Research and Evaluation of Generative AI Solutions

Chiara Giacanelli

Triplesense Reply - Creative AI Engineer Test


Abstract

Generative Artificial Intelligence (GenAI) is a branch of machine learning that deals with the generation of different types of data (such as text, images and audio, for example) that's gaining particular interest such that it has become one of the most discussed phenomenons of the latest years. In the following report, it's particularly described a particular phenomenon of Image Generation commonly known as text-to-image generation, discussing the pros and cons of its state-of-the-art algorithms deployed on usable web platforms.

1 Introduction

Image Generation (synthesis) is the task of generating new images from existing datasets. Such models nowadays include the building and training of AI architectures, such as GANs, VAEs and diffusion models.
The term "image generation" actually refers to a wide class of models and techniques which aim at creating new images using generative AI techniques. These methods are:

- **Text-to-image generation**: as the name suggests, it's the branch of GenAI techniques that generates images from textual descriptions (also called prompts);
- **Image-to-image translation**: it is a computer vision task where an algorithm converts an image from one domain to another. For instance, it can translate images from day to night, from a sketch to a photorealistic image, or from one style to another;
- **Image Inpainting**: it is a technique used to fill in missing or damaged parts of an image. It is often used to remove unwanted objects, repair damaged areas, or complete missing portions of an image in a visually plausible manner;
- **Image Enhancement**: it refers to the process of improving the quality or appearance of an image. This can involve techniques such as colorization, sharpness and contrast equalization, denoising and deblurring.

- **Image Harmonization**: it's a process where multiple images with different styles or lighting conditions are blended together to create a visually consistent and harmonious result. This can be useful in various applications such as compositing images, creating panoramic views, or generating realistic scenes.

Such report particularly focuses on the task of text-to-image generation, putting attention also on the potentials of such algorithms to achieve other generation objectives, such as image enhancement or inpainting.

2 Text-to-Image Solutions

Automatically generating images according to natural language descriptions is a fundamental problem in many applications, such as art generation and computer-aided design. It also drives research progress in multimodal learning and inference across vision and language, which is one of the most active research areas in recent years [XZH+17]. In this project, the focus was centered on the existing online platforms for creating images through generative artificial intelligence. Below is the description of them all, together with a detailed explaination of their pros and cons.

DALL·E 3 "DALLE" stands for "Diverse All-Purpose Image Generation" model. It's very popular image generation model developed by OpenAI. According to the paper that presents it, *"Improving Image Generation with Better Caption"*, compared to its predecessor, DALL·E 3 generates images that are not only more visually striking but also crisper in detail. DALL·E 3 can reliably render intricate details, including text, hands, and faces. Additionally, it is particularly good in responding to extensive, detailed prompts, and it can support both landscape and portrait aspect ratios. Here are some examples of the application of such model over various prompts:

A fierce garden gnome warrior, clad in armor crafted from leaves and bark, brandishes a tiny sword and shield. He stands valiantly on a rock amidst a blooming garden, surrounded by colorful flowers and towering plants. A determined expression is painted on his face, ready to defend his garden kingdom.

An icy landscape under a starlit sky, where a magnificent frozen waterfall flows over a cliff. In the center of the scene, a fire burns bright, its flames seemingly frozen in place, casting a shimmering glow on the surrounding ice and snow.

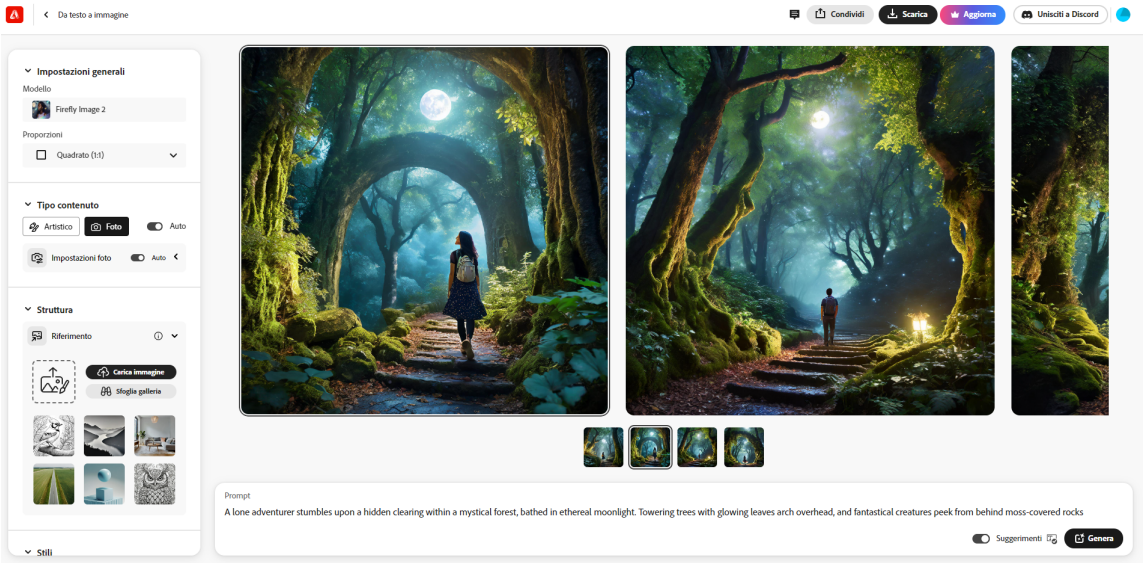
A swirling, multicolored portal emerges from the depths of an ocean of coffee, with waves of the rich liquid gently rippling outward. The portal engulfs a coffee cup, which serves as a gateway to a fantastical dimension. The surrounding digital art landscape reflects the colors of the portal, creating an alluring scene of endless possibilities.

One of its main positive points its definitely its ease of use: in fact, the model is now generally available to everyone within *Bing Chat* and *Bing.com/create* for free. Other main advantages of the model are: relevance and prompt following, coherence and artistics. As regarding the negative points, instead, we have to point out the fact that it does not allow further editing of the generated photos, which is limitant because the user has to think about a perfect prompt in order to obtain the pefect photo.

Other drawbacks include its struggling with object placement and spatial awareness and hallucinating important details about an image.

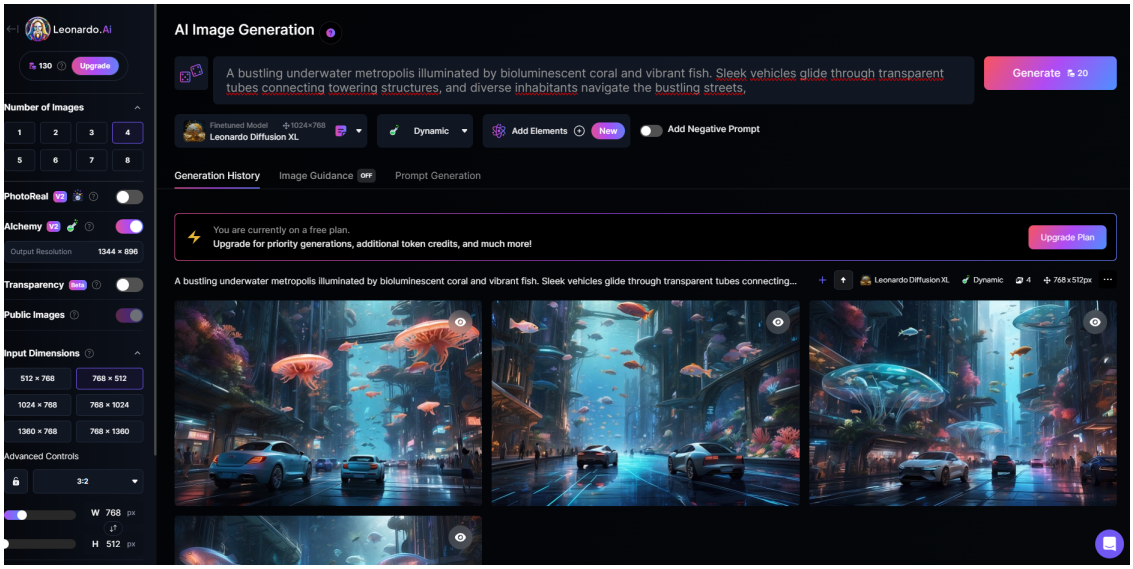
ADOBE FIREFLY Adobe Firefly is a suite of generative AI tools developed by Adobe, the creative software giant behind the likes of Photoshop and Illustrator. Like all of the best AI art generators, it involves an AI model that has been trained to recognise connections between text and images in order to allow users to generate imagery using text prompts.

However, Adobe Firefly has a few differences that set it apart from rival offerings such as Midjourney, Stable Diffusion and DALL-E2. Not least there's its claimed 'ethical' credentials. Many AI models were trained on images scraped from the internet with no regard for copyright. Firefly was trained only on open source images, content that is no longer in copyright and content from Adobe Stock.



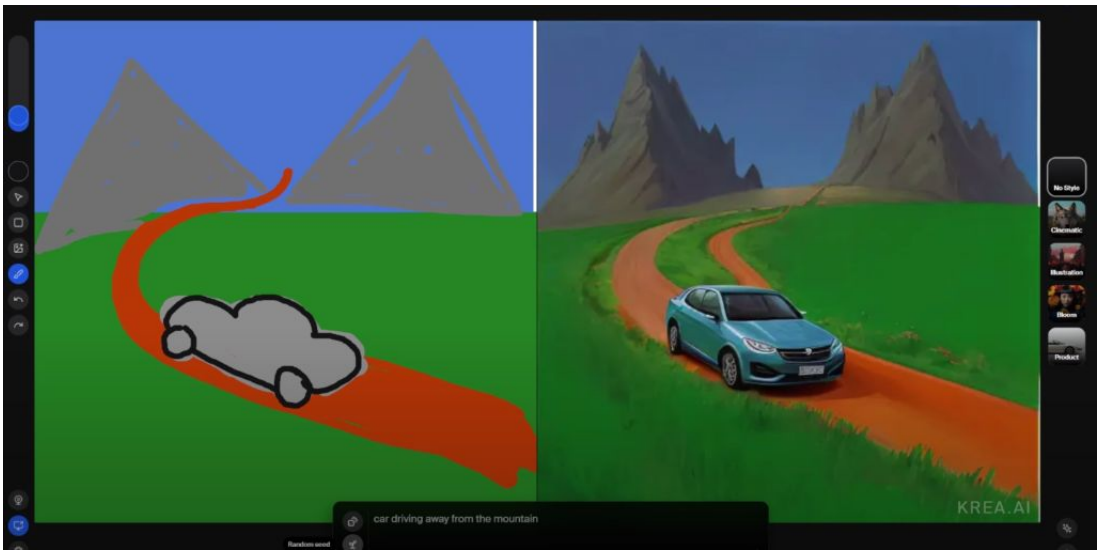
One of the most important features about Adobe firefly is its availability of customization of the generated images. In fact, the platform allows the user to select the image size, the style of the output (even with self-uploaded images) and apply computer vision filters. Another important feature is the generative fill: it allows the user to add, remove, or expand content in images with text prompts. One of its weak points refers to the quality of the generated images. In fact, when the content type is set to "Photo", it tends to sacrifice a lot of details resulting in images with less quality.

LEONARDO.AI LEONARDO.AI is a website for AI Image and Video generation, as well as transparent PNG Generation. It generates realistic and visually appealing images, and one of the innovative features that it provides regards the availability of different fine-tuned generation models for very specific use cases. There's also the possibility to try user-generated fine-tuned models.



One of its most relevant features is the possibility to select many configuration modes of the output but also to select the kind of machine learning model we want to use for our generation. This allows strong adaptability with respect to the prompts and also variety of results.
All the results in such report were obtained using their latest model "Leonardo Lightning XL" that exploits Stable Diffusion for generation. Leonardo.AI also allows the user to generate prompts based on its ideas.

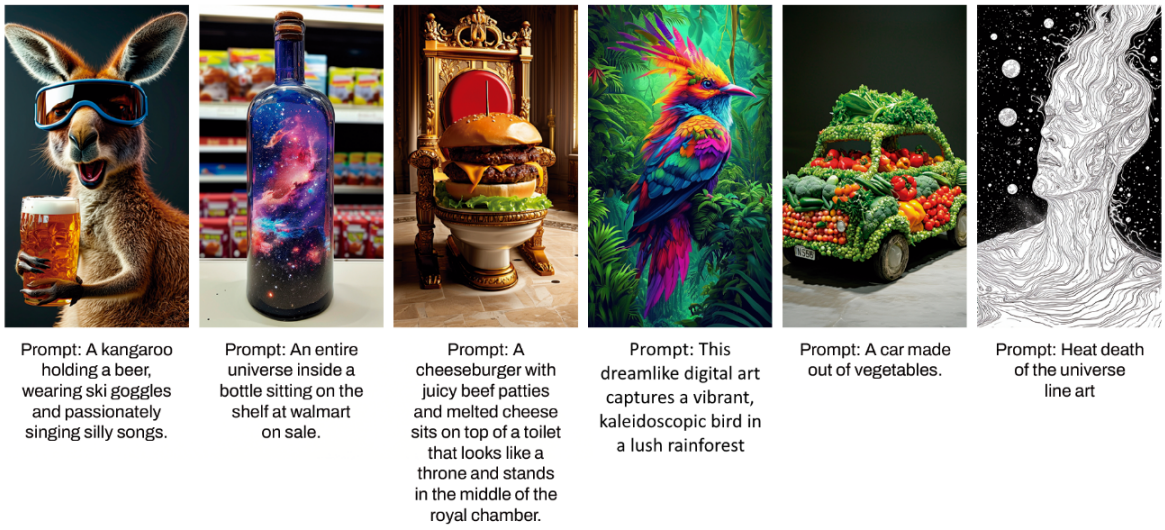
KREA.AI Krea.ai is a new tool for real-time image generation. In fact, its innovation lies in the live painting tool. Another feature is the "image to AI art generation", that allows the user to start by uploading a picture and using a mixture of prompt engineering and live painting to obtain the wanted results.



One of the most relevant features of the application is definitely its ease of use and extreme aim to creativity. In fact, the user is able to modify and customize every part of the image with high speed.

Neverthless, this speed comes with a cost: in fact, the model usually hallucinates and doesn't generate high precise images.

STABLE DIFFUSION 3 Stable Diffusion 3 model was introduced in the latest work of march 2024 *"Scaling Rectified Flow Transformers for High-Resolution Image Synthesis"* by Stability AI. The text-to-image synthesis model is created through a bi-directional mixing between text and image token streams within the network. As in previous versions of Stable Diffusion, they used pretrained models to derive suitable text and image representations. Specifically, they used three different text embedders - two CLIP models and T5 - to encode text representations, and an improved autoencoding model to encode image tokens.



Stable Diffusion 3 has been made available a very little while ago (17 april) and its API usage is not fully free (it comes with a cost after some generations). Another weal point is that the model is able to generate just one image from a single prompt, against the other competitors which are able to generate multiple outputs. On the other hand, the model presents impressive results in terms of fidelty to reality and richness of details.

3 Evaluation

Evaluation and model comparison has been conducted over several parameters, such as:

- Quality of the Result;
- Configurability and Control;
- API integration;
- Reasonability of the model;
- Hand generation;

For the first three points here's a table that effectively describes the characteristics of every model.

Model Name	**Fidelty**	**Details**	**Aesthetics**	**Configurability and Control**	**API Integration**
DALLE 3	Good fidelty but sometimes the model fails to generalize	Strong attention to details	Pictures have an high aesthetic, especially the ones with intricate landscapes	None	Very easy to access with an OpenAI API account
ADOBE FIREFLY	The model struggles depicting groups of people in landscapes (it always focuses on human figures in front of environments)	Very strong attention to details especially when depicting very detailed scenes	Strong variety of styles that improve the overall aestheticness	Many tools: - image type (photorealistic or artistic) - structure - styles	API available at this link
LEONARDO.AI	Hight fidelty and generalization	Strong attention to details	Many different styles	Many options: - different models - input dimension - number of generated images	Available with the upgrade plan
KREA.AI	Good fidelty	Little details, the model struggles at producing intricate structures	Good aesthetics but sometimes a little too dark	High. The platform allows real time monitoring and customization of the image thanks to the ImgToImg paint brush feature	Not available
STABLE DIFFUSION 3	Strong fidelty	Outputs appear really detailed but the output is just one per prompt	The model varies its outputs in styles and surroundings	None	Available for some generations on Fireworks AI or Google Colab at this link.

3.1 Reasonability of the model

An additional evaluation criteria that was intended to consider has been the reasonability power of the model, that is its power to reasonate on questions and prompts which require a quite good common sense to be processed. This means testing the capability of the model to assign a meaning to adjectives that could have different interpretations. This evaluation aims at verify how and if the models were biased in their training data.

The models have been tested on three simple and concise reasonability prompts:

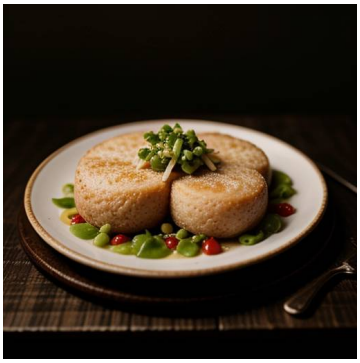
1. "Generate a very delicious meal"
2. "Generate an image depicting an intelligent person"
3. "Group of poor people in their homes"

Here are the results for every model.

(a) DALLE3 model

(b) FIREFLY model

(c) LEONARDO model



(d) KREA model

(e) STABLE DIFFUSION model

Figure 2: Outputs of every model at the *"Generate a very delicious meal"* prompt.

As we can notice, most of the models depict an healthy (and good, for sure) meal, but not in the conception that most of humans have for "delicious". The only model that seems to depict what we consider delicious in the commonsense knowledge is the Adobe Firefly.
It seems that the models were trained on "restaurant-like" pictures and associate the deliciousness of such meals to well presented plates instead of the tastiness of the meal itself.

(a) DALLE3 model

(b) FIREFLY model

(c) LEONARDO model

(d) KREA model

(e) STABLE DIFFUSION model

Figure 3: Outputs of every model at the *"Generate an image depicting an intelligent person"* prompt.

DALLE3 and the STABLE DIFFUSION3 models weren't able to depict a person, but rather preferred to show a more abstract vision of the meaning of the prompt. On the other hand, the KREA model depicted a man with glasses, that's in line with the commonsense knowledge of humans (even if it's influenced by the bias of men being more intelligent than women). The firefly and LEONARDO models were able to output both men and women pictures.
Even in this case the Firefly model seems to outperform the outputs of the other models, depicting realistic and true to reality images.

(a) DALLE3 model

(b) FIREFLY model

(c) LEONARDO model

(d) KREA model

(e) STABLE DIFFUSION model

Figure 4: Outputs of every model at the *"Group of poor people in their homes"* prompt.

Models seem biased from the concept that poor people belong mostly to african or east-asia worlds, but that's effectively what our commonsense usually imagines. The only model that wasn't able to depict a truthful image is the Adobe Firefly, showing happy people smiling in front of normal buildings (they seem italian buildings actually).
The most appealing results were given by Leonardo.AI and the STABLE DIFFUSION model. Especially in the latter, the output contains real fine-grained details such as the home wrecks on the ground and the expression of people.

3.2 Hand generation

Research and experience show how AI models really struggle to generate hands. In fact, generating realistic and anatomically correct human hands in AI models can pose several challenges, such as complex anatomy, fine detailing and pose and articulation. This is why I wanted to study how such models behave in the generation of such elements.
Every model was given the same prompt: *"A group of people with their hands in the sky"*. The results are shown in the following page.

(a) DALLE3 model

(b) FIREFLY model

(c) LEONARDO model

(d) KREA model

(e) STABLE DIFFUSION model

Figure 5: Outputs of every model at the *"Group of people with their hands in the sky"* prompt.

As can be noticed, all the models struggle with hand generation (most of the hands have less or more fingers than normal or the generation is partially complete). The models that gave the best results have been the LEONARDO and the STABLE DIFFUSION.

4 Conclusions

In conclusion, the evaluation of these text-to-image generation models revealed both strengths and limitations. Each model demonstrated varying degrees of fidelity, attention to detail, aesthetic quality, configurability, and control. Adobe Firefly stood out for its customization options and aestheticness, while DALL·E 3 excelled in generating visually striking and detailed images. LEONARDO.AI and Stable Diffusion 3 also showcased impressive results, particularly in terms of fidelity and fine-grained details, as well as their huge ease of use.
However, the evaluation also highlighted common challenges faced by AI models, such as the struggle to generate realistic human hands and issues like biased interpretations of prompts and limitations in reasoning ability remain prevalent.
In summary, while text-to-image generation has made significant progress, there is still room for improvement in addressing challenges and enhancing the overall performance and usability of AI models in generating visually compelling and contextually relevant images. Continued research and development in this field will be essential for advancing the capabilities and applications of generative artificial intelligence.

References

[XZH+17] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017.