Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/pj5g7k56

Limitations of AI-Generated Content and Strategies for Enhancement

Weihang Gao*

School of Intelligent Systems Science and Engineering/JNU-Industry School of Artificial Intelligence, Jinan University, Zhuhai, Guangdong, 519000, China

**Corresponding author: Weihang Gao*

Abstract

With the rapid advancement of artificial intelligence (AI) technologies, tools such as Doubao and DeepSeek have become indispensable elements in everyday life. Many users turn to these AI platforms for secondary editing of images and videos or to generate visuals that align with specific textual descriptions. However, the outcomes frequently fall short of expectations: the generated content often bears little resemblance to the user's prompts and is riddled with unacceptable flaws. To mitigate these issues and enable AI tools to produce satisfying results that meet given specifications, this paper explores the underlying principles of AI content generation. It proposes methods to enhance the accuracy of outputs and reduce inherent defects, concluding with a discussion of the study's limitations.

Keywords

AI-generated content (AIGC), limitations, enhancement strategies

1. Introduction

In the current era, artificial intelligence (AI) applications such as Doubao, along with built-in AI features in platforms like Douyin and Baidu, commonly offer capabilities for generating text, images, or videos. Users simply need to provide a textual description of their requirements and upload relevant reference materials, and the AI can produce the corresponding content. Given the remarkable convenience, efficiency, and versatility of using AI tools for image creation, individuals across various fields increasingly rely on them to assist in completing creative tasks. In academic settings, for example, students often turn to AI to help with homework assignments. In the realm of business and economics, people leverage these tools to design promotional posters for products or events, which they then display in public spaces or share on social media. For leisure and entertainment, users generate a wide array of customized images that suit their preferences, leading to the emergence of numerous artists on platforms like Pixiv and X who specialize in creating and sharing AI-generated content (AIGC), particularly in the style of anime or manga illustrations.

Nevertheless, the vast majority of users express significant dissatisfaction with the outputs when employing these AI tools, as AIGC frequently exhibits substantial flaws. For instance, when users request a generated text based on a description, the resulting content may bear no relation to the prompt. Similarly, in creating an anime

character, the produced image might feature distorted limbs, extra or missing fingers, or attributes that deviate entirely from the user's specifications. In cases where the image should incorporate text, the embedded words often appear as nonsensical gibberish that belongs to no recognizable language. These issues persist widely across various AI tools today, and this paper primarily aims to propose solutions for mitigating them by examining the underlying principles of AIGC generation.

2. Principles of AIGC Generation

To minimize the shortcomings and defects in AI-generated content (AIGC) and produce higher-quality results, one must first grasp the category of the AI tool in question and its foundational mechanisms.

2.1 Generative Artificial Intelligence

All the AI tools discussed earlier belong to the subcategory known as generative artificial intelligence (Generative AI, or GenAI). This represents a class of AI models rooted in generative adversarial networks (GANs) [1], capable of interpreting user inputs like text or voice as directives, incorporating supplied references such as images or videos, and independently producing entirely novel content—including text, images, audio, videos, and code—that did not exist before [2]. Given that the majority of GenAI systems can generate text, their operational principles can be extrapolated from the models outlined below, and thus will not be elaborated upon separately here. What follows is an overview of two prevalent GenAI models designed specifically for image generation.

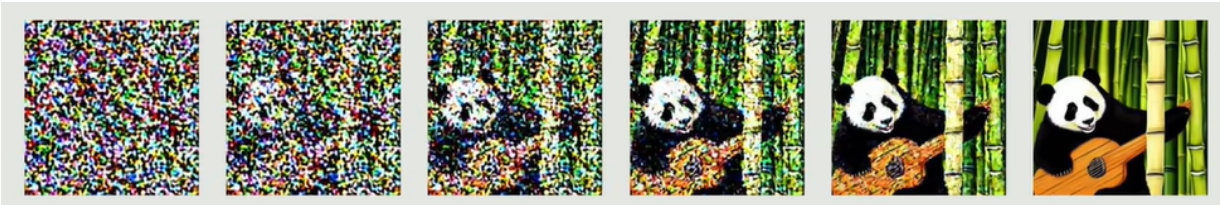
2.1.1 Text-to-Image Models

Text-to-image models represent the most fundamental approach to generating images using artificial intelligence. In this method, users provide the AI tool with a descriptive text prompt, and the AI then interprets this input to produce an image that aligns with the description.

The underlying mechanism of text-to-image models operates as follows: upon receiving a textual input from the user, the AI employs a Transformer-based architecture to tokenize the text. It subsequently utilizes a text encoder to convert the tokenized text into a vector representation, which is fed into a diffusion model. Within the diffusion model, the text vector serves as a conditioning guide to progressively denoise a noisy initial image, as illustrated in Figure 1. Through multiple iterations, this process yields a clear image. Finally, an image decoder scales the resulting image to a standard resolution. In this manner, the AI accomplishes the text-to-image generation task.

The training and inference processes of diffusion models in pixel space entail exceptionally high computational complexity. To address this challenge, latent diffusion models (LDMs) incorporate an encoder that compresses images from the pixel domain into latent variables. The diffusion model's training and inference then occur within this latent space. Once the latent variables are generated, a decoder reconstructs them back into images. This approach significantly reduces the computational cost of the diffusion model [3].

Figure 1: The image denoising process (sourced from BiliBili)



2.1.2 Multimodal Models

Another frequently utilized model is the "multimodal" model. This paradigm involves users providing AI tools with data encompassing various information formats, such as text, visual images, video, and audio. The AI then generates content tailored to specific requirements based on this diverse input. Compared to text-to-image models, multimodal models typically produce higher quality content. Their adoption has garnered

increasing attention from users in recent years, driven by growing societal demands and the enrichment of information data collection methods [4].

The underlying principle of multimodal models is as follows: image encoders, such as Convolutional Neural Networks (CNNs) or Vision Transformers (ViT), transform visual content like images and videos into sets of vectors. Subsequently, an alignment module performs "modal alignment" between these image and video vectors and text vectors. This process involves calculating a contrastive loss function to minimize the distance between paired vectors and maximize the distance between unpaired vectors, thereby enhancing the semantic congruence between text and images [5]. If the desired output is text, the corresponding text vectors are directly generated by a text decoder. Conversely, if the generation of images, videos, or other visual content is required, these vectors are then fed into a diffusion model, and the subsequent process mirrors that of text-to-image models.

Notably, modal alignment necessitates an extensive volume of data for training. For instance, the multimodal perception model CLIP (Contrastive Language-Image Pre-training) achieved text-image alignment through contrastive learning, having been trained on over 400 million pairs of text and image data [6].

2.2 Reasons for Deficiencies in AIGC

The models discussed above often require an extensive number of training iterations before they can be deployed. However, in many domains, it is frequently challenging to collect sufficient material for comprehensive model training. This scarcity of data often results in significant deficiencies when Generative AI (GenAI) attempts to produce content in these areas. The following sections will discuss several common GenAI shortcomings and their underlying causes.

2.2.1 Low Relevance in Text-Based AIGC

A common defect arises when users request AI to generate text, and the resulting output bears no relation to the expectation, or when an AI provides an entirely incorrect answer to a posed question. This deficiency primarily stems from the following reasons:

Ambiguous Reference Material. Consider, for instance, submitting an image like Figure 2 to an AI and asking, "How many numbers are in this picture?" In this scenario, the AI might offer three possible answers: one (123), two (1 and 23), or three (1, 2, 3). Presenting such an ambiguous image to human respondents would similarly elicit a variety of answers from different individuals.

Figure 2: A Set of Numbers (Author-created)

Ambiguous Text Content Provided by the User. To illustrate this further with Figure 2, even if a user explicitly informs the AI that the image contains three distinct numbers, "1," "2," and "3", and subsequently asks, "Which number is the smallest?", the AI might still furnish two potential answers: "1" or "2." The ambiguity in this scenario resides in the phrase "which number is the smallest." Does this question intend a comparison of the inherent numerical values or the physical font size of the digits? Such interpretational discrepancies can frequently occur when the AI processes and comprehends user queries.

2.2.2 Deficiencies in Image-Based AIGC

When users require AI to generate an image, the resulting content may bear little resemblance to the user's textual description, a problem analogous to the low relevance issues observed in text-based AIGC. Beyond this, the most significant challenge with image-type AIGC is often syntactic distortion or malformation within the generated visuals. The following elaborates on this phenomenon.

2.2.2.1 Deformed Human Figures

When users employ AI to generate images of human figures, the resulting pictures frequently exhibit several characteristic flaws: deformed fingers and toes, or an incongruence between the number of heads and bodies. These are among the most unacceptable defects for users, and their root causes are discussed below:

(1) Certain human body parts inherently possess extremely complex structures. Consider the hand, for example: an adult hand comprises 17 bones and 19 joints, allowing for a multitude of possible poses. When these poses are captured in images, some fingers may be obscured; for instance, a "V" sign gesture only clearly displays two fingers. Providing AI with a diverse set of images featuring varying numbers of visible fingers for training significantly increases the probability of the AI being misled.

(2) Insufficient training data. Humans often lack detailed descriptions of hands. Regardless of the pose, a hand is generally referred to broadly as "a hand," which provides insufficient nuanced information for AI training. These combined factors make it challenging for AI to generate a complete and accurate hand.

The "Uncanny Valley" effect leads to zero tolerance for "deformity." The "Uncanny Valley" refers to the psychological phenomenon where humans experience feelings of unease or revulsion towards anthropomorphic objects that are highly realistic but not perfectly so [7]. Taking Figure 3 as an example, we generally feel no discomfort when observing the anime character on the left. However, upon seeing the realistic human rendering on the right, which is based on the same character, we experience an inexplicable sense of unease. Should the hand of the figure on the right exhibit any deformity, this discomfort would only intensify. Conversely, AI machines do not possess human "aesthetics"; without human intervention and processing, Generative AI cannot refine image details according to human cognitive logic based solely on its inherent algorithms [6].

Figure 3: Anime Character Image from Cygames

2.2.2.2 Textual Malformations

When a user requests that an AI generate an image containing text, it is highly probable that the text within the image will appear in an unknown or illegible script. The reasons for this deficiency are as follows:

Text possesses both "image" characteristics and semantic meaning. For instance, if an AI is asked to generate an image of "apple", it faces a dilemma: whether to generate an image of the fruit itself or to render the two Chinese characters for "apple." Compounding this, AI tools generally receive less specific training on text as an image component, making it challenging for them to produce coherent or normal-looking text.

The strokes and structure of characters are complex, and typefaces are diverse. This inherent complexity often leads AI, when generating textual images, to merely arrange strokes randomly. Consequently, textual malformations are a common outcome.

3. Methods for Reducing AIGC Deficiencies

By analyzing the operational principles of AI tools and the underlying causes of AIGC defects, we can now propose the following actionable solutions.

3.1 The AIM Methodology

"AIM" stands for "Actor," "Information," and "Mission." The "AIM methodology" requires users to communicate their role (Actor), the necessary background information and data for completing a task (Information), and the specific task instructions (Mission) to the AI.

In most scenarios, the "Actor" component can be omitted. However, "Information" holds a critically important position, as it is key to the quality of AIGC. Firstly, the information provided by the user must avoid ambiguity, similar to the example illustrated in Figure 2. Secondly, the more comprehensive and detailed the user's content, the higher the quality of the AIGC and the fewer the defects.

Furthermore, if a user requests AI to generate an image, in addition to providing detailed content, it is crucial to emphasize complex drawing aspects. For example, if specific clothing, accessories, or a particular artistic style needs to be retained for a character, repeated emphasis by the user will significantly reduce the probability of AI errors. If the AI-generated image must include text, one simply needs to instruct the AI to output the required text as a graphical element.

Finally, when issuing task instructions, it is equally important to express requirements clearly, ensuring that no textual ambiguities are present.

3.2 Reshaping Thought Processes

If users find it challenging to provide all detailed information to the AI in a single input, they might consider reshaping the AI's thought process. Below are three methods for achieving this.

3.2.1 Chain-of-Thought Pattern

Users can simply issue a command to the AI, such as "show your thought process." This prompts the AI to reveal the logic and reasoning behind its generated results, allowing users to quickly identify the root cause of any deficiencies. This pattern is particularly applicable to the generation of textual content.

3.2.2 Validator Pattern

In this pattern, the user instructs the AI tool: "Before generating content, ask me a few questions that will help you accurately understand my requirements." After the AI poses its questions, the user can then further refine and enrich the information they initially provided based on these prompts.

3.2.3 Optimization Pattern

Prior to content generation, users can ask the AI to formulate several more precise expressions, derived from the initial information provided, which are more likely to yield high-quality results. The user then selects from these options and can further refine them. This process optimizes the information supplied by the user, thereby reducing the probability of AIGC deficiencies.

4. Conclusion

This paper first systematically outlined the fundamental principles of two categories of Generative AI (GenAI): “text-to-image models” and “multimodal models.” It highlighted that both are fundamentally based on Generative Adversarial Networks (GANs), utilizing encoders to transform source material into vectors, and subsequently employing diffusion models and decoders to complete content generation and reconstruction.

The paper then identified that current AI technology still exhibits shortcomings in semantic alignment for both text and images. Generated images frequently suffer from structural deformities and detail distortions. The underlying causes for these issues can primarily be attributed to four aspects: the ambiguity of user input, the complexity of generation tasks, the limitations of training data, and the AI’s insufficient capacity to handle human perceptual boundaries such as the “Uncanny Valley effect.”

The innovation of this paper lies in its novel approach from the user’s operational perspective, systematically categorizing the causes of AIGC deficiencies. It further proposes two solutions: the “AIM methodology” and “reshaping thought processes,” providing a theoretical reference for subsequent optimizations.

5. Limitations and Future Work

Despite providing a relatively comprehensive analysis of AIGC deficiencies, this paper acknowledges certain limitations. Firstly, it exclusively offers methods to reduce AIGC defects and enhance quality from the user’s operational standpoint, without delving into more refined and feasible measures for improving GenAI from the perspective of its underlying generation principles. Secondly, the paper does not deeply explore different model architectures (e.g., comparing diffusion models with autoregressive models) and lacks a systematic evaluation of content quality control mechanisms. Furthermore, while the proposed “AIM methodology” and “reshaping thought processes” offer practical guidance, their effectiveness and universality warrant further empirical validation through subsequent research.

In this era of rapid technological advancement, AI researchers and practitioners must continue to investigate, discover, or develop more sophisticated AI models to address existing defects and shortcomings. This will enable AI tools to offer even greater convenience in people’s lives. Future research could unfold in several directions: first, exploring more efficient prompt optimization strategies; second, promoting the co-evolution of multimodal models in semantic understanding and image generation consistency; and third, integrating user behavior data to construct intelligent generation quality feedback mechanisms, thereby further enhancing the application value and user experience of AIGC.

References

[1] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. Generative adversarial nets. In Advances in neural information processing systems, Montreal, QC, Canada, 2014; p. 2672−2680.

[2] Miao, F. C. Examination of the Technique Principle of Generative AI and Its Educational Applicability. Modern Educational Technology. 2023, 33(11), pp. 5-18.

[3] Zhang, L. Y., Yang, S., Wang, W. J., Gao, X. and Liu, J. Y. AIGC-Based Image and Video Generation Method: A Review. Journal of Computer-Aided Design & Computer Graphics. 2025, 37(3), pp. 361-384.

[4] Zeng, C., Ge, Y. J., Zhao, L. C. and Wang, Q. Survey of Multimodal Vision-Language Representation Learning Models and Their Adversarial Examples Attack and Defense Techniques. Journal of Computer Research and Development. 2025, 62(09), pp. 2208-2232.

[5] Shao, S. Y., Du, Y. and Fan, X. L. Non-Autoregressive Sign Language Translation Technology Based on Transformer and Multimodal Alignment. Journal of Electronics & Information Technology. 2024, 46(7), pp. 2932-2941.

[6] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P. and Clark, J. Learning transferable visual models from natural language supervision. In International conference on machine learning, Virtual Conference, 2021; pp. 8748-8763.

[7] Jiang, L. Interpreting and Addressing the “Uncanny Valley Effect” in AI-Generated Art. Journal of Nanjing University of the Arts(Fine Arts & Design). 2025(4), pp. 187-191.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access