DALL·E 3论文公布、上线ChatGPT，作者一半是华人

作者：机器之心

原文链接：https://mp.weixin.qq.com/s/xLvJXe2FDL8YdByZLHjGMQ

论文地址：https://cdn.openai.com/papers/dall-e-3.pdf

Improving Image Generation with Better Captions

James Betker[*†] jbetker@openai.com
Gabriel Goh[*†] ggoh@openai.com
Li Jing[*†] lijing@openai.com
Tim Brooks[†]
Jianfeng Wang[‡] **Linjie Li**[‡] **Long Ouyang**[†] **Juntang Zhuang**[†] **Joyce Lee**[†] **Yufei Guo**[†]
Wesam Manassra[†] **Prafulla Dhariwal**[†] **Casey Chu**[†] **Yunxin Jiao**[†]
Aditya Ramesh[*†] aramesh@openai.com

Abstract

We show that prompt following abilities of text-to-image models can be substantially improved by training on highly descriptive generated image captions. Existing text-to-image models struggle to follow detailed image descriptions and often ignore words or confuse the meaning of prompts. We hypothesize that this issue stems from noisy and inaccurate image captions in the training dataset. We address this by training a bespoke image captioner and use it to recaption the training dataset. We then train several text-to-image models and find that training on these synthetic captions reliably improves prompt following ability. Finally, we use these findings to build DALL-E 3: a new text-to-image generation system, and benchmark its performance on an evaluation designed to measure prompt following, coherence, and aesthetics, finding that it compares favorably to competitors. We publish samples and code for these evaluations so that future research can continue optimizing this important aspect of text-to-image systems.

1 Introduction

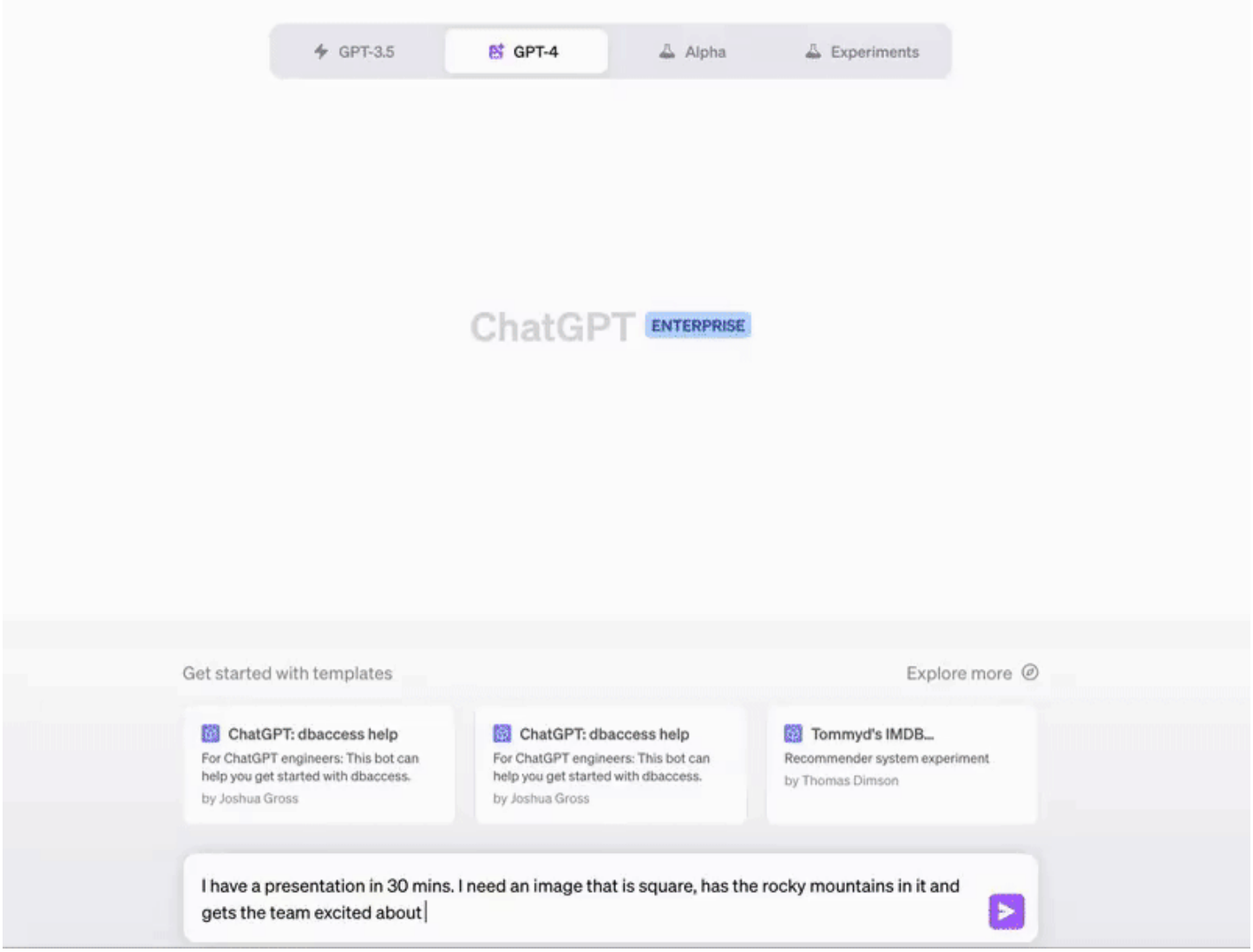
Recent advances in generative modeling have allowed text-to-image generative models to achieve drastic performance improvements. In particular, tackling the problem with sampling-based approaches such as autoregressive generative modeling[27, 2, 1, 20, 30] or using diffusion processes[25, 6, 11, 12, 19, 22] have allowed us to decompose the problem of image generation into small, discrete steps which are more tractable for neural networks to learn.

In parallel, researchers have found ways to build image generators out of stacks of self-attention layers[15, 3, 4]. Decoupling image generation from the implicit spatial biases of convolutions has allowed text-to-image models to reliably improve via the well-studied scaling properties of transformers.

Combined with a sufficiently large dataset, these approaches have enabled the training of large text-to-image models which are capable of generating imagery which is rapidly approaching the quality of photographs and artwork that humans can produce.

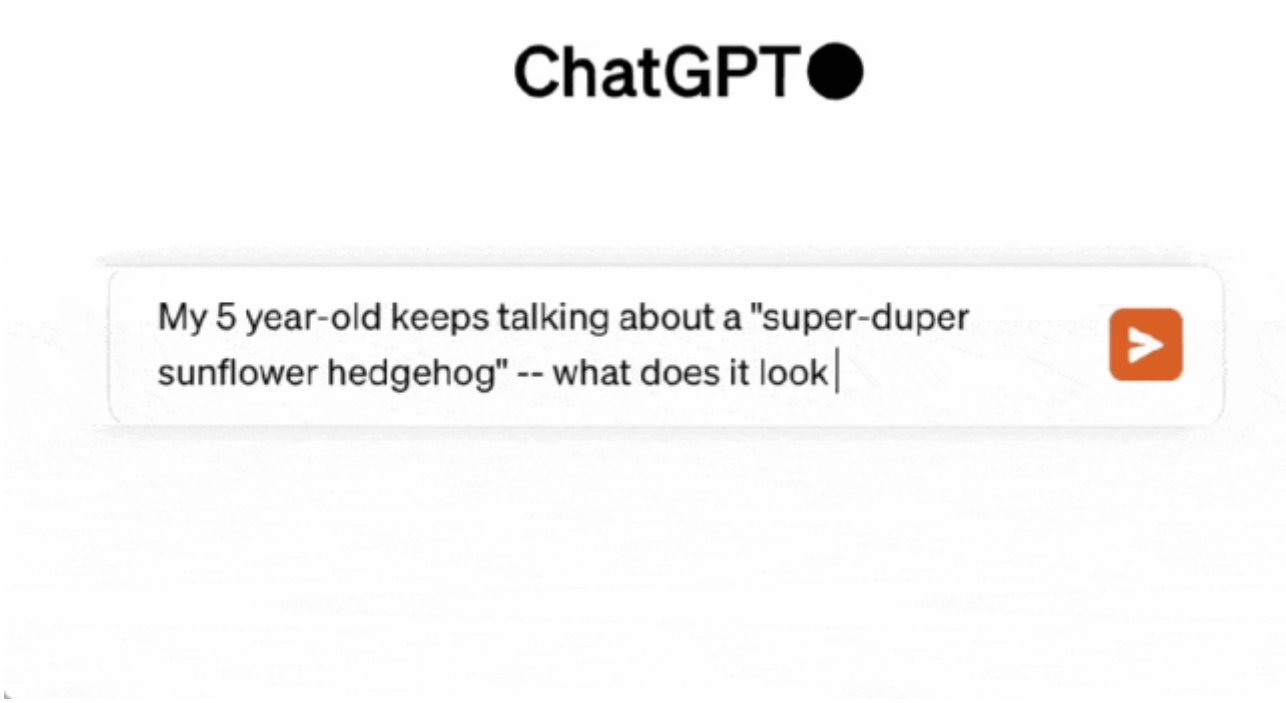
[*] Equal contribution
[†] OpenAI
[‡] Microsoft

打开 ChatGPT 就能用 DALL・E 3 生成图片了，OpenAI 还罕见地发布了一些技术细节。



终于，「OpenAI 又 Open 了」。在看到 OpenAI 刚刚发布的 DALL・E 3 相关论文后，一位网友感叹说。

DALL・E 3 是 OpenAI 在 2023 年 9 月份发布的一个文生图模型。与上一代模型 DALL・E 2 最大的区别在于，它可以利用 ChatGPT 生成提示（prompt），然后让模型根据该提示生成图像。对于不擅长编写提示的普通人来说，这一改进大大提高了 DALL・E 3 的使用效率。



此外，与 DALL・E 2 相比，DALL・E 3 生成的图质量也更高。

DALL・E 2 与 DALL・E 3 的生成效果对比。对于同样的 prompt「一幅描绘篮球运动员扣篮的油画，并伴以爆炸的星云」，左图 DALL・E 2 在细节、清晰度、明亮度等方面显然逊于右图 DALL・E 3。

即使与当前最流行的文生图应用 Midjourney 相比，DALL・E 3 也能打个平手甚至超越 Midjourney。而且与 Midjourney 相比，DALL・E 3 不需要用户自己掌握复杂的 Prompt 编写知识，使用起来门槛更低。

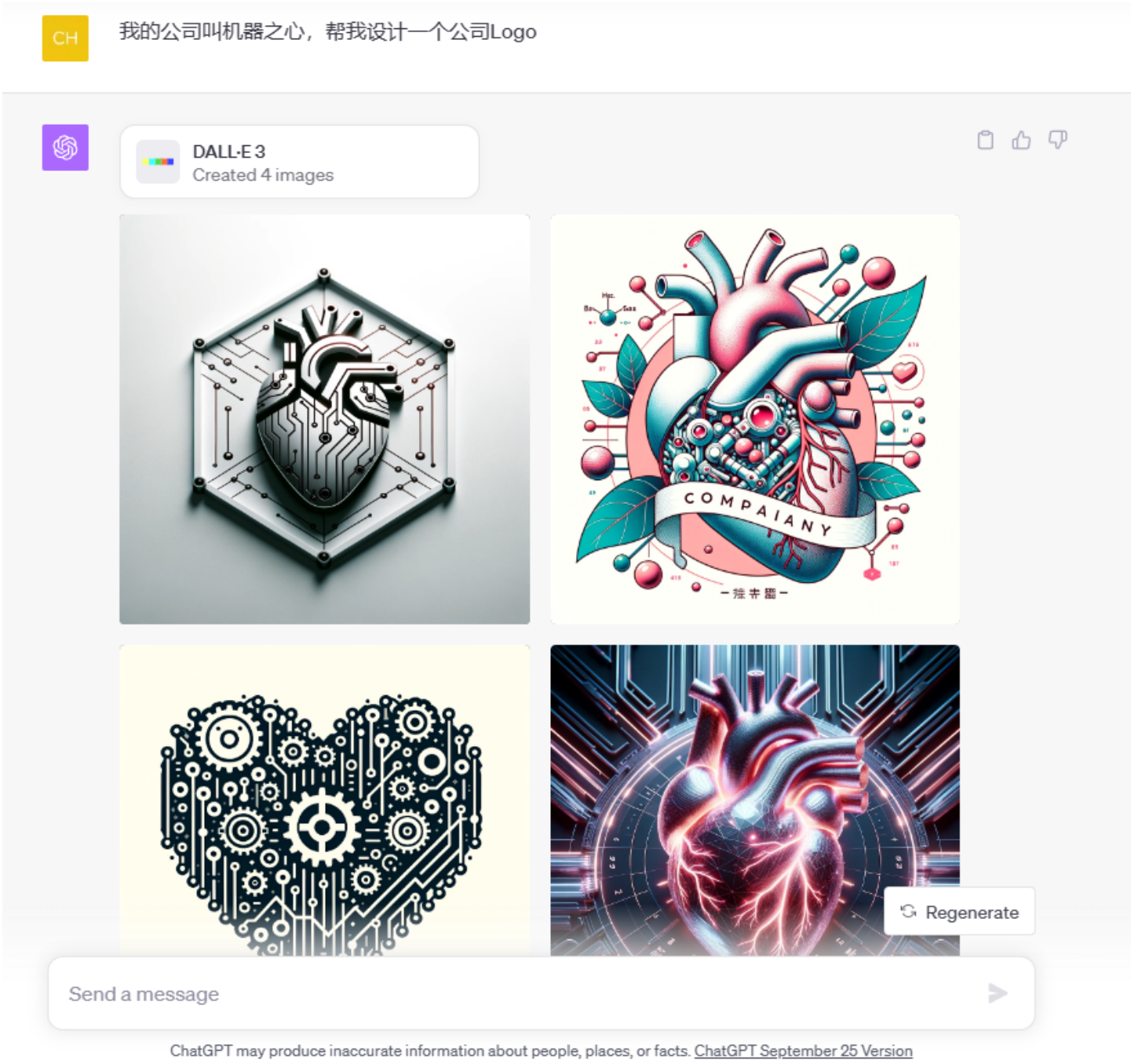
DALL・E 3 vs. Midjourney 文生图效果。prompt：这幅插画描绘了一颗由半透明玻璃制成的人心，矗立在惊涛骇浪中的基座上。一缕阳光穿透云层，照亮了心脏，揭示了其中的小宇宙。地平线上镌刻着一行醒目的大字「Find the universe within you」。

这一模型的发布引发了不小的轰动，也再次巩固了 OpenAI 技术领头羊的形象。一时间，所有人都很好奇，这么炸裂的效果是怎么做到的？不过，令人失望的是，当时 OpenAI 并没有透露技术细节，就像之前发布 GPT-4 时一样。

不过，一个月后，OpenAI 还是给了大家一些惊喜。在一份篇幅达 22 页的论文中，他们阐述了针对 DALL・E 3 所做的改进。论文要点包括：

- 模型能力的提升主要来自于详尽的图像文本描述（image captioning）；
- 他们训练了一个图像文本描述模型来生成简短而详尽的文本；
- 他们使用了 T5 文本编码器；
- 他们使用了 GPT-4 来完善用户写出的简短提示；
- 他们训练了一个 U-net 解码器，并将其蒸馏成 2 个去噪步骤；
- 文本渲染仍然不可靠，他们认为该模型很难将单词 token 映射为图像中的字母

除了论文之外，OpenAI 还公布了一个重要消息：DALL・E 3 已经正式上线 ChatGPT，Plus 用户和 Enterprise 用户都可以使用。以下是机器之心的试用效果：



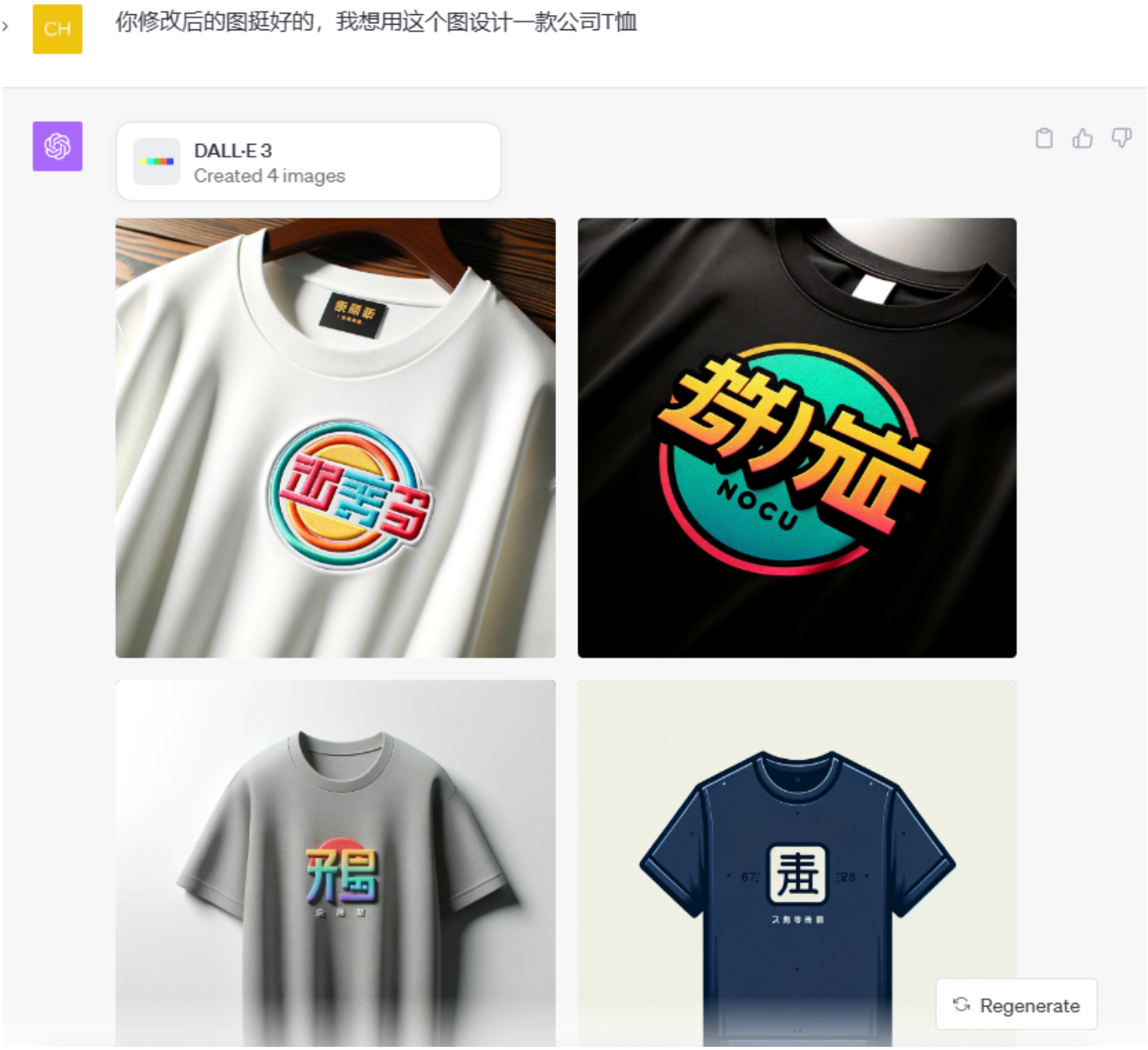
如果对结果不满意，你还可以直接让它在原图的基础上修改：

CH 我喜欢第一个图，但是我觉得它的颜色太单调了，帮我改一下

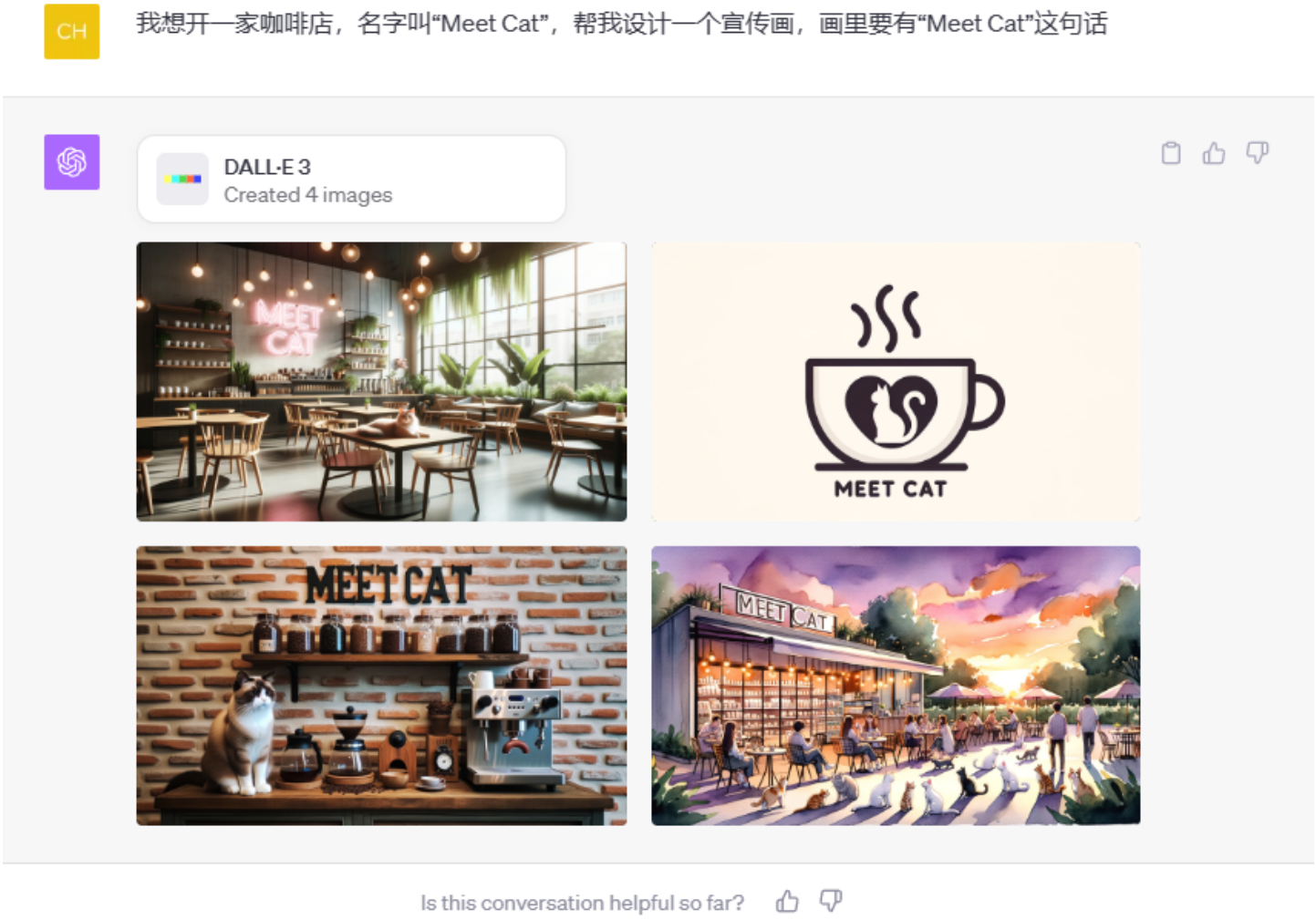
DALL·E 3
Created image

I've adjusted the color palette of the logo you liked, incorporating shades of blue, gold, and silver. Please take a look at the enhanced design and let me know if it meets your expectations or if you have any further adjustments in mind!

不过，随着对话长度的增加，生成结果变得有些不稳定：



在文字生成方面，DALL·E 3 已经提升了不少：



不过，在面对中文时，它的表现仍然较差：

为了保证 DALL・E 3 输出内容的安全性和合规性，OpenAI 也做了一些努力，确保模型输出的内容是被检查过的，而且不侵犯在世艺术家的版权。

当然，要了解 DALL・E 3 背后的技术，还是要详细阅读论文。以下是论文介绍：

论文概览

OpenAI 发布的 DALL・E 3 相关论文总共有 19 页，作者共有 15 位，半数为华人，分别来自 OpenAI 和微软。

Improving Image Generation with Better Captions

James Betker*† jbetker@openai.com
Gabriel Goh*† ggoh@openai.com
Li Jing*† lijing@openai.com
Tim Brooks†
Jianfeng Wang‡ Linjie Li‡ Long Ouyang† Juntang Zhuang† Joyce Lee† Yufei Guo†
Wesam Manassra† Prafulla Dhariwal† Casey Chu† Yunxin Jiao†
Aditya Ramesh*† aramesh@openai.com

论文地址：https://cdn.openai.com/papers/dall-e-3.pdf

论文提出了一种解决提示跟随（prompt following）问题的新方法：文本描述改进（caption improvement）。本文假设现有的文本 - 图像模型面临的一个基本问题是：训练数据集中的文本 - 图像对的质量较差，这一问题在其他研究中也已经被指出。本文建议通过为数据集中的图像生成改进的文本描述来解决这个问题。

为了达到这一目标，该研究首先学习了一个具有稳健性的图像文本生成器，它可以生成详细、准确的图像描述。然后，将此文本生成器应用到数据集以生成更详细的文本。最终在改进的数据集上训练文本 - 图像模型。

其实，用合成数据进行训练并不是一个全新的概念。本文的贡献主要在于研究者构建了一个新颖的具有描述性的图像文本系统，并对用合成文本训练生成的模型进行了评估。该研究还为一系列评估建立了一个可重复的基准性能概要文件，这些评估用于测量提示执行的情况。

在接下来的章节中，第 2 节对训练图像文本生成器的策略进行了全面概述，第 3 节对在原始文本和生成文本上训练的文本到图像模型进行了评估，第 4 节对 DALL-E 3 进行了评估，第 5 节讨论了限制和风险。

下面我们看看每个章节的具体内容。

数据集重描述（Recaptioning）

OpenAI 的文本到图像模型是在大量 (t, i) 对组成的数据集上进行训练的，其中 i 是图像，t 是描述图像的文本。在大规模数据集中，t 通常源于人类作者，他们主要对图像中的对象进行简单描述，而忽略图像中的背景细节或常识关系。

更糟糕的是，在互联网上找到的描述往往根本不正确或者描述与图像不怎么相关的细节。OpenAI 认为所有的缺陷都可以使用合成描述来解决。

- 构建图像描述生成器

图像描述生成器与可以预测文本的传统语言模型非常相似。因此，OpenAI 首先提供了语言模型的简单描述。这里先用分词器（tokenizer）将字符串分解为离散的 token，以这种方式分解之后，语料库的文本部分就表示为了序列 t = [t_1, t_2, . . . , t_n]。然后通过最大化以下似然函数来构建文本语言模型。

$$L(t) = \sum_{j} \log P(t_j | t_{j-k}, \ldots, t_{j-1}; \Theta) \tag{1}$$

接下来若想将该语言模型转换为描述生成器，只需要对图像进行调整即可。因此给定一个预训练的 CLIP 图像嵌入函数 F (i)，OpenAI 将语言模型目标做了如下增强。

- 微调描述生成器

为了改进在图像生成数据集上的描述效果，OpenAI 希望使用描述生成器来生成图像描述，这有助于学习文本到图像模型。

在首次尝试中，他们构建了一个仅能描述图像主对象的小规模描述数据集，然后继续在这个数据集上训练自己的描述生成器。该过程诱导的更新到 θ 使得模型偏向于描述图像的主对象。OpenAI 将这种微调生成的描述称为「短合成描述」。

OpenAI 做了第二次尝试，创建了一个更长的、描述更丰富的文本数据集，来描述微调数据集中每个图像的内容。這些描述包括图像的主对象，以及周围对象、背景、图像中的文本、风格、颜色。

他们在该数据集上对基础文本生成器进行进一步微调，并将该文本生成器生成的文本称为「描述性合成描述」。下图 3 展示了真值、短合成和描述性合成描述的示例。

评估重描述（re-captioned）数据集

OpenAI 利用重描述数据集，开始评估训练模型对合成文本的影响。他们尤其试图回答以下两个问题：

1. 使用每种类型的合成描述对性能有什么影响
2. 合成描述与真值描述的最佳混合比例是多少？

- 合成与真值描述混合

像文本到图像扩散模型这样的似然模型都有一个不好的倾向，即对数据集中的分布规律过拟合。当说到在合成描述上训练时，则需要考虑这个问题。

OpenAI 的描述生成器模型可能有很多难以检测的模态行为，但如果该模型基于描述进行训练，则这些行为将变成文本到图像模型的偏差。

解决这一问题的最佳方法是：将「输入」正则化为更接近人类可能使用的风格和格式的文本分布。使用真值描述时，你可以「自由」获得，这是由于它们实际上是从人类文本分布中提取的。此外，为了在使用合成描述时将正则化引入到自己的模型训练中，OpenAI 选择将合成描述与真值描述混合使用。

混合操作在数据采样时进行，这时 OpenAI 以固定的百分比随机选择真值或合成描述。

- 评估方法

在评估时，OpenAI 在相同的图像数据集上训练了相同的 T5-conditioned 图像扩散模型。所有的模型均以 2048 的 batch 大小训练了 500000 步，相当于 1B 张训练图像。

训练完成后，OpenAI 使用评估数据集上的描述来为每个模型生成 50000 张图像。接着使用 Hessel et al. (2022) 的 CLIP-S 评估指标对这些生成的图像进行评估。他们选择 CLIP 分数作为指标，该指标与文本图像相似度有很强的相关性。

OpenAI 首先使用公共 CLIP ViT-B/32 图像编码器来生成一个图像嵌入 z_i，然后使用文本编码器来为图像描述 z_t 创建一个文本嵌入，最后将 CLIP 分数计算为余弦距离 C。

$$C(z_i, z_t) = 1 - \frac{z_i \cdot z_t}{\|z_i\| \|z_t\|} \tag{3}$$

接下来针对为所有 50000 个文本 / 图像对计算的余弦距离，OpenAI 执行了平均操作，并做了 100 倍重缩放（rescale）。

在计算 CLIP 分数，选择使用哪个描述非常重要。对于 OpenAI 的测试，他们要么使用真值描述，要么使用描述性合成描述。同时，每次评估时都注明使用了哪个描述。

- 描述类型结果

OpenAI 首先分析了基于三类描述训练的模型之间的性能差异，为此训练了以下三个模型：

1. 仅在真值描述上训练的文本到图像模型
2. 在 95% 短合成描述上训练的文本到图像模型
3. 在 95% 描述性合成描述上训练的文本到图像模型

OpenAI 进行了两次评估，一次使用根据真值描述计算的 z_t，一次使用根据描述性合成描述计算的 z_t。这里没有选择短合成描述的原因是，它们与本次评估中的真值情况非常相似。

结果如下图 4 所示，其中在合成描述上训练的模型会得到比在真值描述上评估的基线模型好一些的 CLIP 分数性能，并且在描述性合成描述上评估时性能会明显更好。这表明在训练文本到图像模型时使用合成描述没有缺陷。

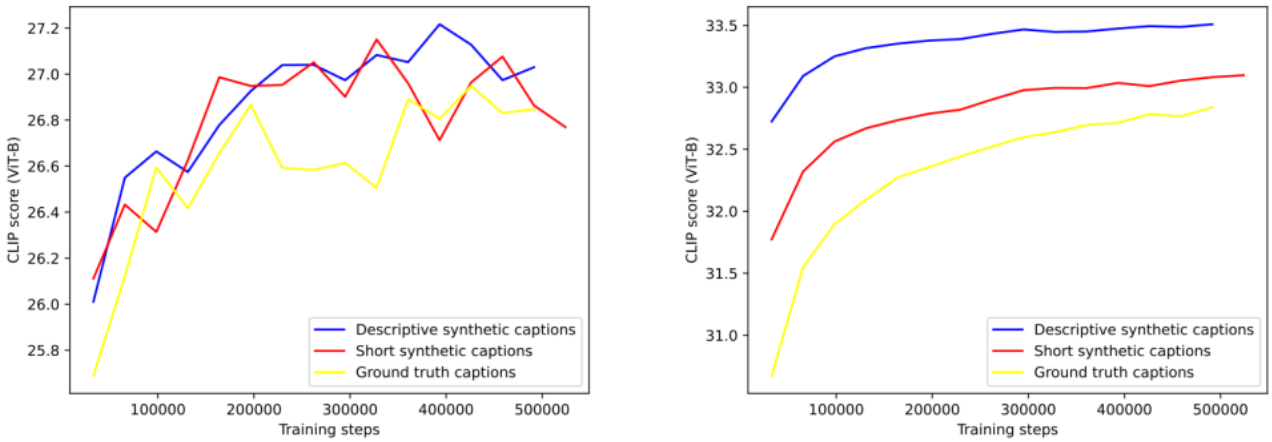


Figure 4 – CLIP scores for text-to-image models trained on different caption types. Left is evaluation results with ground truth captions on our evaluation dataset. Right uses the descriptive synthetic captions from the same dataset.

When computing CLIP scores, the choice of which caption to use when performing the above calculation is important. For our tests, we either use the ground truth caption or we use the descriptive synthetic caption. Which is used is noted in each evaluation.

- 描述混合比例

为了评估描述混合比例，OpenAI 使用不同混合比例的描述性合成描述，训练了四个图像生成模型。他们分别选择了 65%、80%、90% 和 95% 的合成描述混合比例。他们发现，实验进行到一半时，65% 的混合比例在所有评估中远远落后于其他比例，因此放弃不用。

下图 5 中的结果表明，合成描述混合比例越高，CLIP 分数往往越高，两者呈正比关系。

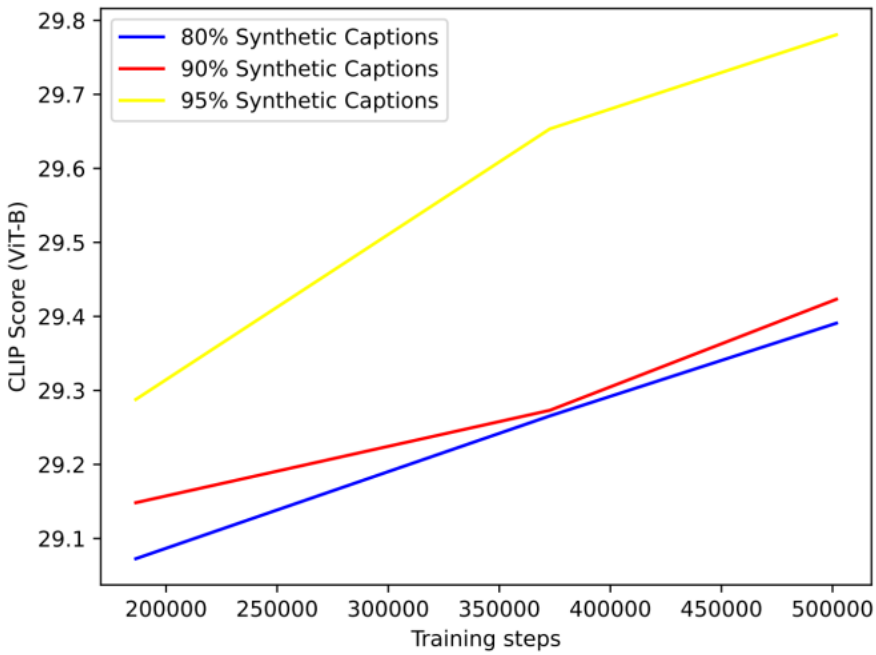


Figure 5 – CLIP scores for text-to-image models trained on various blending ratios of descriptive synthetic captions and ground-truth captions. Evaluation performed using ground truth captions.

DALL-E 3

为了大规模测试合成文本，本文对 DALL-E 3 进行了训练。训练过程中，本文混合使用了 95% 的合成文本和 5% 的真实文本。比较模型包括 DALL-E 2 以及 Stable Diffusion XL 1.0。

在 CLIP 得分评估中，DALL-E 3 优于 DALL-E 2 和 Stable Diffusion XL；在 Drawbench 基准评估中，DALL-E 3 同样优于 DALL-E 2 和 Stable Diffusion XL。

本文还将 DALL-E 3 生成的样例与其他模型生成的结果进行了对比。他们通过向人类评分员展示由相同描述生成的两张并排的图像进行评分，评分中包括三个方面：提示跟随（Prompt following）、风格（Style）、连贯性（Coherence）。

- 提示跟随：给评分 yuan 提供完整的图像描述内容，要求评分员选择更符合文本描述的图像；
- 风格：让评分员想象一下自己正在借助一些工具根据文本生成图像。如果你自己正在使用此工具，请选择你希望看到的图像；
- 连贯性：让评分员选择哪张图像包含更连贯的对象，例如从人的身体部位、面部和姿势、对象的位置等方面做出判断。

结果显示，DALL-E 3 在所有三个方面，尤其是在提示跟随方面，DALL-E 3 生成的图像在大多数情况下都比所有竞争对手更受人类评分者的青睐。

限制与风险

本文的最后一章是大家比较关心的关于限制与风险的问题。虽然 DALL-E 3 在 prompt 跟随方面表现出色，但它仍然在空间感知等方面表现不佳。例如，DALL-E 3 不能很好的理解左边、下面、后面等表示方位的词语。

此外，在构建文本描述生成器时，本文着重考虑了一些突出的引导词（prominent words），这些引导词存在于原本图像以及生成的描述中。因此，DALL-E 3 可以在出现 prompt 时生成文本。在测试过程中，本文注意到此功能并不可靠。本文怀疑这可能与使用 T5 文本编码器有关：当模型遇到 prompt 中的文本时，它实际上会看到代表整个单词的 token，并且将它们映射到图像中出现的文本。在未来的工作中，本文希望进一步探索字符级语言模型，以帮助改善 DALL-E 3 面临的这种限制。

最后，本文还观察到，合成的文本还会让生成的图片在重要细节上产生幻觉。这对下游任务产生了一定的影响，本文也表示，DALL-E 3 在为特定术语生成图像方面并不可靠。不过，该研究相信，对图像文本描述的完善能进一步改进 DALL-E 3 的生成结果。