 5907–15. Available from: https://openaccess.thecvf.com/content_iccv_2017/html/Zhang_StackGAN_Text_to_ICCV_2017_paper.html
[45] Xu Zhang Huang Zhang Gan Huang et al. *AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks*. In 2018 [cited 2024 Jun 22]. p. 1316–24. Available from: https://openaccess.thecvf.com/content_cvpr_2018/html/Xu_AttnGAN_Fine-Grained_Text_CVPR_2018_paper.html
[46] Kang Zhu Zhang Park Shechtman Paris et al. *Scaling Up GANs for Text-to-Image Synthesis*. In 2023 [cited 2024 Jun 22]. p. 10124–34. Available from: https://openaccess.thecvf.com/

content/CVPR2023/html/Kang_Scaling_Up_GANs_for_Text-to-Image_Synthesis_CVPR_2023_paper.html

[47] Dosovitskiy Beyer Kolesnikov Weissenborn Zhai Unterthiner et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* [Internet]. arXiv; 2021 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/2010.11929

[48] Chang Zhang Barber Maschinot Lezama Jiang et al. *Muse: Text-To-Image Generation via Masked Generative Transformers* [Internet]. arXiv; 2023 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/2301.00704

[49] Ramesh Pavlov Goh Gray Voss Radford et al. *Zero-Shot Text-to-Image Generation*. In: Proceedings of the 38th International Conference on Machine Learning [Internet]. PMLR; 2021 [cited 2024 Jun 22]. p. 8821–31. Available from: https://proceedings.mlr.press/v139/ramesh21a.html

[50] Radford Kim Hallacy Ramesh Goh Agarwal et al. *Learning Transferable Visual Models From Natural Language Supervision*. In: Proceedings of the 38th International Conference on Machine Learning [Internet]. PMLR; 2021 [cited 2024 Jun 22]. p. 8748–63. Available from: https://proceedings.mlr.press/v139/radford21a.html

[51] Nichol Dhariwal Ramesh Shyam Mishkin McGrew et al. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models* [Internet]. arXiv; 2022 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/2112.10741

[52] Saharia Chan Saxena Li Whang Denton et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. Advances in Neural Information Processing Systems. 2022 Dec 6;35:36479–94.

[53] Raffel Shazeer Roberts Lee Narang Matena et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Journal of Machine Learning Research. 2020;21(140):1–67.

[54] Radford Narasimhan Salimans Sutskever others. *Improving language understanding by generative pre-training*. 2018;

[55] Radford Wu Child Luan Amodei Sutskever et al. *Language models are unsupervised multitask learners*. OpenAI blog. 2019;1(8):9.

[56] Brown Mann Ryder Subbiah Kaplan Dhariwal et al. *Language Models are Few-Shot Learners*. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2020 [cited 2024 Jun 22]. p. 1877–901. Available from: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[57] Achiam Adler Agarwal Ahmad Akkaya Aleman et al. *Gpt-4 technical report*. arXiv preprint arXiv:230308774. 2023;

[58] Chen Tworek Jun Yuan Pinto Kaplan et al. *Evaluating Large Language Models Trained on Code* [Internet]. arXiv; 2021 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/2107.03374

[59] Bass News. *Microsoft is rolling out generative AI in Windows and Office app* [Internet]. [cited 2024 Jun 22]. Available from: https://techxplore.com/news/2023-09-microsoft-generative-ai-windows-office.html

[60] *What's next for AI in 2024Proc.Symp.Int.Conf.2nd* [Internet]. MIT Technology Review. [cited 2024 Jun 22]. Available from: https:// www.technologyreview. com/2024/01/04/1086046/whats-next-for-ai-in-2024/

[61] Ramesh Dhariwal Nichol Chu Chen. *Hierarchical text-conditional image generation with clip latents*. arXiv preprint arXiv:220406125. 2022;1(2):3.

[62] *Colorize Photo | Try Free | Realistic ColorsProc.Symp.Int.Conf.2nd* [Internet]. [cited 2024 Jun 22]. Available from: http://www.palette.fm

[63] Lewkowycz Andreassen Dohan Dyer Michalewski Ramasesh et al. *Solving Quantitative Reasoning Problems with Language Models*. Advances in Neural Information Processing Systems. 2022 Dec 6;35:3843–57.

DOI: 10.54254/2755-2721/95/2024BJ0056

[64] *HeyGen Raises $60M Series A to Scale Visual Storytelling for Businesses | HeyGen BlogProc. Symp.Int.Conf.2nd* [Internet]. [cited 2024 Jun 22]. Available from: https://www.heygen.com/article/announcing-our-series-a

[65] Child Gray Radford Sutskever. *Generating Long Sequences with Sparse Transformers* [Internet]. arXiv; 2019 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/1904.10509

[66] Zhang Lemoine Mitchell. *Mitigating Unwanted Biases with Adversarial Learning*. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society [Internet]. New York, NY, USA: Association for Computing Machinery; 2018 [cited 2024 Jun 22]. p. 335–40. (AIES '18). Available from: https://dl.acm.org/doi/10.1145/3278721.3278779

[67] Lapuschkin Wäldchen Binder Montavon Samek Müller. *Unmasking Clever Hans predictors and assessing what machines really learn*. Nat Commun. 2019 Mar 11;10(1):1096.

[68] Bourtoule Chandrasekaran Choquette-Choo Jia Travers Zhang et al. *Machine unlearning*. In: 2021 IEEE Symposium on Security and Privacy (SP). IEEE; 2021. p. 141–59.

[69] Su Zhu Gao Song. Utilizing Greedy Nature for Multimodal Conditional Image Synthesis in Transformers. IEEE Transactions on Multimedia. 2024;26:2354–66.

[70] Patterson Gonzalez Le Liang Munguia Rothchild et al. *Carbon Emissions and Large Neural Network Training* [Internet]. arXiv; 2021 [cited 2024 Jun 22]. Available from: http://arxiv.org/abs/2104.10350

[71] Henderson Hu Romoff Brunskill Jurafsky Pineau. *Towards the systematic reporting of the energy and carbon footprints of machine learning*. Journal of Machine Learning Research. 2020;21(248):1–43.