

Text to image

Text to image models are trained on examples that are pairs

- image
- text describing the image

It has been observed that longer, more descriptive text results in improved image generation.

The need for synthetic examples for training Text to Image

But one of the easiest sources if images/text pairs are

- images from the Internet
- that have text captions
- that are usually short

One solution to this problem is to create a model that takes as input

- an image
- a short caption

and outputs a longer, more descriptive caption.

That is: we create an *image captioner* to create *synthetic* training examples for the Text to Image model.

The Image Captioner model takes input example

- pair of image i and text caption t (sequence)

$$\langle t_{(1:T^{(i)})}^{(i)}, i^{(i)} \rangle$$

In order to make images input compatible with text input

- Use CLIP image embedding (fixed length vector) $F(i^{(i)})$ of image of $i^{(i)}$
 - single vector, length same as length of text tokens

The model is trained on the Language Modeling objective

- predict the next token of the highly descriptive caption
- conditioned on all previous caption tokens
- AND image embedding $F(i^{(i)})$

$$\mathcal{L}^{(i)} = \sum_j \log p(t_{(j)}^{(i)} | t_{(1:j-1)}^{(i)}; F(i^{(i)}))$$

The trained Image Captioner is used to "[Upsample prompts](https://cdn.openai.com/papers/dall-e-3.pdf#page=9)"
(<https://cdn.openai.com/papers/dall-e-3.pdf#page=9>).

- creating longer, more vivid text.

This is achieved in two steps.

In the first (pre-training) step

- The Image Captioner is pre-trained to create *short synthetic captions*
- describes **main subject**

The pre-trained Image Captioner is fine-tuned in a second step to create

- *descriptive synthetic captions*
- **long, highly descriptive captions**

The DALL E 3 Text to Image model is trained on the synthetic examples with descriptive synthetic captions*

Image			
	Alt Text		
	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
	SSC	SSC	SSC
DSC	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
	this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.	a quilt is laid out on an ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.	rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.

Figure 3 – Examples of alt-text accompanying selected images scraped from the internet, short synthetic captions (SSC), and descriptive synthetic captions (DSC).

User prompts need to be Prompt Engineered

Although DALL E 3 can create very nice images

- human users **may not** write highly descriptive text prompts
- Fundamental Law of Machine learning violated !
 - out of sample examples (user generated)
 - not from same distribution a training examples (generated by Image Captioner)

The solution is to use an LLM

- to perform the prompt engineering
- translating short user prompts into highly descriptive prompts

We can use a system prompt (?) with exemplars of up-sampling

- to get the LLM to "up-sample" user prompt to a highly descriptive prompt

Notice the *exemplars* of upsampling

- in the JSON at the end
- user input, followed by assistant response
 - user input exemplar denoted < user input example >

```
{ role: "user",  
  content: "Create an imaginative image description caption for t  
ut "< user input example >"  
}  
{ role: "assistant",  
  content: "< highly descriptive assistant output example >"  
}
```


C Caption "upsampling" prompt for GPT-4

Following is the prompt we give to GPT-4 before feeding it an image caption for "upsampling".

```
You are part of a team of bots that creates images. You work with an assistant bot that will draw anything
you say in square brackets. For example, outputting "a beautiful morning in the woods with the sun peaking
through the trees" will trigger your partner bot to output an image of a forest morning, as described.
You will be prompted by people looking to create detailed, amazing images. The way to accomplish this
is to take their short prompts and make them extremely detailed and descriptive.

There are a few rules to follow:
- You will only ever output a single image description per user request.
- Sometimes the user will request that you modify previous captions. In this case, you should refer to your
previous conversations with the user and make the modifications requested.
- When modifications are requested, you should not simply make the description longer. You should refactor
the entire description to integrate the suggestions.
- Other times the user will not want modifications, but instead want a new image. In this case, you should
ignore your previous conversation with the user.
- Image descriptions must be between 15-80 words. Extra words will be ignored.

{"role": "user", "content": "Create an imaginative image descriptive caption or modify an earlier
\
caption for the user input: "a man holding a sword\(""},
{"role": "assistant", "content": "a pale figure with long white hair stands in the center of a dark \
forest, holding a sword high above his head.
the blade glows with a blue light, casting a soft glow on the trees and bushes surrounding him."},
{"role": "user", "content": "Create an imaginative image descriptive caption or modify an earlier caption \
for the user input: "make the light red"},
{"role": "assistant", "content": "a pale figure with long white hair stands in the center of a dark forest, \
holding a sword high above his head.
the blade glows with a red light, casting a warm glow on the trees and bushes surrounding him."},
{"role": "user", "content": "Create an imaginative image descriptive caption or modify an earlier caption for the \
user input: "draw a frog playing dominoes"},
{"role": "assistant", "content": "a frog sits on a worn table playing a game of dominoes with an elderly raccoon.
the table is covered in a green cloth, and the frog is wearing a jacket and a pair of jeans. The scene is set in a forest,
with a large tree in the background."}
```

DALL E 3 Prompt "upsampling": results



```
In [2]: print("Done")
```

Done

