Artificial Intelligence Final Report: A Research Study of AIGC

Yuwei XIAO
12012902
dept. CSE

I. What's AIGC

Artificial Intelligence (AI) has been under the spotlight for the recent years in the area of computer science and it has extended many important sub-areas such as Machine Learning, Natural Language Processing, and Computer Vision. Recently, there's one thriving idea in the AI area that attracted global attention, **Artificial-Generated Content (AIGC)**.

AIGC uses artificial intelligence technology to generate content. It's regarded as a brand new way of contetn creation after professional-generated content (PGC) and user-generated content (UGC). It is able to give full play to its technological advantages in creativity, expressiveness, iteration, dissemination, and individuation. Before 2021, AIGC generated mainly text, but the new generation of models can handle: text, speech, code, images, video, robot action, and so on. And in 2022, the AIGC developed at an astonishing pace. At the beginning of the year, it was still out of practice, but just in a few months it became professional enough to pass off the fake as the real thing. This phenomenon creates anxiety and tension among practitioners who have spent a lifetime learning to create. At the same time, the iteration speed of AIGC shows an exponential explosion, among which, the continuous improvement of the deep learning model, the promotion of the open source model, and the possibility of large model exploration and commercialization become the acceleration of the development of AIGC.

II. How AIGC Becomes popular

Fig. 1. Theatre d'Opera Spatial

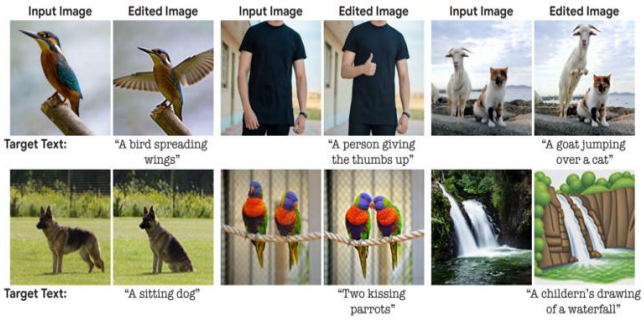
In August 2022, an AI-generated artwork titled *Theatre d'Opera Spatial* won first prize in the "Digital Arts/Digitally-Manipulated Photography" category of the Colorado State Fair fine arts competition. After the author, Jason M. Allen, claimed that his piece was simply generated by the AI art platform called Midjourney, which simply works by creating art pieces using keywords entered by users, this competition became a hot issue on the internet. On one hand, people thought that Allen's submission was not within the spirit of the competition because other entrants created their pieces by virtue of their own skills. On the other hand, the mass was shocked by the power of AI since the composition and style of that painting were so impressive, and it was indistinguishable from human work.

This type of AI is extremely astounding since, in our conventional thinking, machines cannot compete with people in the realm of creative endeavor. It excels at calculation and mining with a concentration on the study of large amounts of data. Humans excel at creative endeavors like poetry, design, programming, and other forms of innovation. Therefore, AIGC has a more obvious impact on the public compared to previous AI applications, such as AI playing chess, and AI predicting prices, because it allows people without professional skills to be capable of creating professional works just by simply inputting text.

With the increasing attention and investment in this emerging area, more and more innovations in AIGC come into our lives. The main ideas and applications in recent AIGC area are listed as follows:

- **Generate Text**
  A vivid example of this aspect is definitely the recent craze on the internet, ChatGPT, an AI-powered chatbot from Silicon Valley startup OpenAI, which is an unbelievable chat robot. It is trained using human feedback-based reinforcement learning, which uses human intervention to enhance machine learning to achieve more realistic results.
- **Generate Image**
  There are so many popular ideas and applications in this aspect. For example, the *DALL-E model* of *OpenAI*, *Stable Diffusion model* of *Stability AI*, and other models and applications that are based on them. We can input text to generate an image from nothing. Or, we can input text to change or add features in the given image.

(a)



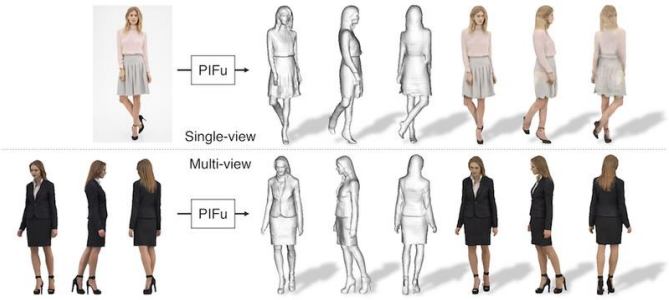
(b)

Fig. 2. Text to Image

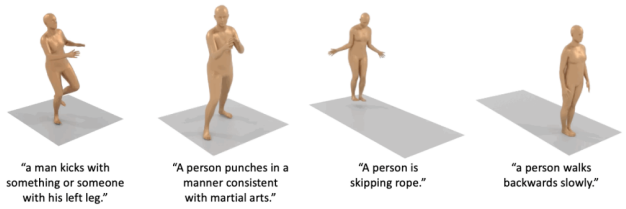
- **Generate Video**
  After the idea of text-to-image, the idea of text-to-video was soon proposed and implemented by *Google*, *Meta*, and other companies. On platforms like *Tik Tok*, the filters of portrait style transferring are widely used. And using tools like *Ebsynth*, you can bring your painting style to the video.

Fig. 3. Generating video by Ebsynth

- **Generate Model**
  Stepping out of 2-D, AI is also able to generate the 3-D model. *IMAvatar* generates 3-D face models by video or image, *PIFu* creates a 3D character model with clothing material by a single image and *MDM* directly drives the human body 3D model to do the corresponding action by inputting the text.



(a) PIFu



(b) MDM

Fig. 4. Generating 3-D Model

III. Stable Diffusion Model

Behind the rise of AIGC, the magical technology that turns text into the image is indispensable. There are two remarkable models that support this technology, *DALL-E*, and *Stable Diffusion*. The *Stable Diffusion* is a milestone in the development of AIGC. It enables the mass to use a high-performance model, the *DALL-E*, which has a good performance in image quality, speed and resource usage. So in this article, we will focus on the open-source, lightweight model, the **Stable Diffusion Model**.

A. Components

Stable Diffusion system consists of multiple components and models.

1) Text Encoder:

The very first component is Text Encoder. It can first capture the ideas in the text information, and turn it into a representation of numeric. Specifically, it's a special Transformer language model, which actually is the CLIP model's text encoder. Firstly it takes input text from user, then it produce a list which contains the numeric representation of the words in the text. Finally, it passes the information to the Image Information Creator.

2) Image Information Creator:

It is the core of the Stable Diffusion model. The *image information space* , as known as the *latent space*, is where it works, which makes it more efficient than the previous models, which has to work in pixel space. Specifically, it consists of a neural network named UNet and a algorithm, which is used to schedule. And this component is where "diffusion" happens. The information is processed step by step and finally leads to an image with high quality, which is produced by the Image Decoder.

3) Image Decoder:

The image decoder is an Autoencoder Decoder that paints a picture according to the information that the Image Information Creator passes to it. It only works once to generate the final image at the end of process.

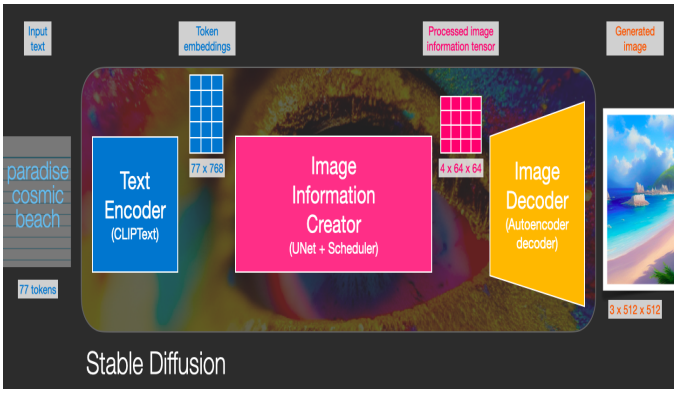


Fig. 5. Stable Diffusion Components

B. Diffusion

After knowing the components of Stable Diffusion, we should focus on how it works. The most important part of this model is *Diffusion*. Step by step, this process happens. It starts with a information array of a random image and the token embedding that represents the input text. Each step of this process adds more relevant information to the produced information array so that it can better resemble the input text.

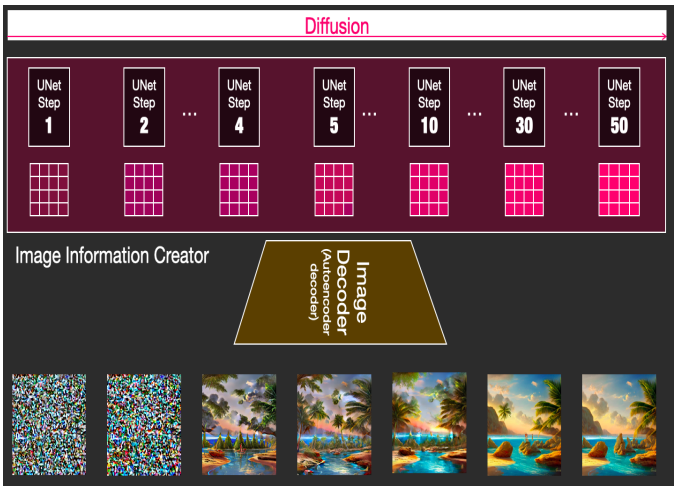


Fig. 6. Diffusion Steps

The main idea of producing images with models like Stable Diffusion depends on the powerful computer vision models that we already have. To be specific, we first create tens of training examples for each image in the data set by generating a random amount of noise and adding it to the image. And we use these training examples to train a noise predictor.

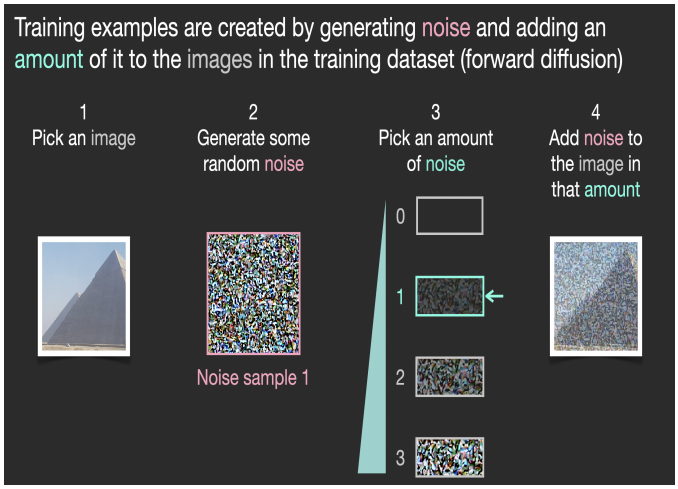


Fig. 7. Noise Predictor Training Example

By using this noise predictor, we predict the noise in a noisy image and subtract the noise in the image. In that case, we can get an image which is more alike the images that we used to train the model. If we use aesthetical images as the training dataset to train the model(e.g., LAION Aesthetics), then the generated image is likely to be aesthetically pleasing. Likewise, if we use images of logos as the training dataset to train the model, we would get a logo-generating model in the end.

C. Transformer Language Model

With the diffusion model only, we are not able to control the generated images. Therefore, we have to cooperate with the text encoder. A text understanding component is important in this process since it can take input text prompt and get its token embedding. A Transformer language model is what we use to represent the text understanding component. As the Imagen paper shown, it's important for us to choose a good language model. Comparing to larger image generation components, larger language models have a more significant effect on the quality of generated images. And since Stable Diffusion uses CLIP, we focus on it.

1) How CLIP is trained: CLIP consists of a text encoder and an image encoder. We use a dataset of images and their corresponding captions to train it. During the training process, we take an image and its corresponding caption, then encode them both with the picture and text encoders separately. From that point onward, we think about the subsequent installing utilizing cosine similitude. The comparability will be low toward the start of the preparation interaction, regardless of whether the text portrays the picture accurately. We update the two models so the following time we embed them, the result of embedding is similar.

We eventually have encoders that can produce embeddings where an image of a dog and the phrase "a picture of a dog" are similar by repeating this across the data set and with high batch sizes. Similar to word2vec, the model must assign low

similarity scores to negative samples of photos and captions that don't match throughout the training phase.

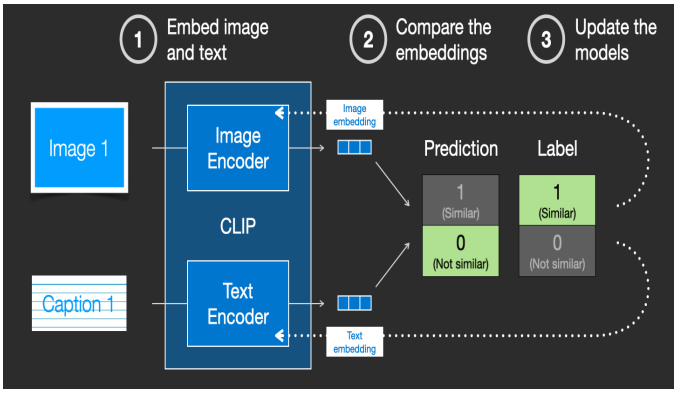


Fig. 8. Clip Training Steps

2) Cooperate With Diffusion: In order to let the text take part in the process of image generation, we need to adjust the noise predictor that we mention, to take the text as the input. Specifically, we add an attention layer between each original layer in our noise predictor model. The attention layers merge the text information with the image information.

D. Speed up Image Generation

Instead of using the actual pixel images to execute the diffusion process, the Stable Diffusion uses a compressed version of the image to speed up the image production process. Using an auto-encoder, this compression is carried out. The auto-encoder compresses the image using its encoder, places it in the latent space, and then uses its decoder to reconstruct it using only the compressed data.

The compressed latents are currently undergoing the forward diffusion process. The latents themselves are subject to the noise slices, not the pixel picture. In order to anticipate noise in the compressed representation, the noise predictor has been trained.

To create the data needed to train the noise predictor, we employ the forward process (using the auto-encoder's encoder). After training, we can create images by doing the reverse operation (using the auto-encoder's decoder).

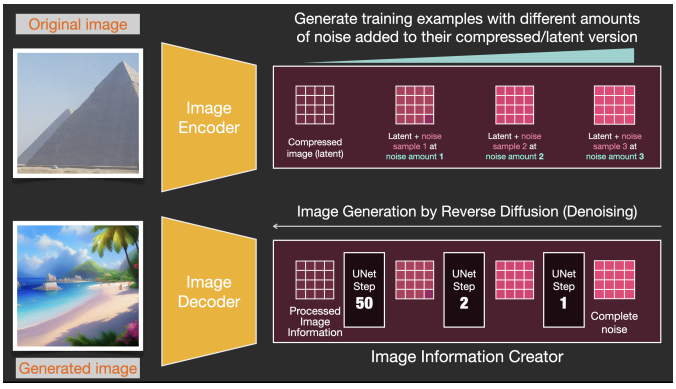


Fig. 9. Forward and Reverse process using auto-encoder

IV. ChatGPT Model

Another key factor in the rise of AIGC is ChatGPT, which draws millions of people's attention in the past few months because of its incredible intelligence. It is a deep learning model that produces human-like language using a transformer-based neural network, enabling the model to process input text and produce results similarly to how a human might.

A. Breakthroughs

Previous conversational AI systems fail to generate responses that are both relevant and coherent and are unable to keep the flow of the conversation unchanged and stop the conversation from becoming repetitive or stilted.

ChatGPT solved these two challenges. ChatGPT is able to understand the patterns of real conversation by employing a sizable conversational text dataset as training data, making it more likely to produce pertinent and cogent responses. Also, ChatGPT can generate multiple potential responses to a given prompt at the same time and make the best decision based on the context of the conversation.

B. How it works

Language models like ChatGPT use a decoder, where it is fed with the questions, produces some responses, and then uses the responses as inputs back into itself. However, the problem with this technique is that it can cause unpredictable and unintended behaviors, such as not following the user's prompt properly, making up facts, and generating biased text.

ChatGPT improves this technique by using Reinforcement Learning from Human Feedback (RLHF). That is to say, using human feedback in the training loop minimizes harmful, untruthful, and biased outputs. RLHF overall consists of three steps:

1) Supervised Fine-Tuning:

The first step consists in collecting demonstration data in order to train a supervised policy model, referred to as the Supervised Fine-Tuning (SFT) model. First, a list of prompts is sampled from the prompt dataset and human labelers are asked to demonstrate the expected responses. Then, the data is used to fine-tune a pre-trained language model.

2) Mimic Human Preferences:

The goal of this step is to learn an objective function (reward model) from the data. And the purpose of the reward model is to rate how desirable the SFT model outputs are for humans. Firstly, prompts and several model outputs are sampled. Then, human labelers will rank the outputs from best to worst. After that, the data is used to train the reward model. So after all training processes are done, an automatic system, which is supposed to mimic human preferences, is going to be extracted from the data.

3) Fine-tuning via Proximal Policy Optimization:

First of all, Proximal Policy Optimization (PPO) is an algorithm used to train reinforcement learning agents. It is an "on-policy" algorithm that learns from and updates the current policy directly, rather than learning from past experiences. It also uses a value function to estimate the expected return of a

given state, and the value function is used to compute the advantage function, which represents the difference between the expected return and the current return. Then the difference in advantage function is used to update the policy by comparing the actions taken by current and previous policies.

In this step, the PPO model is initialized from the SFT model, and the value function is initialized from the reward model. So, after a prompt is sampled, the policy generates an output, and the reward model calculates a reward for the output, then uses that reward to update the policy using PPO.

V. Pros and Cons of AIGC

Every coin has two sides, and AIGC is no exception. Here is a list of the advantages and disadvantages of AI-Generated Content.

A. Pros

1) Create Content Immediately:

With the state of art models and tools, AI is able to generate text, images, or videos within a small amount of time. This can give you a headstart on content creation. In this case, you don't need to worry about researching and brainstorming for content ideas, you just need to provide AI with a topic. Then, you can start your work by modifying the AI-Generated Content.

2) Cost Efficient:

Since gaining substantial content with AIGC is easier, it saves time, effort, and resources. AIGC can make your regular working process faster and more insightful. In that case, you can use the time and resources saved by AIGC to focus on more critical work. Although the costs are reduced, you still get the best of the AIGC's work. Therefore, you can gain not only rich but also high-quality content using AIGC.

3) Personalization:

The development of AIGC gives everyone the ability to create. You can turn any thought or idea that is in your head into reality in just a few seconds. And the result is unique. Also, the extended functionality of AI means that you are able to create content in your style or tailor certain content for different situations.

B. Cons

1) Unnatural Content:

The AI content generator will create content that doesn't always sound like a human. What's worse, the content may read as incomprehensible. As well as the generated images, which sometimes are totally irrelevant to the text given.

2) Too Much Human Involvement:

Whether the process of training the models or the process of generating qualified content, it needs human involvement. The process of Reinforcement Learning needs human feedback to tune the AI, and when you use AI to generate content, you need to continually adjust the text you input and modify the generated content to fit the requirements. So, the AIGC isn't good enough yet to finish the whole task on its own, and that's what researchers are going to work on.

3) No New or Neutral Ideas:

Firstly, AIGC is generated from existing content, which means it can only recycle the points that it analyzed, no new ideas will be generated. Secondly, flawed algorithms may generate biased content that favors a particular stance. Therefore, the generated content may have points that don't align with your initial goals.

4) No emotion or depth: Without doubts, AI-Generated Content is informative and substantial. However, you can't expect the content to empathize with a character, experience, or other aspects of the text. It can provide opinions and exposition about the data it's studied. But it can't provide in-depth views. So, it's unable to be subjective, nor can it make an argument for its views.

VI. Concerns about AIGC

The contest-winning AIGC painting we mention has aroused wide public concern. And with the rapid development of AIGC and its wide application, more and more concerns are aroused.

A. Creativity

As we mention, AIGC is not able to create new ideas since the content is generated from existing content. However, whether it is creative still remains controversial.

Some insist that the result of the creation is the most important part, while others regard the intention and way of the creation as the most important part. According to Margaret Boden, a prominent artificial intelligence researcher and philosopher at the University of Sussex in the UK, creativity can be defined by three main factors: An idea or item must be original, unexpected, and valuable to qualify as creative. Beyond that, you can instantly recognize originality when you see it. When using a computer to create something, researchers in the subject of "computational creativity" say their work qualifies as creative if it can be assumed that only humans could have produced it.

The text-to-image model does indeed go beyond prior definitions, but Maria Teresa Llano, a computational creativity researcher at Monash University in Australia, does not consider it to be creative. She draws attention to the fact that when consumers use these tools frequently, the outcomes can seem to look the same. That indicates that they fall short of any or all of the criteria for innovation, which could indicate a fundamental flaw in the technology.

In fact, the text-to-image algorithm creates images based on the billions of existing photos. It's possible that machine learning will only create (or mimic) images that it has already seen. It might not matter for computer graphics. Text-to-image creation has been added to Photoshop by Adobe; Blender, an open-source alternative to Photoshop, already has a Stable Diffusion plugin; and OpenAI is collaborating with Microsoft to create a text-to-image widget for the Office productivity suite. Users will truly feel the impact of robots that don't replace human ingenuity but rather augment it in this interaction, in future iterations of these well-known productivity products.

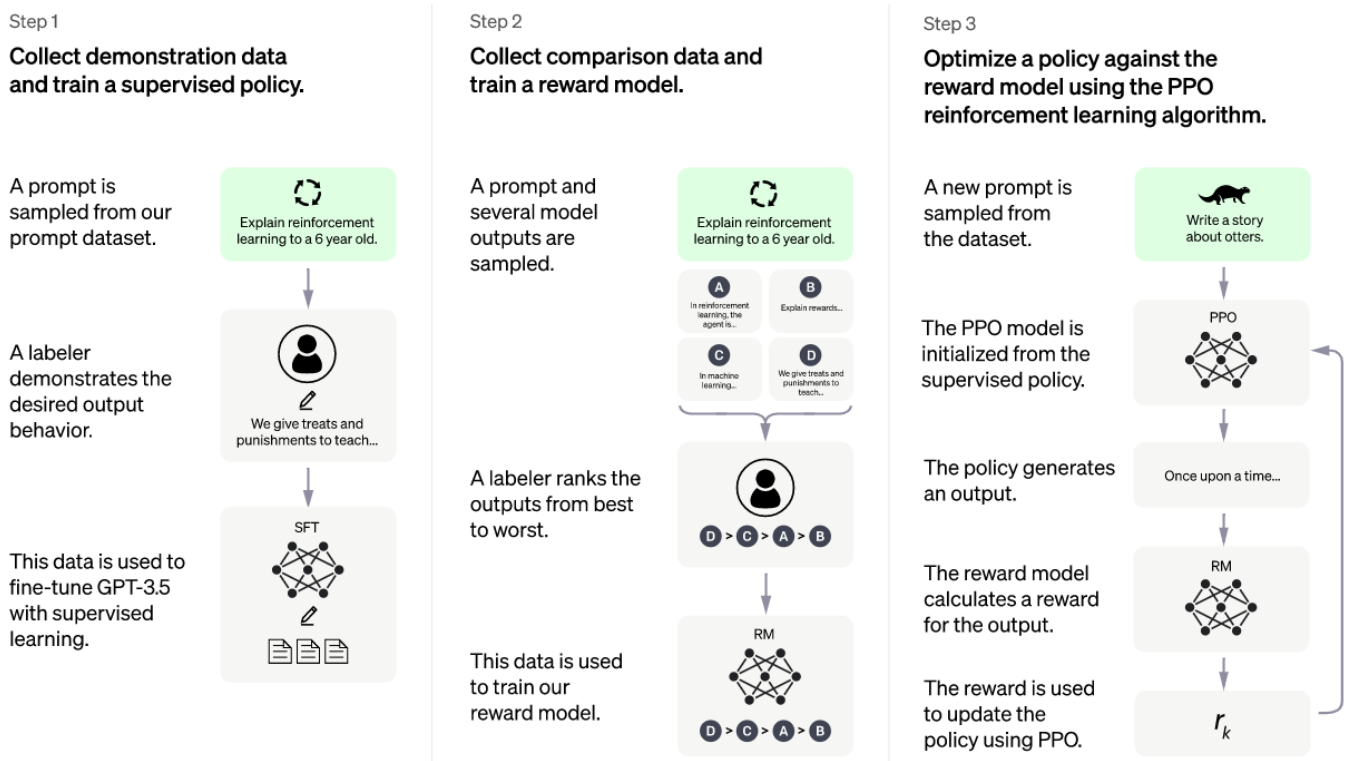


Fig. 10. ChatGPT's workflow

According to Lallano, "the creativity we see today comes from the application of systems, not from the systems themselves." Other researchers that study computational creativity share this viewpoint. It's more important to consider how machines operate than merely what they perform. Giving them creative responsibilities, encouraging them to be more autonomous, and allowing them to plan and produce are all necessary to transform them into true creative partners.

B. Ethics

AI ethics have been discussed over the past few years. But the ethical discussion about generative AI is relatively new. It has been accelerated by the popularization of generative models. Below we discuss some prevalent ethical concerns.

1) Reliability:

AIGC approaches such as deepfake, can be used to generate synthetic media. This kind of AIGC can be difficult to distinguish from the real one. And that will pose serious ethical implications, such as spreading misinformation and manipulating public opinion. What's worse, AIGC may be used to harass or defame individuals by creating fake images or videos that depict them in a negative or embarrassing light.

2) Truthfulness and Accuracy:

AIGC uses the machine learning method to infer information, which would possibly bring the inaccuracy problem to acknowledge. And the pre-trained large language models, e.g. ChatGPT, are not keeping up with new information. Although language models have become more persuasive and eloquent, their proficiency has been utilized to propagate inaccurate details or lies. They can make up convincing conspiracy theories that may cause harm and spread superstitious information.

3) Copyright:

The ambiguities over authorship and copyright of AIGC are another ethical concern. This determines who the rights of the created works belong to and how they can be used. This concern is focused mainly on 3 questions:

- Should AIGC be eligible for copyright protection?
- Who should have the rights over the created content?
- Can copyrighted data be used for training purposes?

For the first question, one answer is AIGC is not the product of human creativity so that it is not eligible for copyright protection. However, others argue that AIGC is the product of complex algorithms and programming that are made by humans. In that case, it should be eligible for copyright protection.

For the second question, we all know that users can enter the name of a specific artist and have the AI create a replica of their style using AIGC. And if not mentioned, it's hard to tell the difference. Although it's the model who creates this painting by copying the style of the painter, people familiar with the style of the painter can assume that this painting is one of his/her works. Given this. we should decide if it's ethical for AI to generate art or other content that is similar to others' work. Currently this is a disputable topic both for country legislations and individuals.

For the third question, it's acceptable of using generated date to train machine learning models. But the use of copyrighted generated data in compliance with fair use doctrine is ambiguous. Fair use generally accepts academic and nonprofit purposes and forbids commercial purposes. For example, Stability AI doesn't directly use such generated data. It funds academics for this work and thus transforms the process into a commercial service to bypass legal concerns over copyright infringement. But recently, GitHub released an AI programmer named Copilot, which can helps user write better code by suggesting a few lines of code or the functions. However, it draws developers hate. The major reason why developers are criticizing it is this: Microsoft, a company that owns GitHub, has access to all the repositories. Microsoft took an advantage of the GPL license, which is a license for open-source software which perimits the use of software in other projects without any copyright restrictions, to train an AI model using all the public repositories and, in the end, charging a subscription fee for others to use it. In this case, Microsoft uses developers' trust of GitHub to use the repositories to train the Copilot for their own good and gives nothing to developers in return. In fact, it even charges developers for using the AI that is trained by their own code.

C. Increase of Biases

Recent research suggests systems that are larger and more sophisticated are likely to absorb underlying social biases from the training data. These biases include racist, sexist, or ableist approaches.

D. Misuse

Generally speaking, AI could produce misleading or harmful content in any context.

In the education context, AIGC could be misused by students to prepare their homework on a wide variety of topics. And it could be misused by generating misleading information that possibly leads to students being misinformed. In addition, it can be misused to generate material that is ideologically biased and factually incorrect.

In the marketing context, AIGC can be misused for unethical business practices. For example, manipulating online reviews or mass-creating accounts with false identities.

In the societal context, AIGC can be misused to generate convincing social attacks, such as phone calls or phishing emails. These attacks are designed to trick individuals into revealing sensitive information, such as financial information, login credentials, or malware downloading.

VII. Future of AIGC

In particular for images and videos, the AIGC generation is moving away from "cutting costs and boosting efficiency" and toward "generating value." The research paper states that the domains where the underlying technology is evident and anticipated to be widely used in 1-2 years are mostly text picture generation, creative function, and functional fields.

The three main infrastructures of the metacoverse and Web3.0 are the AIGC, NFT, and VR / AR. Writing, editing, painting, and video production are examples of creative industries that AI is steadily encroaching onto as data accumulates, processing power advances, and algorithm iteration rises.

So what can AIGC do in the future?

- With manufacturing capabilities and knowledge levels superior to humans, AIGC can perform fundamental mechanical tasks like information mining, material calling, and copy editing. It can also meet a lot of customized needs at a low marginal cost and high efficiency.
- By encouraging the multi-dimensional interaction, integration, and penetration of digital content and other industries, AIGC can promote new business forms and new models.
- Contribute to the creation of the "metaverse." Infinite content generation and accelerated physical world reproduction are made possible by AIGC, leading to spontaneous organic expansion.

Specifically, there are several areas:

- AIGC + media: writing robot, interview assistant, video subtitle generation, voice broadcast, video collection, artificial intelligence synthesis anchor
- AIGC + e-commerce: commodity 3D model, virtual anchor, virtual goods yard
- AIGC + film: AI script creation, AI synthesis face and voice, AI creation role and scene, AI automatic generation of film and television trailer.
- AIGC + entertainment: AI face changing applications (such as FaceAPP, ZAO), AI composition (such as Hatsune Future virtual song), AI synthetic audio and video animation
- AIGC + education: AI synthetic virtual teachers, AI make historical figures according to textbooks, AI convert 2D textbooks into 3D
- AIGC + finance: AIGC automates the creation of financial information and product introduction videos, and it shapes virtual digital people's customer service.
- AIGC + medical treatment; AIGC creates synthetic body projections for people with disabilities, linguistic audio for people with voice loss, and medical companionship for people with mental illness.
- AIGC + industry: Using AIGC, finish off repetitive low-level engineering design chores, produce derivative designs, and motivate engineers.

The explosion of AIGC has undoubtedly received a lot of attention, and huge models' commercialization potential is becoming more and more obvious. Large model businesses can carry out commercial transformation, "offer services on demand" in accordance with the actual demands of B and C-end consumers, and use cloud computing and cloud storage more extensively. AIGC is anticipated to open up new avenues for the investigation of long-term value and the commercialization of large models.

The biggest companies in the world are also attracted by the rise of AIGC and its wide application.

Google opened the first test of their own text picture model Imagen and released Imagen Video, an AI-generated video model. Since its release, Imagen has been likened to Stability AI's Stable Diffusion and OpenAI's DALL-E 2; nevertheless, Google has been careful about not making the model available to the general public. In order to get early input on the technology, Google has now revealed that it will be integrating Imagen into its AI Test Kitchen app. In addition, Google has made available four of its most recent AIGC technologies, which may be used to generate text, code, music, extended films in high quality, and more based on text commands.

According to Robin Li, the founder and CEO of Baidu, AI has advanced significantly over the past year in terms of both technology and practical applications, with some even changing course. AIGC is one of them, and it represents a shift in artificial intelligence technology. The "advance force" behind the Baidu mobile ecological AIGC application is Baijia. Baijia has introduced a digital host in addition to the AI painting feature and AIGC graphic to video technology in order to give creators more scene application experience. It is capable of 24-hour automatic broadcast, multilingual broadcast, sign language hosting, and other responsibilities in addition to automatically broadcasting news.

The WIMI Hologram Cloud has AIGC technology reserves, has long adhered to the integration of AI and AR to develop technical capabilities for the future, developed AI face recognition and AR live broadcasting technology, gradually investigated and verified AR remote collaboration, and made strides in focused research and development. The business will continue to concentrate on market trends and cutting-edge technology in the future and will offer more tools and content support for user creation in the Web 3.0 age. It is important to note that the AIGC's capabilities rely heavily on artificial intelligence technologies, including big models that have already been developed, deep learning platforms, and AI chips. In the area of artificial intelligence, WIMI established the holographic academy of sciences, which, at the algorithm level, gives the AI algorithm a new role that isn't just command and control but also more image structure and big data analysis. This enhances the platform's overall efficiency and aids in the development of digital content creation engines and universe scene development tools.

Alibaba has also introduced its first AI model community in Chinese with the goal of lowering the entry barrier for AI applications and creating the "infrastructure" for the AI era. The magic build community ModelScope was launched under the direction of Alibaba's Alidamo Institute. In the beginning, more than 300 models, including more than 100 Chinese models, were released to the public. These models covered more than 60 common tasks in more than 60 major AI fields, including vision, speech, natural language processing, and multimodal. All of these models are completely open-source and free to use.

VIII. Conclusion

In this essay, we first introduce the AIGC and explain why it's so popular. Then, we take a look at the wide and amazing applications of AIGC. After that, we focus on two specific and essential models, Stable Diffusion and ChatGPT. We explain what these models can do and how them work. Moreover, we analyze AIGC by its advantages and disadvantages. In addition, we discuss some concerns about AIGC, such as the ethical problems and the possible misuses. Finally, we have a glance at the future development of AIGC and the pioneers in this area.

From all above, we can easily conclude that AIGC is an amazing technology. It's able to turn a incredible thought or scene in our mind to a concrete image or story within seconds. This experience is a dream to everyone, especially the ones who works in content-creating field, it enables them to verify their inspirations and get a prototype rapidly.

Although the current methods of AIGC still have some shortcomings and AIGC still faces some public concerns, it's obvious that AIGC has a bright future. And I believe in the near future, with more attention and investment on this area, AIGC will become more powerful and it will be applicated in more areas, such as business, medical treatment, and so on. Meanwhile, the society should be more open but cautious to this technology. Law should be made to restrict or ban the inappropriate usage of AIGC. Ahd public should be aware of the copyright, as well as the ethical implications.