Text-to-Image Synthesis Powered by Automatic Generated and Transferable Text-Prompt

Yuanbiao Wang **Ziye Tao** **Li Sun** **Yanqi Luo**

Abstract

Text-to-image synthesis is critical in generating customized and copyright-free images for conversational ChatBots and search engines. In order to generate high-quality images, creating informative and reusable text prompts are necessary. So text augmentation through GPT-2 model is first used to automatically enrich the texts and save user's trouble of typing explicit description. Three versions of stable diffusion models are then utilized to improve the model's adaptability of generating images to the specific style of MidJourney data. Similarity and aesthetic scores from CLIP embedding are calculated to compare the models. The research found out the fine-tuned stable diffusion model based on augmented prompts from GPT-2 achieves the best performance on aesthetic scores, which shows the endorsement of people on the generated images. Therefore, text augmentation through GPT-2 model could facilitate enriching the content of human prompts and bring convenience to users when generating images. Stable diffusion model after fine-tuning also has the effectiveness in capturing useful textual information and generate high-quality and customized images.

Our code can be found at: bluehttps://github.com/agil27/PromptGen

1 Introduction

In recent years, text-to-image generation techniques is trending, where large language models and generative models are tamed to work together in order to generate appropriate images only based on descriptive phrases. The nature of the text-to-image model indicates that in order to generate a suitable image, the input text should be as informative as it could be. Simple words or unclear statements might confuse the model and lead to low quality images. Thus, we purpose to develop a text-completion model to generate high quality descriptive text inputs for further image generalization. With the help of this model, the users won't need to write every statement explicitly. Instead, they just need to enter some deterministic keywords, and the model would complete and converge everything to a style preferred by the text-to-image model.

2 Related Work

2.1 Text-to-Image Synthesis

Text-to-image synthesis has always been a heated topic as it has great potential for application in a variety of fields such as art, creative industries, and education. Recently, diffusion models captured unprecedented public attention for the spectacular quality of generated images, especially when conditioned on multi-model inputs by introducing the powerful multi-model encoder CLIP [9]. Among them [8][10] [12], Stable Diffusion stood out with publicly available open-source implementations and has a widespread impact on the creative image generation community and business.

2.2 Automated Prompt Refinement

However, these text-to-image synthesis model has several defects: they still have some limitations in binding attributes to objects, producing coherent text, and producing details in complex scenes. Also, they are also very sensitive to the input text prompts. Adding or removing some depictions will result in notable distinctions in the generated images. And the automation of text prompt generation and refinement is necessary if we want to deliver the model as a user-accessible product.

While prompt engineering for text generation purposes has been more and more frequently dis-

cussed, less work has been done to rigorously examine how users can prompt generative frameworks with natural language for image generation purposes. One related work is [3], which uses large language and image generation models to produce conceptually blended images for an input object. It has two phases: reasoning and generation. In the reasoning phase, the model identifies a relevant object of the given input and generates a description of the blend of the two. In the generation phase, it explores BigSleep (BigGAN+CLIP) and DeepDaze (SIREN+CLIP) which use CLIP to guide the BigGAN[1] and SIREN[14] models for text-based image generation for visual conceptual blending.

Vivian Liu and Lydia B. Chilton conduct a series of experiments and present design guidelines to help people produce better outcomes from text-to-image generative models [5]. Some of the experiments are related to hyperparameters such as random initialization and some of the experiments are related to the content of the prompts such as the style and subjects. It provides some useful suggestions for an end user to provide text to the text-to-image model. For example, more attention should be put on subject and style instead of connecting words.

2.3 Cross-model Prompt Adaptation

The gap between different models and the reusability of refined text prompts is another concern for the commercialization of image synthesis model. Some recent research shows that it is practical to adapt such text-to-image synthesis model to specific domains with intricate design by leveraging the capacity of large text encoders. [11]. In this work, we propose to utilize such methods to perform cross-model prompt adaptation and further research into finishing this task at a lower cost. Lately, there is a trend to add trainable latent-space "prompts" (this is different from the text prompts used to generate the images) to perform few-shot transfer learning with large language models. [6][4]. In our work we use a similar idea and append a fixed-length trainable postfix to the encoded text representations to achieve cross-model adaption of text prompts.

2.4 Our Contributions

Our contributions in this work are three-fold:

1. We introduced a text-completion model that can add substantial content to any given input text, yielding a better image.

2. We propose two methods, namely finetuning and postfix tuning to adapt the Stable Diffusion model with the text-image pair data collected from another image-synthesis service.

3. We use the DrawBench aesthetic score as well as the CLIP cosine similarity score to measure the images generated with our proposed text augmenter and the tuned Stable Diffusion models, compared with the original version of Stable Diffusion, and confirm significant improvements.

3 Methodology

3.1 Data

The data we used is Midjourney-prompts, a moderate-scale text-image pair dataset. Unlike large text-image pair datasets like LAION, the texts in Midjourney-prompts are actual inputs to the image-synthesis service. Midjourney is an independent research lab. In 2022, it launched a text-to-image service on Discord where users can interact with a bot, provide some descriptions, and then get visual depictions from the bot that are faithful to their description. MidJourney-prompts dataset is collected and provided by Succinctly AI, which contains the text-image pairs that scrapped from a text-to-image service on public Discord server during the period June 20,2022 to July 17, 2022. There are 246,381 prompts in the dataset which are further divided into a train set (221743), a validation set (12318), and a test set(12320).

For the text-completion model, we only took the text data in this dataset to perform language model finetuning. For the transfer learning with the pretrained Stable Diffusion, due to the constraints of computational resources, we randomly sampled 2048 text-image pairs from the original dataset, meanwhile cropping and resizing them to 256×256 aspect ratio. We only selected the images that are not "upscaled" to rule out any potential repetition.

3.2 Automated Text-prompt Completion

The very primitive goal of our project is to build a language model that is able to auto-complete the text prompts given by user to invoke the AI engineer to generate images. Since large language models are proved effective in being transferred to

all kinds of downstream natural language processing and understanding tasks, we choose to finetune the GPT-2 model which is a large transformer-based language model with at most 1.5 billion parameters, trained on a dataset of 8 million web pages. The pretrained objective of the GPT-2 is to predict the next words based on the previous given text and the large scale pretrained work has enabled the GPT-2 model to perform text generation tasks well in many situations but it needs to be further finetuned to fit different downstream tasks. To explore if GPT-2 model can help generate high quality input text for the Midjourney text-image synthesis model based on user prompts, the GPT-2 model is finetuned on the MidJourney-prompts dataset(see section 3.1).

We used the GPT-2-medium model whose number of layers is 24, embedding size is 1024, number of heads is 16 and total 345M parameters. We finetune GPT-2 on the MidJourney-prompts dataset, using the same unsupervised pretraining algorithm and the corresponding pretrained GPT-2 tokenizer is used to tokenize the prompts. Since the dataset is a collection of high-quality text prompts to generate images, we expect the obtained finetuned model will have the capacity to generate decent prompts to a certain extent. For the baseline model, we plan to make no or only minor changes on the hyperparameters of the GPT-2 model and evaluate its performance by the perplexity metrics on this dataset as the baseline.

The finetuning algorithm follows the general training procedure of large language models by maximizing the log joint probability of the input sequence:

$$\log p(x) = \sum_{k=1}^{n} p(x_k | x_1, x_2, \cdots, x_{k-1}) \qquad (1)$$

where the conditional probabilities are calculated by large neural networks with fixed-length input (truncated the sequence elements out of the sliding window).

3.3 Cross-Model Adaptation

Based on the enriched text prompts, how to capture the useful textual information to generate the images is the next focus. Below the stable diffusion model is introduced to improve the model's ability of generating images to the specific type of MidJourney text prompts.

3.3.1 Finetuning the Stable Diffusion with DreamBooth

Similar to Dreambooth[11], we fixed the image encoder, VAE and the noise scheduler and only trained the text encoder as well as the noise-prediction UNet. The training target is to maximizing the cosine similarity between the predicted noise by UNet and the VAE-encoded image latents with noise added.

3.3.2 Postfix Tuning

We appended a fixed-length continuous trainable postfix to the text representation (output of the text encoder). If the text representation is of $T \times d$ shape, where T is the sequence length and d is the embedding dimension, the postfix would be of shape $T_0 \times d$. If the concatenation of text embedding and the postfix exceeds the maximum length allowed, the end of the text embedding will be truncated.

During the training, almost all the modules are frozen except for the postfix. The postfix will be continually optimized to find the best fit in hope of transitioning the original prompts from MidJourney to be more suitable for Stable Diffusion in the latent semantic space.

The continuous postfix is initialized with the encoded representation of a semantically meaningful sentence: `from midjourney style to stable diffusion style`.

3.3.3 Concept Injection

For both finetuning with DreamBooth and Postfix Tuning we added several new tokens. We call this concept injection because we inject some new concepts into the Stable Diffusion model through new words or control sequences. This includes the word "MidJourney" as well as several control commands in its API, such as `--uplight` to increase the brightness of the scene, and `--ar` to change the aspect ratio, `--chaos` to control how chaotic the image would be, and `--stylize` to create more artistic instead of photorealist output.

For DreamBooth finetuning, the text encoder will be optimized to better understand these new concepts. For the postfix tuning, we allow the embedding layer parameters related to these injected concepts to be trainable.

3.4 Evaluation and Metrics

The evaluation can be divided into two parts: quantitative evaluation and qualitative evaluation

which will both be based on DrawBench dataset. The DrawBench dataset is created by Google research team [13], which aims to evaluate text-to-image performance in different aspects such as cardinality, compositionality, spatial relations, color combinations and rare words. In total, DrawBench comprises 200 prompts across 11 categories, which is both comprehensive and small enough for human evaluation.

3.4.1 Quantitative

To quantitatively evaluate the final quality of generated images, two kinds of scores were introduced: similarity score and aesthetic score, which are derived from CLIP embedding. Contrastive Language-Image Pre-Training (CLIP) is a neural network model optimized to minimize the difference between image and text embedding[7]. Simple dot product score between the text and image embedding is commonly used to compare the performance of text-to-image process. To better evaluate the endorsement level by users of generated images, aesthetic score predictor, which is based on a simple neural net taking CLIP embedding as input, is also utilized to measure how much people like an image on average[2].

3.4.2 Qualitative

In addition to evaluate the correlation between the full input prompt and generated image directly, it is meaningful to see how the model react on the newly adding control sequences. Comparative experiments will be conducted to evaluate the model performance on each control sequence and some combinations of them. Using the prompts from DrawBench dataset and adding some control sequences manually enable us to visually check if the expected effect such as `--uplight` and `--chaos` show up in the image.

4 Experiments and Results

4.1 Text-prompt completion

Our baseline model is GPT-2-medium finetuned on the MidJourney-prompts dataset. We use the original GPT-2 Tokenizer with BOS, EOS and padding token. For the baseline, the model was finetuned for 3 epochs which takes about 6 to 7 hours on a NVIDIA V100 GPU with batch size being 6. GPT2-medium is a moderate-sized language model proposed by Open AI with 12 Transformer blocks and 355 million trainable parameters. Compared to the full-sized GPT-2, GPT-2-medium is more training-efficient considering the fact that we have limited computational resources. The max length of the generated text in this baseline model is 50 by default, which can be further tuned in the future.

The original data source comes with a train, validation and test partition. The average loss on the training data is 1.31 and the average perplexity is 3.72. The average loss on the test data is 1.592 and the average perplexity is 4.916, which shows satisfactory generalization ability on new data. To more intuitively demonstrate the model's ability on promoting image generation process, we apply some example texts to the baseline model, get the auto-completed text which contains more rich details. Then both texts are input into the MidJourney text-image synthesis system (where the data is collected) and same quality images are generated. We will qualitatively evaluate the richness and artistic expressiveness of both images.

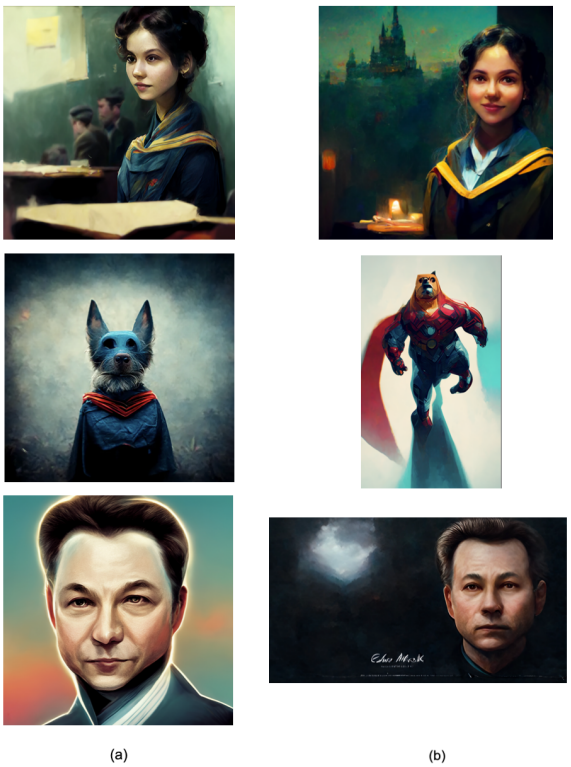

Figure 1: (a) Generated image based on original text. (b) Generated image based on auto-completed text.

Both texts are applied to the MidJourney text-image synthesis service (where the data is initially collected) to generate sythesized images. Several examples are shown in figure 1.

The first row of figure 1 compares the images

generated from the original text `a beautiful girl in the classroom` and the auto-completed text by the model `a beautiful girl in the classroom at Hogwarts teaching English, beautiful, happy, soft lighting. English world background. In the style of Ivan Aivazovsky, trending on artstation+Deviantart, style of artstationorea`. Compared with the original text, the auto-completed text add more concrete descriptions and defines the artistic style. As of the synthesized image, the one generated from the completed prompt is more vivid, natural and includes more convincing details, while the image generated from the origin text has many artifacts and looks eccentric.

The second row compares the images generated from the original text `a photo of a dog as a superhero` and the auto-completed text by the model `a photo of a dog as a superhero in the Marvel Cinematic movie Rogue. :: pixar. :: photo. Marvel style movie artwork. :: Marvel Comics style look. :: Marvel Comics colors. :: --ar 9:16 --uplight.` Compared with the original text, the auto-completed text enriches the prompt by specifying the style, lighting, format and even the aspect ratio. As a result, the generated image from the auto-completed prompt has more superhero-related elements (such as the Iron Man's suit) and is in more compliance with the theme.

The last row compares the images generated from the original text `Elon Musk in Twitter` and the auto-completed text by the model `Elon Musk in Twitter, album Cover, hyper-realistic, insanely detailed and intricate, dark, 8k, cinematic --ar 5:3 --uplight --iw 1.8 --stop 85 --uplight.` Similar with the previous example, the auto-completed text supplement the original prompt with detailed style, lighting, and the aspect ratio. As a result, the generated image from the auto-completed prompt is more realistic, and the background is more consistent with the serious expressions on his face.

4.2 Cross-model Adaptation

4.2.1 Quantitative Result

The raw and augmented DrawBench dataset are applied to the three versions of stable diffusion models, and similarity and aesthetic scores are calculated respectively. Summaries are shown in the tables below.

Similarity Scores			
	Pre-trained Stable Diffusion Model	Fine-tuned Model on MidJourney Data	Stable Diffusion Model with Postfix
Raw Data	0.303	0.309	0.292
Augmented Data	0.289	0.286	0.276

Table 1: Similarity scores across all the models

Aesthetic Scores			
	Pre-trained Stable Diffusion Model	Fine-tuned Model on MidJourney Data	Stable Diffusion Model with Postfix
Raw Data	5.128	5.137	5.112
Augmented Data	5.195	5.245	5.188

Table 2: Aesthetic scores across all the models

Table 1 shows that generated images based on augmented text have smaller similarity with the embedding of original prompts on average, which makes sense considering the fact that extra elements and details are added to the original prompts through text augmentation. However, images based on augmented text have higher aesthetic scores, which indicates they are more liked by people on average. Whether the generated image win users' affection or not is an important factor to evaluate the quality of text-to-image process. Therefore, the aesthetic score is attached more importance in our research when it comes to the final evaluation of models. Based on this criterion, the comparison shows that the fine-tuned stable diffusion model based on augmented prompts from GPT-2 model is the best.

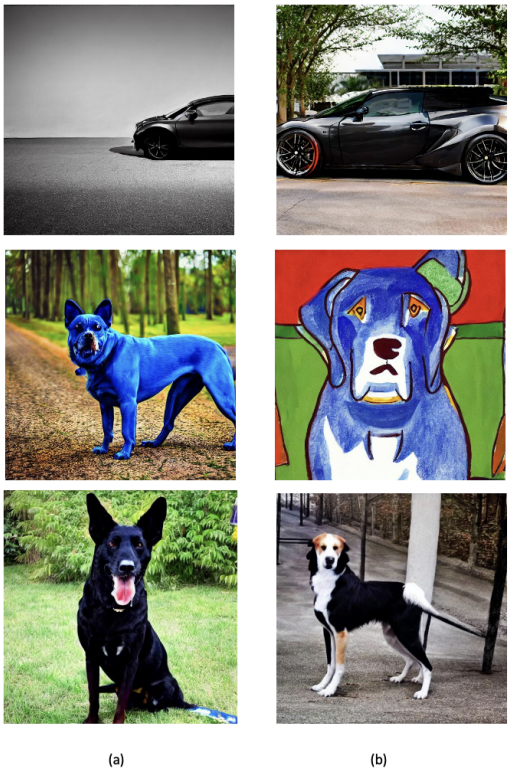

Figure 2: (a) Generated image with no control sequence. (b) Generated image with control sequence.

4.2.2 Qualitative Result

Figure 2 shows the comparison between images generated from the original prompts and images generated from prompts with control sequence. From the first row to the third row of figure 2, `--uplight`, `--stylize 5000`, and `--chaos 100` are added at the end of the original prompts ("`a black car`", "`a blue dog`", and "`a black dog`") respectively.

Control sequence `--uplight` will add less details during upscale process in order to be closer to the original image which usually applies to face or smooth surfaces. This could be shown in the first row of the figure 2 that the car surface in image (b) is smoother glossier compared to image (a) which is frosted more angular.

Control sequence `--stylize` followed by a parameter changes the degree of artistic the result would be. 5000, twice as much as the default value, makes the generated image more cartoonish compared with the realistic style image generated by the original prompt.

Control sequence `--chaos` followed by a parameter in range [0,100] controls the degree of variation of the result would be. The higher then parameter, the more unusual the generated image would be. Without setting any chaos degree, the model would generated a normal black dog. However, by setting the parameter to be 100, a dog with uncommon colors (with main color black) is generated.

5 Conclusion

The better performances on aesthetic score of generated images based on augmented text show the validity of text augmentation process through GPT-2 model. The text augmentation automatically enriches the original prompts by adding more image details, promotes the final aesthetic level and increases the images' popularity among users. It also brings convenience to users by saving their time and efforts from typing out descriptive prompts. Based on the augmented text, the stable diffusion model finetuned on MidJourney data achieves the best performance of generating customized images and shows the best adaptability to the special style of MidJourney prompts. Therefore, the finetuned stable diffusion model based on text augmenter of GPT-2 is proved to have the potency of augmenting the prompts itself and capturing meaningful textual information, which makes it possible to generate large quantity of high-quality and customized images efficiently.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2018.

[2] christophschuhmann. Clip+mlp aesthetic score predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor.

[3] Songwei Ge and Devi Parikh. Visual conceptual blending with large-scale language and vision models, 2021.

[4] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[5] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. *CHI Conference on Human Factors in Computing Systems*, 2022.

[6] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and et al. Learning transferable visual models from natural language supervision, Feb 2021.

[8] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[9] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[11] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[12] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[14] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473. Curran Associates, Inc., 2020.

A Impact Statement

A.1 Ethical

Without any human intervention(e.g. human filtered dataset and extra model results filter), the stable diffusion model and so many other models will produce sexual imagery if the user deliberately provide a related prompt, and will occasionally produce such images when this is not intended. Therefore, developers of stable diffusion propose the definition: "Not Safe for Work" (NSFW) for images and provide an safety checker on the model output to determine if it is safe to release the generated image. However the tool can be easily disabled, rendering it useless with one line code `--no-nsfw_checker`.

Our work shows feasibility of generating image on relative simple prompts from a given dataset by finetuning the stable diffusion model and the combination with GPT2 model even reduce users' work on the input end. This means that it may be also feasible to train the model on a dataset with unethical prompts and images easily. Without safety checker, the stable diffusion model which is open source has immediately brought the possibility of deepfakes much closer than they were. Images with nude celebrities can be quickly and freely in a batch.

Even when the checker is turned on, the checker is not perfect. Innocuous images can be occasionally flagged (false positives) and violent and gory imagery are frequently missed (false negatives). Bias is introduced through those mistakes which could lead to other ethical issues. Just like the famous crime detection algorithm, the checker might have a tendency to mistakenly flag or miss some certain groups of images. If that happens, it would be unfair to those minority groups and might hurt the right of those people. Thus, it is necessary to do more comprehensive research to verify the fairness of the checker.

A.2 Social

Text-to-Image synthesis has burst onto the scene and gained attention across the society. This advanced technology undoubtedly powered a wide variety of fields. For example, it allows generating customized and copyright-free images efficiently and easily for conversational ChatCots and search engines. This increases the variation of existing image database and brings benefits for researchers and normal users.

However, any new technology that brings profound revolutions will also unavoidably generate frictions and challenges with existing societal norms and policy frameworks. Text-to-Image generation is no exception. The booming of this technology may threaten the current job positions of many original art creators. It also intensifies the debate of what constitutes true art and whether Text-to-Image AI can ever create art at all. The validity and justifiability of utilizing this kind of technology to carry out art creation is still to be determined. For example, Jason M. Allen of Pueblo West, Colo. won the blue ribbon in the Colorado State Fair's annual art competition by creating a hyper-realistic graphic using Midjourney. Although he clearly stated the usage of technology from Midjourney, he was accused of cheating and a heated controversy over the legality of utilizing AI technique to replace the traditional art creation was sparked.

Based on the OpenAI's policy, DALL-E's individual users have full rights to commercialize the images that they create with the model, but OpenAI retains ultimate ownership over the original images. However, the court may not see it this way when high-stakes disputes involving these images get litigated. The regulation about this new field still needs improving in the future. Ultimately, these issues and uncertainty should not be seen as a stop hint of this technology, but rather as a temporarily unresolved barrier that would be eventually overcomed and become a pointer to a better future.