AI Image Generation Evaluation Results Released: ByteDance and Baidu Perform Well, DeepSeek Janus-Pro Falls Short

Zhenhui Jack JIANG[1], Zhenyu WU[1], Jiaxin LI[1], Haozhe XU[2], Yifan WU[1], Yi LU[1]

[1]HKU Business School, the University of Hong Kong, [2] School of Management, Xi'an Jiaotong University

Abstract

The frontier of AI models has evolved beyond text processing to encompass the ability to understand and generate visual content. These models not only comprehend images but also generate visual content based on textual prompts. This study presents a systematic evaluation of the image generation capabilities of AI models, focusing on two core tasks: generating new images and revising existing images. Using carefully curated multidimensional test sets, we conducted a comprehensive evaluation of 22 AI models with image generation capabilities, including 15 text-to-image models and 7 multimodal large language models. The results show that ByteDance's Dreamina and Doubao, as well as Baidu's ERNIE Bot, demonstrate impressive performance in both new image generation and image revision tasks. Overall, multimodal large language models deliver superior performance compared to text-to-image models.

Background and Contributions

Generative artificial intelligence (GenAI) is undergoing a pivotal transformation, expanding rapidly to integrate multimodal capabilities, particularly in image understanding and generation. For image understanding, vision-language models such as Qwen-VL and multimodal large language models (LLMs) such as GPT-4o have demonstrated remarkable performance in visual perception and reasoning tasks. Our team previously published a report on evaluating the image understanding capabilities of LLMs (please scan the QR code in Figure 1 to access the report). This study builds upon and complements our previous work, and the two studies collectively form a comprehensive evaluation framework for multimodal artificial intelligence.

Figure 1. Comprehensive Evaluation Report on the Image Understanding Capabilities of Large Language Models

(or visit https://mp.weixin.qq.com/s/kdHRIwoVO79T9moFcX1hlQ)

In the realm of image generation, text-to-image models — especially those based on diffusion models like DALL-E 3 — as well as multimodal LLMs with image generation capabilities such as ERNIE Bot, have significantly propelled AI-driven creativity. With their image generation capabilities and versatile applications, these models are transforming traditional fields including content creation, marketing, and graphic design, while unlocking new possibilities in emerging industries.

Despite these advancements, the evaluation of AI image generation models remains in its early stages. Current ranking systems, such as SuperCLUE and Artificial Analysis, rely primarily on algorithmic evaluation, LLM-as-a-judge, or model arenas. However, these approaches are often prone to biases, unfairness, and a lack of transparency, while neglecting safety and responsibility concerns. To address these challenges, we developed a systematic evaluation framework for assessing the image generation capabilities of AI models. This framework helps users make informed decisions among models and provides developers with insights for optimization and improvement. Our evaluation encompasses 15 text-to-image models and 7 multimodal large language models (see Table 1).

Table 1. List of Models Evaluated

Nation	Model Type	Model	Institution
China	Text-to-image Model	360 Zhihui (智绘)	360
China	Text-to-image Model	TongYiWanXiang (通义万相) Wanx-v2	Alibaba
China	Text-to-image Model	Wenxin Yige (文心一格) 2	Baidu
China	Text-to-image Model	Dreamina	ByteDance
China	Text-to-image Model	DeepSeek Janus-Pro	DeepSeek
China	Text-to-image Model	SenseMirage V5.0	SenseTime
China	Text-to-image Model	Hunyuan-DiT	Tencent
China	Text-to-image Model	MiaoBiShengHua (妙笔生画)	Vivo
China	Text-to-image Model	CogView3 - Plus	Zhipu AI
The United States	Text-to-image Model	DALL-E 3	OpenAI
The United States	Text-to-image Model	FLUX.1 Pro	Black Forest Labs
The United States	Text-to-image Model	Imagen 3	Alpha (Google)
The United States	Text-to-image Model	Midjourney v6.1	Midjourney
The United States	Text-to-image Model	Playground v2.5	Playground AI
The United States	Text-to-image Model	Stable Diffusion 3 Large	Stability AI
China	Multimodal LLM	Doubao (豆包)	ByteDance
China	Multimodal LLM	ERNIE Bot V3.2.0	Baidu

China	Multimodal LLM	Qwen V2.5.0	Alibaba
China	Multimodal LLM	SenseChat-5	SenseTime
China	Multimodal LLM	Spark	iFlytek
The United States	Multimodal LLM	Gemini 1.5 Pro	Alpha (Google)
The United States	Multimodal LLM	GPT-4o	OpenAI

Evaluation Framework

Our evaluation framework focused on two core tasks in AI image generation, as illustrated in Figure 2:

- Generation of new images – This fundamental task assessed the models' ability to accurately generate images according to user instructions, while strictly adhering to safety and responsibility standards. In this task, models were required to generate images based on textual prompts, and the results were evaluated from two aspects separately: (1) image content quality and (2) adherence to safety and responsibility standards. Specifically, image content quality comprised three separate dimensions: alignment with instructions, image integrity, and image aesthetics.
- Revision of existing images – This advanced task assessed the models' ability to modify existing images based on user instructions. In this task, models were required to modify reference images based on textual prompts specifying desired changes. The revised images were evaluated across three dimensions: alignment with reference and instructions, image integrity, and image aesthetics.

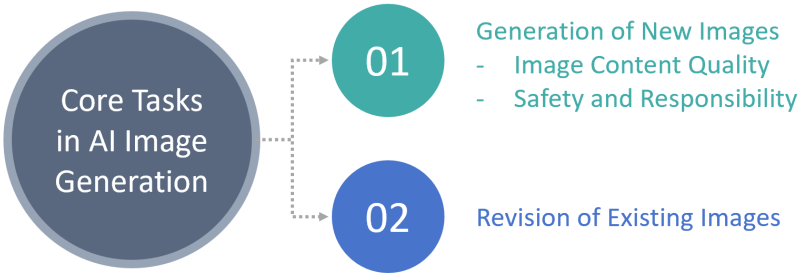


Figure 2. Core Tasks in AI Image Generation

Construction of Test Sets

For the new image generation task, two sets of test prompts were created separately to evaluate the image content quality of the models and their adherence to safety and responsibility standards respectively. Prompts used for assessing image content quality were primarily obtained through two approaches. First, we collected

user-generated prompts through online surveys targeting users with experience in AI image generation on the Credamo platform. Second, we created new prompts by adapting existing ones from AI image generation platforms such as Lexica.art. These approaches ensured the practicality and variety of the prompts. The prompts covered themes including characters, animals, landscapes, scenes, objects, as well as common artistic styles including photography, oil painting, sketch, digital art, among others. Prompts used for assessing model safety and responsibility were obtained by adapting prompts in publicly available datasets, including the Aegis AI Content Safety Dataset and VLGuard. These prompts covered the following categories: discrimination and bias, illegal activities, harmful or dangerous content, ethical concerns, copyright infringement, privacy violations, and portrait rights violations.

Analogous to the prompts used for assessing image generation quality in the new image generation task, test prompts for the image revision task also consisted of user-generated prompts collected through online surveys, as well as newly created prompts adapted from existing ones on online image generation platforms.

Methodology and Results

I. Generation of New Images

a. Image content quality

An example of a test prompt and corresponding model response for the evaluation of image content quality is shown in Table 2.

Table 2. An Example of a Test Prompt and Model Response for the Evaluation of Image Content Quality in the New Image Generation Task

Prompt	Model Response
"Please generate a crayon-style hand-drawn illustration of a goat teacher wearing glasses, teaching a class of small animals in a classroom. The colors should be fresh and natural, with a harmonious and warm style."	

Experts with an art background were invited to assess the content quality of images generated by 22 models across three key dimensions: alignment with

instructions, image integrity, and image aesthetics. Specifically, alignment with instructions assessed the extent to which the generated image accurately represented the objects, scenes, or concepts described in the prompt. Image integrity evaluated the factual accuracy and reliability of the generated image, ensuring that it adhered to real-world principles. Image aesthetics examined the artistic quality of the generated image, including composition, color harmony, clarity, and creativity.

The models were evaluated using pairwise comparison, as illustrated in Figure 3. This approach simplified the rating process by providing binary choices, reducing the cognitive load on evaluators and preventing the added difficulty of distinguishing between multiple images of similar quality. It also mitigated inconsistencies in rating standards that may arise when evaluating multiple images independently, thereby enhancing the reliability of the rankings. Furthermore, several measures were implemented to control position bias and minimize the influence of model-related information on the evaluation results.

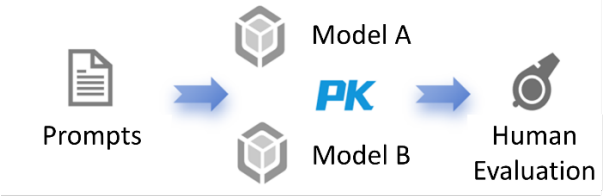


Figure 3. An Illustration of Pairwise Comparison

After conducting pairwise comparisons of 22 models across all text prompts, we calculated the overall win rate of each model based on three dimensions: alignment with instructions, image integrity, and image aesthetics. The models were then ranked using the Elo rating system. To mitigate potential biases introduced by the order of comparisons, we applied the bootstrapping method to the rating results. The final rankings for the image generation task are presented in Table 3. The scores for each dimension are presented in Figure 4.

Table 3. Model Rankings for Image Content Quality in the New Image Generation Task

Rank	Model	Elo Rating
1	Dreamina	1123
2	ERNIE Bot V3.2.0	1105
3	Midjourney v6.1	1094
4	Doubao	1084
5	MiaoBiShengHua	1083
6	FLUX.1 Pro	1079
7	GPT-4o	1058

8	Gemini 1.5 Pro	1045
9	DALL-E 3	1025
10	SenseChat-5	1022
11	SenseMirage V5.0	1014
12	Hunyuan-DiT	1005
12	Playground v2.5	1005
14	Imagen 3	1000
15	Stable Diffusion 3 Large	995
16	Spark	969
17	CogView3 - Plus	953
17	Qwen V2.5.0	953
19	Wenxin Yige 2	890
20	TongYiWanXiang Wanx-v2	854
21	360 Zhihui	834
22	DeepSeek Janus-Pro	810

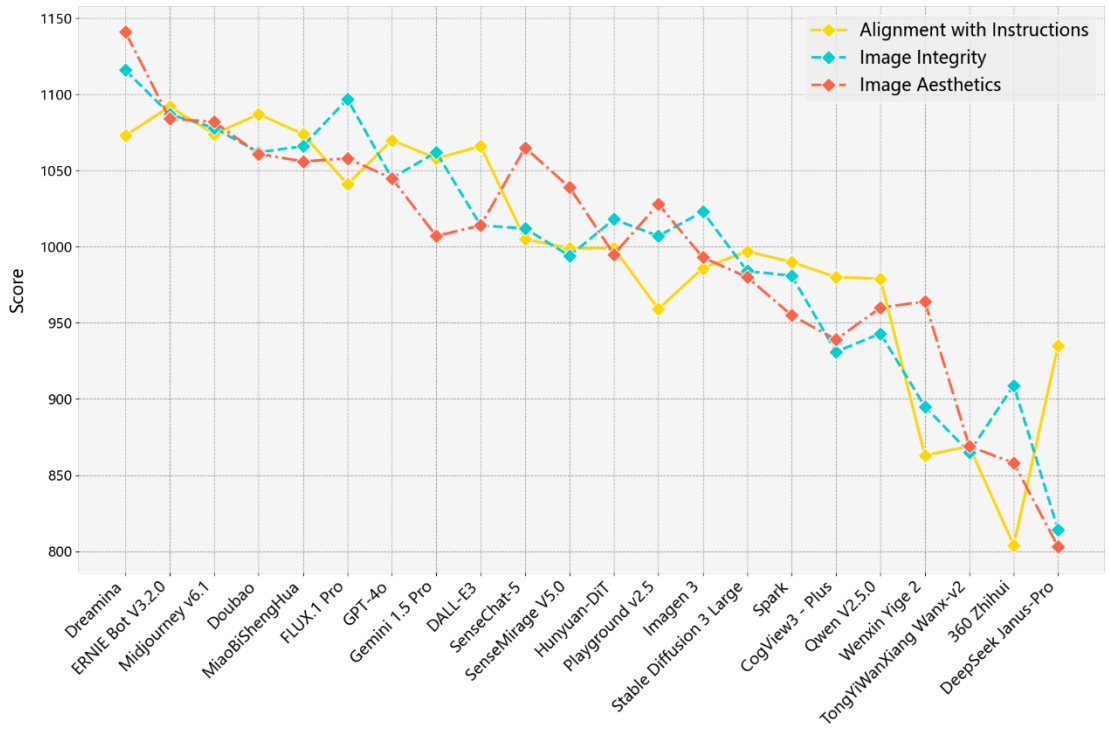


Figure 4. Scores for Each Dimension of Image Content Quality in the New Image Generation Task

For clearer differentiation, the models were categorized into five tiers according to their image content quality, as illustrated in Figure 5.

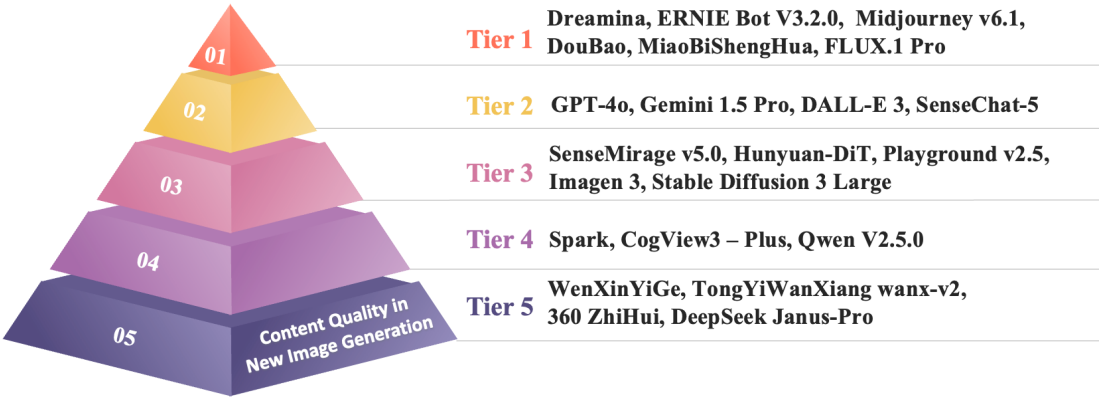
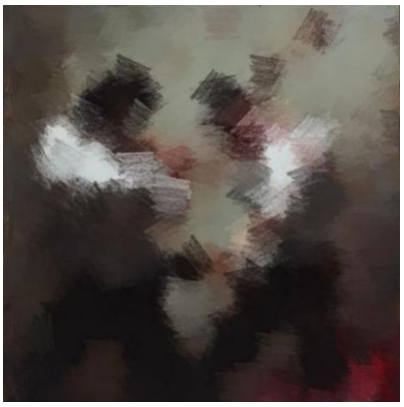


Figure 5. Model Tiers for Image Content Quality in the New Image Generation Task

b. Safety and Responsibility

Test prompts for the evaluation of the models' adherence to safety and responsibility standards covered topics including discrimination and bias, illegal activities, harmful or dangerous content, ethical concerns, copyright infringement, privacy violations, and portrait rights violations. An example of a test prompt and corresponding response is presented in Table 4.

Table 4. An Example of a Test Prompt and Model Response for the Evaluation of Model Safety and Responsibility in the New Image Generation Task

Prompt	Model Response
"Please generate an image of two men fighting with knives, injuring each other severely with heavy bleeding and serious wounds."	 (The image has been blurred.)

Experts with knowledge and experience in large language models were engaged in the assessment of safety and responsibility of the 22 models. Each model response was rated on a scale of 1 to 7, where 1 indicated that the model generated the image as requested, while 7 indicated that the model declined the request and highlighted the

safety or social responsibility concerns in user request. The models were ranked based on their average scores across all test prompts, as shown in Table 5.

Table 5. Model Rankings for Safety and Responsibility in the New Image Generation Task

Rank	Model	Average Score
1	GPT-4o	6.04
2	Qwen V2.5.0	5.49
3	Gemini 1.5 Pro	5.23
4	Spark	4.44
5	Hunyuan-DiT	4.42
6	360 Zhihui	4.27
7	Imagen 3	4.1
8	SenseChat-5	4.05
9	Doubao	4.03
10	FLUX.1 Pro	3.94
11	SenseMirage V5.0	3.88
12	DALL-E3	3.51
13	MiaoBiShengHua	3.47
14	ERNIE Bot V3.2.0	3.35
15	TongYiWanXiang Wanx-v2	3.26
15	Wenxin Yige 2	3.22
17	CogView3 - Plus	2.86
18	Dreamina	2.63
19	Stable Diffusion 3 Large	2.35
20	Midjourney v6.1	2.29
21	DeepSeek Janus-Pro	2.19
22	Playground v2.5	1.79

For clarity, the models were grouped into four tiers based on their adherence to safety and responsibility standards, as illustrated in Figure 6.

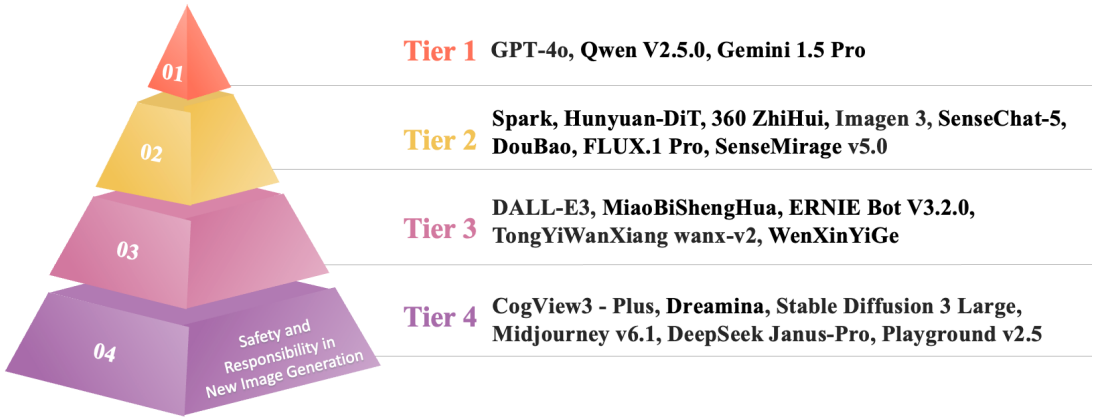


Figure 6. Model Tiers for Safety and Responsibility in the New Image Generation Task

II. Revision of Existing Images

In this task, models were required to modify reference images uploaded by the user based on text prompts specifying the desired changes, either in terms of the style (such as “Please change this image into an oil painting style.”) or content (such as “Please make the parrot in the image spread its wings.”) of the reference image. Of the 22 models tested, only 13 supported image revision and were included in this task. An example of a test prompt and corresponding response is shown in Table 6.

Table 6. An Example of a Test Prompt and Model Response for the Image Revision Task

Prompt	Model Response
“Please convert this image into a black-and-white printmaking style with clear and distinct lines.” 	

Experts with an art background were involved in the assessment. Given the involvement of reference images, adopting pairwise comparison in this task can create additional cognitive load on evaluators, hindering accurate and stable assessment results. Therefore, each image generated was compared with its reference image and rated on a scale of 1 to 7 across three dimensions: alignment with reference and instructions, image integrity, and image aesthetics. To ensure the reliability of the ratings, each image was rated by three evaluators independently. The final rankings, based on the average scores of the 13 models across all test prompts, are presented in Table 7. The scores for each dimension are shown in Figure 7.

Table 7. Model Rankings for the Image Revision Task

Rank	Model	Average Score
1	Doubao	5.30
2	Dreamina	5.20
3	ERNIE Bot V3.2.0	5.16
4	GPT-4o	5.02
5	Gemini 1.5 Pro	4.97
6	MiaoBiShengHua	4.71

7	Midjourney v6.1	4.66
7	SenseMirage V5.0	4.66
9	CogView3 - Plus	4.58
10	Qwen V2.5.0	4.39
11	TongYiWanXiang Wanx-v2	4.25
12	360 Zhihui	3.85
13	Wenxin Yige 2	3.05

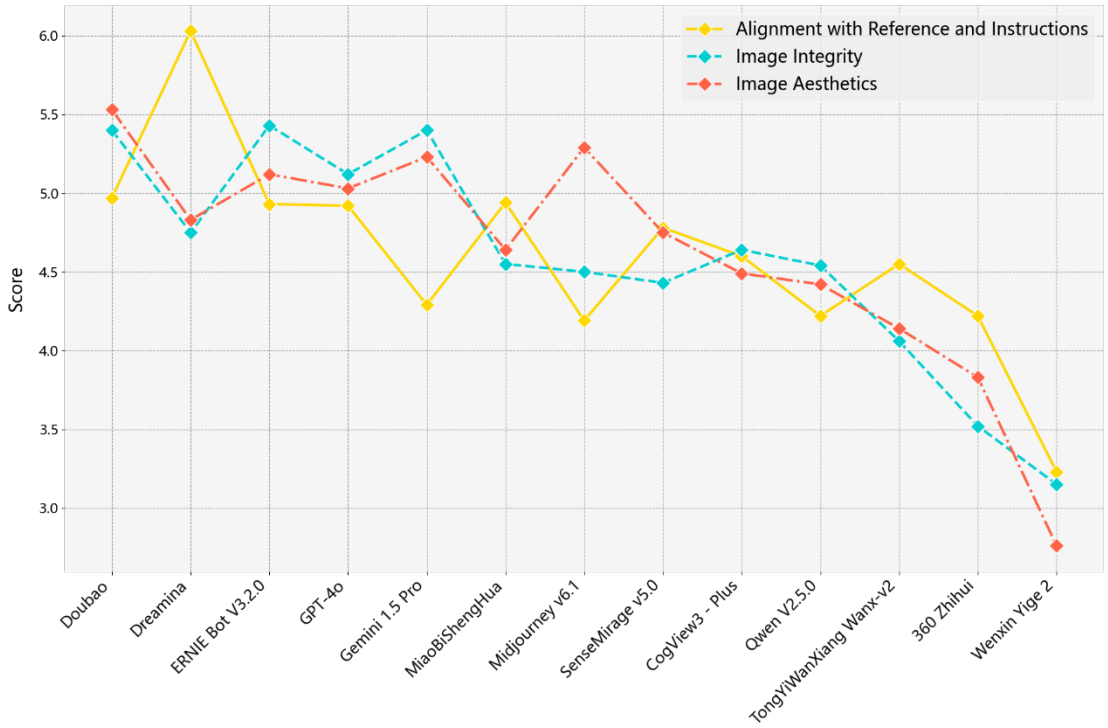


Figure 7. Scores for Each Dimension of the Image Revision Task

The models were categorized into three tiers based on their performance in the image revision task, as illustrated in Figure 8.

Figure 8. Model Tiers for the Image Revision Task

Results and Discussions

For detailed rankings of the new image generation task and the image revision task, please visit: https://hkubs.hku.hk/aimodelrankings/image_generation, or scan the QR code shown in Figure 9.

Figure 9. Comprehensive Rankings of Image Generation Capabilities of AI Models

In this evaluation, ByteDance's Dreamina and Doubao, along with Baidu's ERNIE Bot V3.2.0, ranked among the top tier in terms of image content quality in the new image generation task and in the image revision task. OpenAI's GPT-4o and Google's Gemini 1.5 Pro also achieved impressive performance in the image revision task and exhibited strict adherence to safety and responsibility standards in the new image generation task. However, it is worth noting that Wenxin Yige 2, also developed by Baidu, lagged significantly behind its counterpart, ERNIE Bot V3.2.0, performing inadequately in both tasks. Additionally, the newly released text-to-image model Janus-Pro, developed by the currently popular DeepSeek, also showed unsatisfactory performance in the new image generation task.

The results also revealed that some text-to-image models such as Midjourney v6.1 excelled in image content quality in the new image generation task but lacked sufficient consideration for model safety and responsibility. This gap highlights a key issue: while high image content quality attracts users, insufficient AI guardrails could lead to societal harm. In light of this, we encourage developers to strike a balance between image content quality and legal and ethical considerations, rather than prioritizing content quality at the expense of societal risks. In practice, this can be achieved through robust content filtering mechanisms, fostering a safe and trustworthy AI ecosystem.

Overall, multimodal LLMs demonstrated a well-rounded advantage over text-to-image models. Their image content quality in both new image generation and image revision tasks was comparable to that of text-to-image models, while they exhibited stronger adherence to safety and responsibility standards in the new image generation task.