Development And Challenges of Generative Artificial Intelligence in Education and Art

Junpeng Yang [1, *], Haoran Zhang [2]

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

[2] No.7 middle school, Chongqing, China

* Corresponding Author Email: 23112982g@connect.polyu.hk

Abstract. Thanks to the rapid development of generative deep learning models, Artificial Intelligence Generated Content (AIGC) has attracted more and more research attention in recent years, which aims to learn models from massive data to generate relevant content based on input conditions. Different from traditional single-modal generation tasks that focus on content generation for a particular modality, such as image generation, text generation, or semantic generation, AIGC trains a single model that can simultaneously understand language, images, videos, audio, and more. AIGC marks the transition from traditional decision-based artificial intelligence to generative artificial intelligence, which has been widely applied in various fields. Focusing on the key technologies and representative applications of AIGC, this paper identifies several key technical challenges and controversies in the field. These include defects in cross-modal and multimodal generation, issues related to model stability and data consistency, privacy concerns, and questions about whether advanced generative models like ChatGPT can be considered general artificial intelligence (AGI). While this dissertation provides valuable insights into the revolution and challenge of generative AI in art and education, it acknowledges the sensitivity of generated content and the ethical dilemmas it may pose, and ownership rights for AI-generated works and the need for new intellectual property norms are subjects of ongoing discussion. To address the current technical bottlenecks in cross-modal and multimodal generation, future research aims to quantitatively analyze and compare existing models, proposing practical optimization strategies. With the rapid advancement of generative AI, we anticipate a transition from user-generated content (UGC) to artificial intelligence-generated content (AIGC) and, ultimately, a new era of human-computer co-creation with strong interactive potential in the near future.

Keywords: Generative Artificial intelligence; Education; Art; Cross-modal; Multimodal.

1. Introduction

1.1. Background

The development of generative artificial intelligence can be traced back to early computer science in the mid-20th century [1]. Early generative artificial intelligence mainly focused on automatic language translation and natural language generation. IBM's Georgetown-IBM Laboratory successfully used machines for automatic translation in 1954 , but progress was slow due to computational resource and algorithm limitations. Afterward, the expert systems in the 1960s and 1970s aimed to simulate the human reasoning process by using the knowledge of domain experts. For example, the MYCIN system contributed to medical diagnosis [1]. Their limitations lay in the need for manual maintenance of their knowledge bases and the inability to perform truly creative generation. By the 1990s, statistical methods such as N-gram models and hidden Markov models were introduced in the artificial intelligence(AI) field for natural language processing and language generation. These specific models perform well in short text generation and speech recognition but still have shortcomings in generating long text and diverse content. Until around 2010, the rise of deep learning brought new hope to artificial intelligence-generated content(AIGC) [2]. Models such as recurrent neural networks (RNN) [3] and long short-term memory networks (LSTM) [4] improve natural language generation, allowing them to handle longer contexts. Later, the emergence of variational autoencoders (VAE) [5], generative adversarial networks (GAN) [6], and transformers [7]

(such as the ChatGPT series) further promoted the development of generative models, enabling them to generate more creative and realistic content. Along with previous theories, the emergence of concepts like reinforcement learning [8] and self-attention mechanisms [9] has joined together to further fill the gap in the core technology of generative AI.

In domains of education and art, pre-21st century personalized learning systems draw on rules and knowledge bases to customize teaching content for students [10]. However, these systems are limited by knowledge representation and rule updating. Electronic artists of the same period also experimented with computer-generated images and music [11]. These experimental works are often based on simple algorithms and geometric shapes, limiting creative expression. During the deep learning period, especially after the introduction of GAN and Transformer, generative AI has made significant progress in intelligent-assisted teaching [12]. Virtual teaching assistants and intelligent tutorials can provide students with real-time feedback, answer questions, and adjust teaching content based on student performance[13]. This personalized teaching method helps improve students' learning effectiveness and interest. It can also automate assignment assessment, analyze, and evaluate assignments submitted by large numbers of students, reduce the workload of teachers, and help students better understand concepts. Artists also began to create more complex, diverse, and creative works. Generative art includes image generation, music, and poetry creation, etc [14]. Throughout the entire historical process, the successive evolution reveals the transformation and breakthrough from traditional decision-making AI to generative AI, the latter of which learns the joint probability distribution in the data, analyzes and summarizes the existing data, and then performs deductive creation based on history, as well as generate brand new content via imitation and stitched creation and also inherit the characteristics of decision-making AI to solve practical problems.

Recently, generative AI has achieved great success in domains such as text generation, image generation, and music creation. Generative models have become a breakthrough innovation that can cultivate creativity and produce novel content in various areas. Of particular importance is their profound impact on the fields of education and the arts, revolutionizing traditional ways of learning, creation, and expression [15]. However, these models still suffer from challenges such as bias, insufficient interpretability, and reliance on data [16]. Future developments will focus on increasing the diversity, creativity, and controllability of generated content, as well as addressing the ethical and social implications of the model.

1.2. Objective and Motivation

The integration of deep learning-driven generative models in numerous domains has made significant advances, offering hope and opportunity to empower humans and redefine the nature of teaching and creativity. This dissertation will delve into the multifaceted promise of generative AI, exploring in particular its current state of development and transformative, challenges in education and the arts.

In this review paper, we briefly enumerate the core technologies of AIGC and go through the major technical bottlenecks and controversies that mainly exist. (1) Defects of generative models in cross-modal and multi-modal generation (2) Challenges of model stability and data consistency. (3) Privacy protection issues. (4) Whether the state-of-the-art generative model (ChatGPT) is General Artificial Intelligence (AGI) or not. All of them are valuable research directions in the path of seeking solutions and promoting the development of AIGC. It also surveys the mainstream typical applications (Stable Diffusion, etc.) in the arts and education segments and the algorithm and theory behind them, and then goes on to detail the latest advances in neural networks, natural language processing, computer vision, and other relevant AI technologies that help create complex generative models. These models, with their ability to understand patterns and contexts, open up new avenues for fostering creativity and unleash the imagination of educators and artists. Moreover, we will summarize the previous experiences, review and discuss the above-mentioned progress and challenges in the fields of education and art, as well as put forward speculations on future applications: by combining with new fields such as 3D printing and virtual reality technologies, generative AI will have the potential to

provide more possibilities for artists and educators, thereby promoting the rapid development of these two fields under the collective emergence of generative AI applications, and promoting higher-dimensional artistic innovation and personalized education. Ultimately, we hold a promising view that users can witness the era of user-generated content (UGC) to artificial intelligence-generated content (AIGC), and then to the new intelligent era of human-computer-generated content with strong interaction or human-computer co-creation in the near future.

2. Methodology

2.1. Literature Review

Variational Autoencoder (VAE) is a generative model, proposed by Kingma and Welling, that combines ideas from autoencoder and probabilistic graphical models [5]. It is an unsupervised learning neural network architecture for learning the latent representation of data and is capable of generating new data samples. VAE parameterizes the probability distribution of the latent space by mapping the input data to the mean and variance of the latent space, then draws samples from the latent space by sampling operations, and finally decodes these samples into generated data samples. This process can be regarded as the process of encoding and decoding the input data, while a probabilistic nature is introduced, enabling the model to generate diverse data samples [5]. VAE has achieved remarkable results in tasks such as generative modeling and data dimensionality reduction representation and has become an important research direction in the field of generative modeling. It performs well in image processing tasks such as image generation, image reconstruction, and image interpolation. Generative Adversarial Network (GAN) proposed by Ian Goodfellow and others in the same period also laid the technical foundation of generative AI, and its core idea is to train two neural network models by letting them compete with each other: a generator network (Generator) and a discriminator network (Discriminator) [6]. The generator network is responsible for generating fake data samples from random noise, while the discriminator network is responsible for distinguishing between real data samples and fake samples generated by the generator [6]. The two networks work against each other during training, with the generator trying to generate more realistic samples and the discriminator trying to distinguish between true and false samples more accurately. As training proceeds, the generator and discriminator are optimized until the generator is able to generate increasingly realistic samples and the discriminator's ability to discriminate continues to improve. Eventually, when the samples generated by the generator cannot be distinguished by the discriminator, it can be assumed that the generator has learned to generate realistic samples. The proposal of GAN is considered to be an important milestone in the field of deep learning and generative modeling, which has achieved remarkable results in various tasks such as image synthesis, image enhancement, image super-resolution, and image transformation [17], and has been widely used in computer vision (CV) along with VAE, natural language processing (NLP) and other fields, but suffers from the shortcomings of slow training, low-quality of the generated content, and the lack of adaptability of the model itself to the complex scenes in arts or education systems.

The transformer model put forwards afterwards is to tackle the bottlenecks in previous generative models. Compared with traditional recurrent neural networks (RNN) and convolutional neural networks (CNN), the transformer introduces a self-attention mechanism, which enables the model to better handle long-distance dependencies and parallelize computation more efficiently [7]. Its key feature is that the self-attention mechanism allows the model to perform adaptive weighted aggregation at all positions in the input sequence, thus better capturing important relationships between different positions in the input [9]. Such an attentional mechanism allows the transformer to handle longer sequences and has achieved significant performance gains in NLP tasks. Since its introduction, the transformer has become an important infrastructure in the field of NLP and has been the cornerstone of various new models, such as BERT, the ChatGPT series, and others. The pre-training-fine-tuning approach of transformer saves considerable training time and resources and is suitable for a wide range of generative tasks, enabling the generation of high-quality text and

multimedia content. Moreover, models such as RNN and LSTM are used for sequential data (text, audio), have memory capabilities to capture long-term dependencies, and improve natural language generation to handle longer contexts [3, 4]. As an extension of VAE and GAN, conditional variation autoencoder (CAVE) [18] and conditional generative adversarial network (cGAN) [19] both introduce conditionality to generative AI, which enables the model to be guided by specific conditions during generation, thus enhancing the generated content. guided by specific conditions during the generation process, thus enhancing the control and customization of generated content, which in turn promotes personalized education and creative arts. It is worth mentioning that the application scenarios of both art and education are relatively complex, involving diverse data distributions and unstable environments. Reinforcement learning can help generative models to optimize the generation strategy, improve the performance through automated hyper-parameter tuning [8], and achieve adaptive generation and personalized generation, which further strengthens the diversity of generative art and education, and derives more appealing and interesting educational resources, as well as more creative and emotionally charged artworks.

2.2. Bottleneck and Controversy

AIGC generally stays in the traditional generative model, which only targets a single data type, and has certain defects in cross-modal and multi-modal generation [20]; Cross-modal generation generally requires large-scale multimodal datasets, which are often difficult to obtain and label. The bottlenecks mainly come from data fusion: fusing data from different modalities into a consistent representation, and data alignment: ensuring that data from different modalities are semantically aligned. In addition, the lack of a model structure that can handle multimodal inputs, traditional single-modal evaluation criteria that may not be applicable to cross-modal and multimodal tasks, overfitting of multimodal generative models, and making complex black-box models more interpretable are also unresolved issues.

When generating the model, there is also a possibility of obtaining inaccurate data due to noise and other problems; at the same time, how to predict and generate the data on a long-term time scale, the challenges for both model stability and data consistency. GAN, as an important method in generative modeling, may be unstable in its training process and even difficult to converge to the desired state [17]. Finally, generative models may have security vulnerabilities [16]. The sources of datasets in the training corpus of the pre-trained models are partly unverified and unauthorized by individuals, and the data security, integrity, authenticity, and legality of the private information captured are questioned. Unsupervised models with no filtering mechanism for database information in the pre-learning phase may also generate undesirable, false, or illegal information due to "black box" algorithmic in the runtime phase.

The current controversial topic is whether generative models are AGI or not [21]. As a representative of the state-of-the-art generative models, ChatGPT is based on the GPT-3.5 architecture, with large-scale pre-training parameters and powerful language understanding and generation capabilities, which enables it to understand and generate natural language text, thus having a wide range of general-purpose application capabilities. It can be used for a wide range of natural language processing tasks such as text generation, Q&A, dialog, translation, summarization, etc [20]. Moreover, the model is pre-trained based on a large amount of data and can learn from data and adapt to new tasks. Since it passed the Turing test, and throughout the conversation, ChatGPT demonstrated its rich knowledge, humorous style, logical thinking, and expression of emotions, a considerable number of scholars firmly claim that it is already a general-purpose AI model.

However, ChatGPT, though powerful and miraculous, still has its limitations. Conservative scholars insist on the opposing view that ChatGPT has not yet reached the level of AGI. (1) First, ChatGPT is limited to the natural language domain. It can be called a more generalized natural language model because conversation topics can be freely chosen and various tasks can be accomplished, but multimodal tasks other than text, such as playing Go and image recognition, cannot be accomplished by it. (2) ChatGPT requires a large amount of data for training. General-purpose AI

needs to be able to discover patterns using a small amount of data because unknown domains usually do not have a large amount of known data. ChatGPT has weak learning ability, for example, it still needs to improve in understanding context, reasoning, and common-sense inference, and there is still a considerable gap with human intelligence. (3) There is no closed loop to realize self-iteration, AGI can design itself better than humans for rapid self-evolution. However, ChatGPT enters the technical bottleneck without the development and help of technicians, and its current underlying technology cannot support its self-evolution.

From another perspective, stochastic parrot theory reveals that the creative process may not be entirely predictive [22]. For example, in computer-generated art, the use of stochastic techniques may lead to unique, unexpected works. Similarly, in machine learning and generative modeling, the introduction of randomness can increase diversity and innovation. ChatGPT may employ stochastic parrot theory to add randomness to the content-generation process, but this does not prove that it has the ability to truly understand and create that content. In summary, we remain skeptical of the viewpoint that ChatGPT is a form of general-purpose AI, but it still continues to be recognized by the general public as a prototype of first-ready AGI features.

3. Application

3.1. Milestone Models and Applications for Generative Art and Education

In the period of deep learning, Google released DeepDream, an image generation tool based on CNN that transforms images into works of art full of fantastical and abstract features [23]. Its key feature is to transform images into psychedelic and abstract works of art by applying filters at different levels in the CNN to enhance specific features in the image to create visually stunning effects. Applied to the creation of image art, it can be used to generate psychedelic-style images and form a unique art genre of its own. There are plenty of variants of the GAN technology that have appeared since its introduction, and in the realm of art, Stable Diffusion is a classic variant of GAN for generating images and audio [24]. Its mission is to generate high-quality images and audio by controlling the degree of noise diffusion, it introduces a stabilized diffusion process, which results in superior generation quality compared to the roughness of traditional GAN. It is widely used to generate high-quality images and audio, including image synthesis, music generation, and sound synthesis. Moreover, Neural Style Transfer, leads the way in style migration, allowing artistic styles to be applied from one image to another to create unique artistic effects [25]. CycleGAN is a model for image translation, which can convert images from one domain to another domain [26], such as converting an image of a horse to an image of a zebra.

However, these are still models for a single modality, and in recent years, breakthroughs have been made in cross-modal generation. The DALL-E model, published by OpenAI, generates images that match descriptions based on textual descriptions [27]. It introduces the ability of multimodal generation from text to image, creating a new way of generating images. Afterward, Midjourney uses CGAN, which means that the model not only needs to learn how to generate realistic images but also needs to learn how to generate corresponding images based on textual data inputted by the user. During the training process, it pairs the text data with the corresponding image and inputs it as conditional information into CGAN, which is then able to generate an image that matches the text data based on the input conditional information. Midjourney is then able to generate the corresponding image based on that text data without the need to manually draw or edit it. Such an approach not only improves the efficiency of creation but also generates high-quality, realistic visual effects to meet the user's creative needs. In the field of textual art, the ChatGPT series can also be applied for the generation and creation of literary works with different styles and backgrounds, with a high degree of freedom and complexity.

For education sectors, bidirectional encoder representations from transformers (BERT) models have made breakthroughs in the field of natural language understanding and question answering, providing smarter analytics and feedback for educational technology and materials [28]. The

ChatGPT series can also be employed as an automated question-and-answer system in education, these systems can answer questions posed by students and provide explanations and contextual information. Intelligent Tutoring Systems (ITS) is already a well-established application in the domain of generative educational technology that uses machine learning and natural language processing to provide personalized educational support based on a student's academic level and needs and is designed to help students improve academic performance, fill in knowledge gaps, and provide customized learning experiences. Examples include generating customized textbooks, practice questions, and solution plans to meet students' needs.

To sum up, these are landmark models and applications of generative art and education in the last decade. Along with the maturation of generative technologies and models, the trend is the shift from single-modal to multi-modal input, model, and output data, as well as the generation contents of more and more refined, high-quality, and explanatory, that is closer to the standards and norms of human-intended art and education.

3.2. AIGC Application Scenarios Classification

Table 1. AIGC Application Scenario Classification Based on Modal Type

Modal Type	Technical Details	Usage Examples	Application	
Text Generation	Large-scale pre-training model, Natural language understanding, Dialogue strategy selection	Story, Scriptwriting Academic Chatbot	ChatGPT series,BERT	
Audio Generation	Small sample transfer learning for small amounts of data in tone; Music data annotation	Music generation Lyrics,Composition, arrangement,Vocal recording and Mixing	AIVA,Amper Music	
Image Generation	GAN, Diffusion Model	Picture dress-up, Image face-changing, Expression modification Graphic design, NFT	DeepDream,Stable Diffusion,Midjourney	
3D Model	NeRF(neural radiation field)	Virtual humans, VR/AR, 3D games, Movie special effects	DreamFusion,GET3D	
Video Generation	CV (object recognition and tracking), Image editing technology	Video property editing, quality repair, deletion of specific subjects, automatic tracking of theme clips, Video special effects, Automatic beautification	Synthesia, DeepBrain AI	
Multimodal, Cross-modal Generation	Multi-modal learning, cross-modal understanding; large-scale pre-training models; multi-modal, cross-modal content generation	AI chat painting robot AI generated video Multi-modal intelligent interactive digital human	ChatGPT4.0, DreamFusion, Midjourney, Synthesia, AIVA	

Table 1 demonstrates the AIGC application scenarios classification based on mainstream modal types, as well as enumerating corresponding technical details, usage examples, and typical applications in a concise way.

Innovative trends in generative modeling are also moving toward the idea of integrating traditional technologies into emerging fields, in terms of 3D model unfolding. As multiple industries move toward large-scale 3D virtual worlds under the influence of Metaverse, tools that can generate large amounts of high-quality, diverse 3D content are in industrial demand. Currently, 3D assets on the market are acquired mainly by manual design by modeling software such as Blender and Maya3D, a process that time-consuming and expertise-dependent. GET3D, released by NVIDIA, was originally developed so that Metaverse's content builders could create large and varied 3D objects faster, in the hope of training better 3D generative models to produce textured 3D models that downstream tasks could use directly. To address the trials of AI-generated 3D models with (1) lack of geometric details, (2) lack of textures, and (3) the use of neural renderers during compositing that are not convenient to use in 3D software, GET3D combines new techniques such as micro surface modeling, micro surface renderers, and 2D GANs to train the models, and realizes the ability to generate 3D models with textures, complex topologies, and rich geometric details [29]. However, GET3D still requires a 2D image to be provided as input. DreamFusion smoothly operates cross-modally by combining two new methods: neural radiation field and 2D diffusion to generate models with only text input [30].

Except for modalities stated in Table 1, strategy generation emerges as a common modality. However, it is generally more related to the field of decision-making in business, and less common in the domain of education and art, with a few marginal applications. For example, strategy generators could be curriculum planning tools for planning academic courses and optimizing course schedules; personalized learning platforms for creating personalized learning plans and recommending courses and learning paths based on students' interests, academic levels, and goals. Additionally, in art sectors, it can be in the form of art project management for planning exhibitions, creative schedules, resource allocation and publicity strategies, or creative strategies planner in assisting artists including selection of media, style, subject matter, and creative process.

4. Evolving Cross-modal, Multimodal Generation

4.1. Underlying Technology and Concepts

Evolving from rough and simple single-modal generative models to mature and diverse unimodal and then to multimodal and cross-modal research is the next major optimization, innovation direction for scholars specializing in generative AI. Recent research surveys summarized five core challenges in cross-modal and multimodal machine learning: (1) Representation (2) Translation (3) Alignment (4) Fusion (5) Co-learning [31]. The latest key underlying technologies for cross-modal and multimodal generation are centered around feature fusion, modal alignment, multimodal representation learning, sentiment analysis, multimodal generation, knowledge fusion, and transfer learning. Feature fusion is categorized into early fusion, late fusion, and end-to-end fusion, where early fusion involves fusing features from different modalities together before the data enters the model. This usually involves joining or superimposing feature vectors from different modalities in a specific way to form a larger joint feature vector [33]. Late fusion involves aggregating or fusing features together from each modality after they have been extracted and encoded independently, usually at a higher or task-specific level of the model. It benefits from the use of specialized modal feature extractors to better capture the information of each modality. End-to-end fusion is more advanced and considers not only how to fuse features, but also how to fuse information flows from different modalities. This approach typically uses deep learning models (neural networks). Modal alignment mainly considers semantic, and spatial alignment of data in different modalities (e.g., image, text, audio, etc.). Multimodal representation learning is usually based on deep learning techniques, and the main goal is to find a shared, high-dimensional feature space that helps the model to understand and analyze the correlations between data from different modalities, where the data

from each modality can be mapped to a common representation [31]. Multimodal neural networks, autoencoder, and joint training are the three main approaches, CNNs for image processing, RNN or transformer models for text processing, and convolutional or recurrent layers for audio processing. These multibranch models usually specialize in one modality per branch. Multimodal self-encoders try to learn to map different modal data to a low-dimensional shared representation and then reduce it back to the original data, which can be achieved by retaining important information and removing redundant information. Models can also learn shared representations by jointly training data from multiple modalities. The distance between the represented modalities is minimized to ensure that they are relatively close together in the representation space [32]. Multimodal sentiment analysis aims to combine sentiment information from different modalities to provide a more comprehensive understanding of sentiment, involving recognition of sentiment words and matching of sentiment lexicons in textual data, recognition of facial expressions, poses, and scenes in images, and characterization of acoustic features in audio data, which may be applied to sentiment-driven education and art systems in the future [32]. Knowledge fusion can be personalized by building knowledge graphs, sharing representation space for multi-source data fusion, transfer learning, or through collaborative filtering. Its main goal is to bring together more structured, highly information-dense, and comprehensive knowledge from several different knowledge sources. The main goal of multimodal generation is to create content that includes multiple modalities simultaneously. Data from different modalities can complement each other to provide more comprehensive, diverse perspectives. For example, generating a video description can be combined with text to better convey what is going on in a graphic video. Transfer learning, on the other hand, involves feature extraction migration (features extracted from the model trained in the source task are used as inputs to the target task), model migration (the model trained in the source task is used as an initial model for the target task, and adapted to the target task by fine-tuning), and knowledge migration (knowledge learned in the source task, e.g., weights, attention distributions, and category distributions, etc., can be used to help the target task). Modality mismatch, data imbalance, polysemy, and negative migration are common problems in multimodal and cross-modal techniques, and domain adaptation, quality of generated content, diversity, and consistency are inevitable topics as well.

4.2. Evaluation Metrics

For traditional generative models, the evaluation criteria are typically the following: perplexity, quality of generation, diversity (mostly applying N-gram repetition rate), coverage (judge whether it is possible to generate a variety of different samples from the data distribution), generation rate, stability (whether changing parameters or other conditions affects the output) [34]. Generation quality may have more specific sub-metrics depending on the specificity of the task (image sharpness, resolution, realism, texture detail, color, and contrast in CV or semantic consistency in NLP), and metrics such as structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR) are complemented as key quantitative indicators of image quality. The cross-modal evaluation metrics are somewhat changed from the traditional, common metrics, in particular, the newly introduced multimodal similarity metric, modal alignment metric, and inter-modality correlation (IMC). The first one is used to measure the similarity between different modalities [31]. The second is used to measure the degree of alignment between different modal data to evaluate the performance of modal alignment techniques. The last one is used to determine the degree of correlation between modalities to verify that the model captures cross-modal information correctly. The consistency and diversity of the cross-modal generated content are also as important as the results of the previous experiments evaluating the unimodal generation model, with consistency loss and multimodal diversity being the main metrics used in the experiments with the cross-modal model. Finally, the evaluation of multimodal generative models requires modality-specific metrics for evaluating the quality of the generation of each modality individually, which correspond to image quality assessment metrics for image generative models, NLP assessment metrics for text generative models, and audio quality assessment metrics for audio generative models [32].

5. Discussion

5.1. Revolution and Challenge: Education

5.1.1 Positive effects

(1) **Powerful Learning Assistant**

Education nowadays needs more intelligent assistants, and advanced generative models similar as ChatGPT can provide students with a personalized, comprehensive, and adapted learning solution using the educational resources it already has. Students just need to ask certain questions related to it, and it can recommend books, courses, topics, etc. with a high degree of matching for students. ChatGPT can also be a tool for feedback on students' learning process, e.g., it can intelligently grade certain test questions that students have completed before, and recommend corresponding learning methods for students, so as to achieve the purpose of feedback.

(2) **Efficient Teaching Aid**

Educational generative AI brings convenience to teachers' teaching through its massive resources and efficient interaction. Broadly speaking, this can be done in the following two ways; (a): It helps teachers to correct some of students' assignments when busy period comes. Educational generation model or system can, to a certain extent, help teachers correct questions like fill-in-the-blanks and multiple-choice questions appearing in quizzes or exams, thus reducing the pressure and workload on teachers.2: On the other hand, it can serve as a smart teaching planner, and schedule an interactive classroom outline for teachers in advance through the massive resources it has trained. Teaching staff only need to input information related to the course content, then it can automatically and easily complete the arrangement of the courses.

(3) **Provide Good Teaching Relationship**

Nowadays, students can get information through the Internet, but getting personalized information remains hard, ITS and other generative tools provide students with personalized advice promptly through its good communication methods. Such a phenomenon makes these tools become special support for students in addition to the teacher, and at the same time, students may be independent of the single fixed teacher-student teaching relationship, thus forming a self-learning and self-management learning style with the help of educational generation tools. It is worth mentioning that more studies have shown that it makes the relationship between teachers and students more open [12].

(4) **Enhanced special education**

Cross-modal and multi-modal educational generative models can enrich the forms of special education and assist people with disabilities to complete their special education, and through inter-modal conversion to achieve a more understandable form: for example, a multi-modal generative model can replace the picture of the sea with the corresponding braille text or the sound of waves to make the blind students feel another form of the sea. The cross-modal model can generate a video from the text written by a blind student to show other students or teachers the world he imagines in a more three-dimensional way.

5.1.2 Potential risk

(1) **Information Polarization**After a few effective interactions with educational generation model like ChatGPT, it will keep the data of these exchanges, so that in the future, it will only recommend the communicator's favorite content and no other content. As a result, the communicator's information becomes partially restricted and is prone to go to a certain extreme, which is relatively more obvious in the case of students. While students become dependent on educational generation models or tools, the information they receive is seriously limited.

(2) **Lack of Creativity**

Educational generative AI will have the ability to solve problems based on existing data as well as pre-existing facts and logic, which will lead to its low creativity on unknown data. Although the answers given by it can effectively inspire humans to think creatively, educational generative AI still has more room for progress in the direction of self-innovation in the human context.

(3) **Risk of Technology Misuse**

Based on educational generation AI's efficient interactive capability and its large number of resources, it could be applied to cheat on paper exams like providing answers to some subjective questions. Being addicted to educational generation AI may lure students towards academic misconduct such as AI ghostwriting, hence, educational generation AI systems or tools are suggested to be monitored and regulated by the education administration. What's more, educational generation AI can have problems with its huge database and thus copyright infringement and other issues. This makes the use of it in education partially limited.

(4) **Reducing the Authority of Teachers**

Behind the powerful resources of educational generation AI, the influence of the teacher becomes weaker. After students discover the potential of educational generation AI, they will rely on it to solve problems. Over time, such a phenomenon will have a certain negative impact on the authority of the traditional teacher.

(5) **Bias in Data Specialization**

The data generation of educational generation AI is based on the original database, if part of the database is based on previous wrong experiences or biased data, theories, models, etc., it may lead to the generation of professional information to produce a certain degree of bias, resulting in the emergence of erroneous or invalid information, which will cause negative effects on the quality of education.

5.2. Revolution and Challenge: Art

5.2.1 Enhancement of creative novelty

AIGC is capable of generating unique and novel artworks, thus providing artists with new inspirations and creative ideas. With the aid of AIGC, artists can get out of the information cocoon and inherent thinking, break through their original cultural affiliation, religious beliefs, etc., expand their creativity by leaps and bounds, and try out new artistic styles (combined, separated, or newly created), create previously unimaginable works based on the originally model-generated works.

5.2.2 High-efficiency tool in improving production

For artists, building from scratch is the most time-consuming work for each artwork. From the GANs model onwards, AIGC is capable of generating portraits with different styles and diversity, and the artists only need to optimize the screening and inkjet printing of the final hundreds of generated results, which directly and significantly reduces the workload of prototyping and improves the number of artworks produced per unit of time.

5.2.3 Serious homogenization

GAN shows strong potential in image processing and art generation, but along with more and more art practices, the system may not have the defect of creativity and is only biased towards similar styles of imitation generation. The final artwork generated is often attached to the shadow of the past art, either similar to the Cubist style or similar to the Baroque style, lack of artistic creativity to push the boundaries.

5.2.4 Models lack artistic humanistic understanding

ChatGPT-generated poetry may be influenced by the text and patterns in the training data, and while it can convey emotion in poetry, it may lack the depth of real emotion. Artistic generative models can generate poems that follow certain structures, such as rhymes or specific grammar. However, generated poetry, while grammatically and semantically readable, may have structures of expression that are incoherent or unnatural, as well as insufficiently strict or inconsistent with traditional poetic forms. Generated content is usually not characterized by a specific cultural or social context because it is trained on large-scale generalized data and lacks deep cultural understanding.

5.2.5 Unconventional genres

Art generation models may, due to special design, flaws, or limitations, generate grotesque, bizarre, cyberpunk, surrealist, or even unconventional genres that are not recognized or understood by current human common sense, and that do not conform to traditional aesthetic and artistic concepts. But as an unconventional genre, it is also an AI-led bold innovation. The addition of cross-modal and multi-modal generative models allows artists to create multi-modal artworks combining vision, sound, and touch, deriving new genres, increasing the dimensions of art, and providing audiences with a richer sensory experience.

5.2.6 Artistic authenticity

Generating artworks in an assembly-line manner may make artists over-rely on generative models and lose the patience and craftsmanship to refine their artworks, thus not bothering to give their artworks emotion, design inspiration, and humanistic meaning, which may lead to controversy over the authenticity of their artworks. In addition, a new portrait image generated by fusing a real face with a celebrity can be posted to social media as a fake, and it is hard for human users to recognize whether it is a real person or not, hence, the authenticity of the art is doubtful.

5.3. Conjecture for Future Application

5.3.1 Mass generation of movies and large-scale games via multi-modal generative model

When the multi-modal generative model matures, the single-modal generative model is gradually eliminated. AIGC can integrate image video, music, literary creation, and other art forms to generate comprehensive artistic products: such as large-scale games and movies. Game producers or directors can select the best one from the generated large-scale prototype works and make slight modifications to the finished product. Therefore, the production cost of complex large-scale artworks such as games and movies has been reduced and gradually turned into lightweight and personalized. Each user can set up a studio at home, customize and generate their favorite movies or games, and share them with others on social media or use them for commercial profit.

5.3.2 Lifelong personalized AGI academic assistants

AIGC can generalize the learning ability to all aspects of life, become a multi-modal and multi-tasking competent AGI, achieve self-iteration and self-adaptation without human intervention, and become a lifelong private and personalized AI academic assistant, which can realize all kinds of customer educational needs (including various cultural, religious, gender, personality, etc., and even the special case of disabled people) and provide all the assistance in life and academics, and completely replace the traditional school education system, and there is no longer a need for professors, teachers, trainers, and other similar occupations, which are dominated by AGI to help human beings to upgrade their intelligence.

5.3.3 Combination with other new technology fields

Enriching the forms of art generation via integrating generative AI with other emerging areas(3D printing, VR/AR, bionics), the generated works of art could appear immediately in the form of 3D printing. Countless virtual characters with unique characteristics can also be generated in virtual reality under the metaverse concept, and even in the distant high-tech future can be batch-generated bionic robots with different personalities to serve human beings.

6. Conclusion

The development history and background of generative AI can be seen from the traditional decision-making AI to generative AI transformation and breakthroughs, by listing the core methodologies and mainstream applications of AIGC, we find that along with the maturity of generative technology and models, the generated content is becoming more and more refined, high-quality, more explanatory, and closer to the standards and norms of the human intentional art and

education. The innovation trend of generative modeling is also moving towards the idea of integrating traditional technology with emerging fields. From the rough and simple single-modal generative model to the mature and diversified unimodal, and then evolving to the multi-modal and cross-modal research, is the next major optimization, and innovation direction in the academic community. Meanwhile, this paper also discusses the main technical bottlenecks and controversies that currently exist: (1) Defects of generative models in cross-modal and multimodal generation (2) Challenges of model stability and data consistency. (3) Generative models may have security vulnerabilities that may raise privacy protection issues when handling sensitive data. (4) Whether the state-of-the-art generative model (ChatGPT) is general artificial intelligence (AGI), while insightfully discussing the changes and challenges that generative AI brings to art and education. These contents may help scholars to summarize the development history of generative art and education, peep into the direction of the evolution of new technology of cross-modal generation, and commit to breaking through the existing bottlenecks; at the same time, this dissertation expands horizons and inspire artists and educators to use generative AI to unleash greater vitality and innovative potentials, and to contribute to the advancement of generative AI in the field of education and art.

Limited by length. The content generated by AIGC in the art domains may be sensitive or involve inappropriate topics, and the resulting ethical issues and privacy controversies have not been recounted much in this dissertation. For example, data shared on social media may be stolen and trained to commit AI fraud, these are serious issues. Since the works generated by AIGC are born from algorithms and are not a person or a life entity, the ownership rights may become complicated. The need for new norms for intellectual property rights is also a hot discussion direction. Regardless of the conventional and new technologies, technical bottlenecks around cross-modal and multi-modal generation are also key issues hindering the forward progress of generative AI. This dissertation only concisely outlines the current progress and challenges and has yet to propose concrete solutions or prove new theories to the unresolved technical difficulties. Based on the lack of research summaries of cross-modal and multimodal generation models now, the subsequent research plan is to quantitatively analyze and comprehensively compare the advantages and disadvantages of the existing mainstream and mature cross-modal and multimodal generation models and propose feasible optimization schemes. With the rapid development of generative AI technology, we hold a promising view that users can witness the era of user-generated content (UGC) to artificial intelligence-generated content (AIGC), and then to the new intelligent era of human-computer-generated content with strong interaction or human-computer co-creation in the near future.

Author's contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt, 2023.

[2] Jo, A. The promise and peril of generative AI. Nature, 2023, 614(1): 214-216.

[3] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. Learning representations by back-propagating errors. Nature, 1986, 323(6088): 533-536.

[4] Hochreiter, S., & Schmidhuber, J. Long short-term memory. Neural computation, 1997, 9(8): 1735-1780.

[5] Kingma, D. P., & Welling, M. Auto-encoding variational bayes, 2013.

[6] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139-144.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. Attention is all you need. Advances in neural information processing systems, 2017.

[8] Sutton, R. S., & Barto, A. G. Reinforcement learning: An introduction. MIT press, 2018.

[9] Bahdanau, D., Cho, K., & Bengio, Y. Neural machine translation by jointly learning to align and translate, 2014.

[10] Baidoo-Anu, D., & Owusu Ansah, L. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning, 2023.

[11] Cetinic, E., & She, J. Understanding and Creating Art with AI: Review and Outlook. ACM Transactions on Multimedia Computing Communications and Applications, 2022, 18(2): 1–22.

[12] GAO Lin Qi. A Model of Generative Artificial Intelligence in Personalized Learning. Journal of Tianjin Normal University (Basic Education Edition), 2023, (04): 36-40.

[13] YANG Xiao Zhe, WANG Qing Qing, WANG Ruo Xin. The Limited Capabilities of Generative Artificial Intelligence and Educational Transformation. Global Education Perspectives. 2023, (06): 3-12.

[14] Epstein, Z., Hertzmann, A., Akten, M., Farid, H., Fjeld, J., Frank, M. R., Groh, M., Herman, L., Leach, N., Mahari, R., Pentland, A. S., Russakovsky, O., Schroeder, H., & Smith, A. Art and the science of generative AI. Science (American Association for the Advancement of Science), 2023, 380(6650): 1110–1111.

[15] Goertzel, B. Artificial general intelligence: concept, state of the art, and future prospects. Journal of Artificial General Intelligence, 2014, 5(1): 1.

[16] Sun Q. A Study of Legal Issues in Regulating Providers of Generative Artificial Intelligence Products. Politics and Law, 2023, (07): 162-176.

[17] WANG Kun Feng, Gou Chao, Duan Yan Jie, Lin Yi Lun, Zheng Xin Hu, WANG Fei Yue. Research Progress and Prospects of Generative Adversarial Network GAN. Journal of Automation, 2017, (03),321-332.

[18] Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., & Winther, O. Ladder variational autoencoders. Advances in neural information processing systems, 2016.

[19] Mirza, M., & Osindero, S. Conditional generative adversarial nets, 2014.

[20] Ray, P. P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, 2023.

[21] Goertzel, B. Artificial general intelligence: concept, state of the art, and future prospects. Journal of Artificial General Intelligence, 2014, 5(1): 1.

[22] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, 610-623.

[23] Mordvintsev, A., Olah, C., & Tyka, M. Inceptionism: Going deeper into neural networks, 2015.

[24] Daskalakis, C., Ilyas, A., Syrgkanis, V., & Zeng, H. Training gans with optimism, 2017.

[25] Gatys, L. A., Ecker, A. S., & Bethge, M. A neural algorithm of artistic style, 2015.

[26] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, 2017, 2223-2232.

[27] Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., ... & Olah, C. Multimodal neurons in artificial neural networks. Distill, 2021, 6(3): e30.

[28] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[29] Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., ... & Fidler, S. Get3d: A generative model of high quality 3d textured shapes learned from images. Advances In Neural Information Processing Systems, 2022, (35): 31841-31854.

[30] Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion, 2022.

[31] Peng, C. H. E. N., Qing, L. I., De-zheng, Z. H. A. N. G., Yu-hang, Y. A. N. G., Zheng, C. A. I., & Zi-yi, L. U. A survey of multimodal machine learning. Journal of Engineering Science, 2020, 42(5): 557-569.

[32] Baltrušaitis, T., Ahuja, C., & Morency, L. P. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2): 423-443.

[33] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," ACM Comput. Surv, 2015, vol. 47, Art. no. 43.

[34] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. Improving language understanding by generative pre-training, 2018.