



DALL·E



DALL·E: Creating Images from Text

We've trained a neural network called DALL·E that creates images from text captions for a wide range of concepts expressible in natural language.





OpenAI

TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES



[Edit prompt or view more images↓](#)

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES



[Edit prompt or view more images↓](#)

Table of contents

Overview

Capabilities

Sources

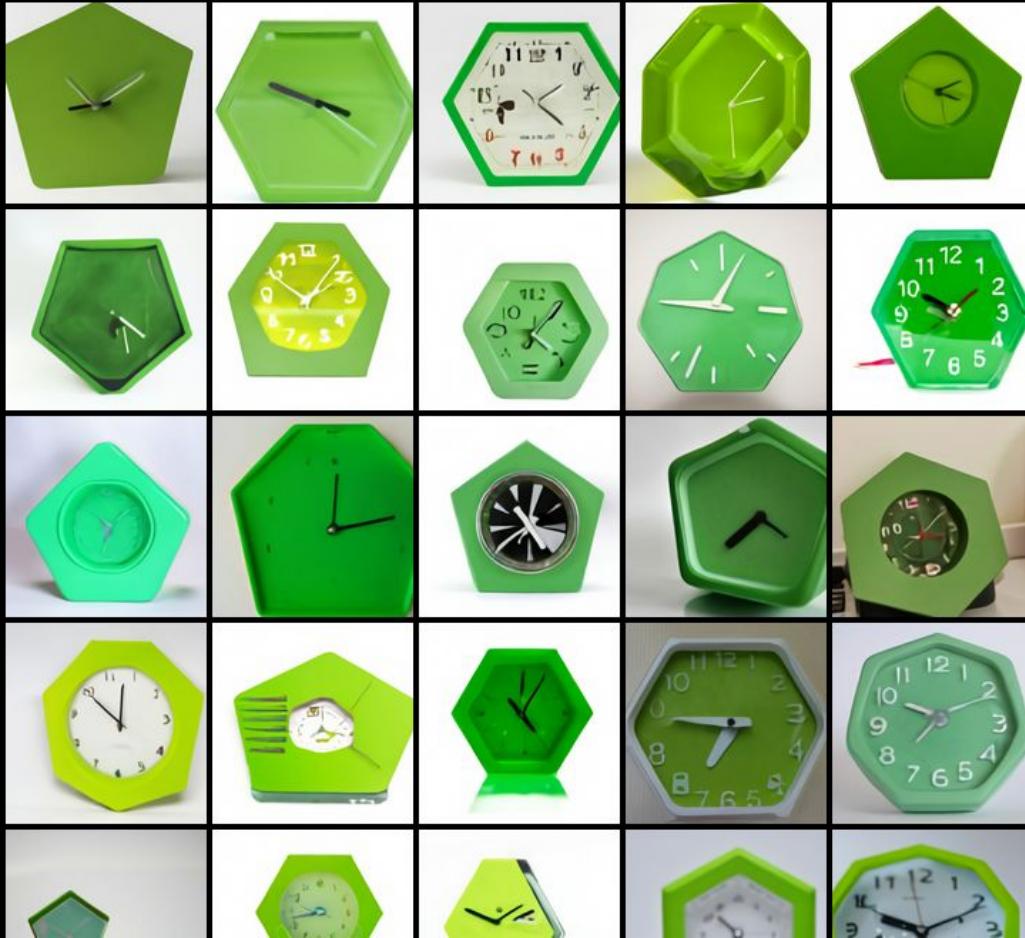
Capabilities

Controlling attributes

TEXT PROMPT

a pentagonal green clock. a green clock in the shape of a pentagon.

AI-GENERATED
IMAGES



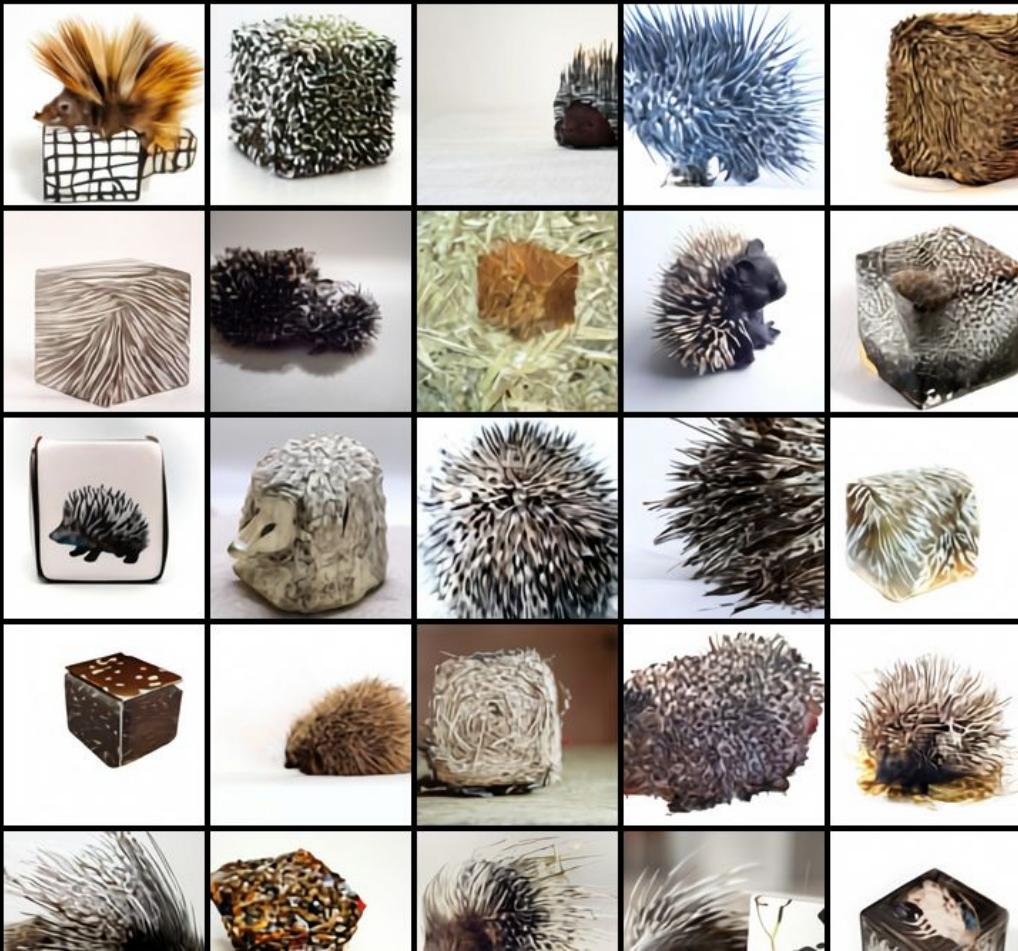
We find that DALL-E can
in polygonal shapes that
to occur in the real world
as "picture frame" and "
reliably draw the object
shapes except heptago
as "manhole cover" and
success rate for more u
"pentagon," is considera

For several of the visual
that repeating the capti
alternative phrasings, in
of the results.

TEXT PROMPT

a cube made of porcupine. a cube with the texture of a porcupine.

AI-GENERATED
IMAGES

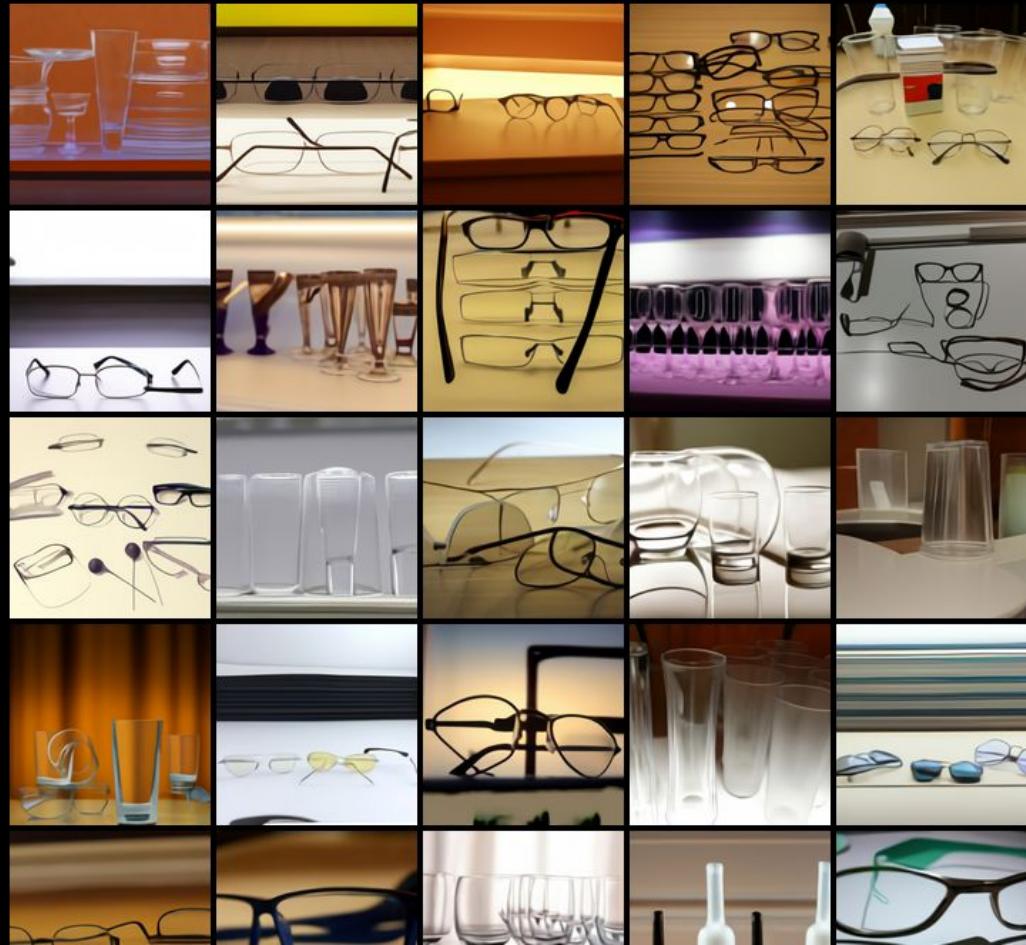


We find that DALL-E can generate various plants, animals, and three dimensional solids. In this visual, we find that repeating alternative phrasing improves the results.

TEXT PROMPT

a collection of glasses is sitting on a table

AI-GENERATED
IMAGES



We find that DALL-E is able to generate multiple copies of an object when it is unable to reliably count. When prompted to draw nouns with multiple meanings, such as "glasses" and "cups" it sometimes generates interpretations, depending on which meaning of the word is used.

Capabilities

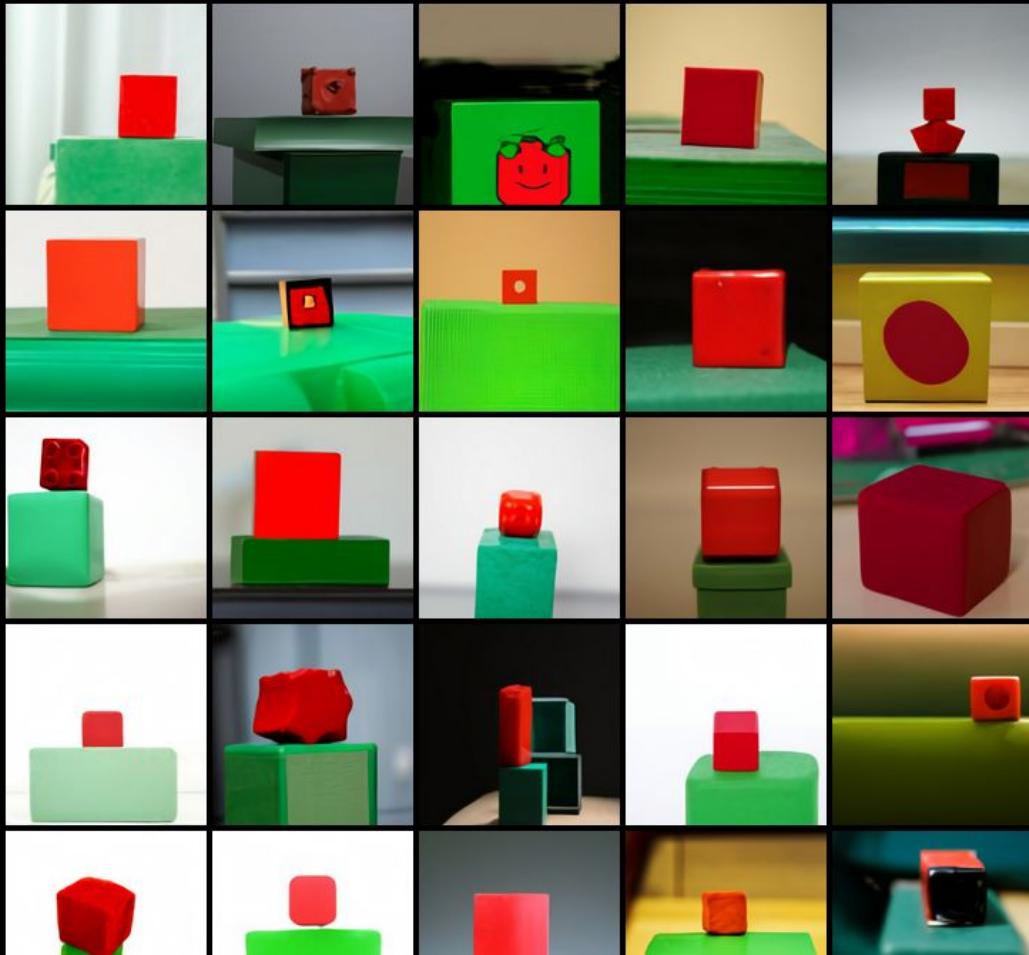
Controlling attributes

Drawing multiple objects

TEXT PROMPT

a small red block sitting on a large green block

AI-GENERATED
IMAGES

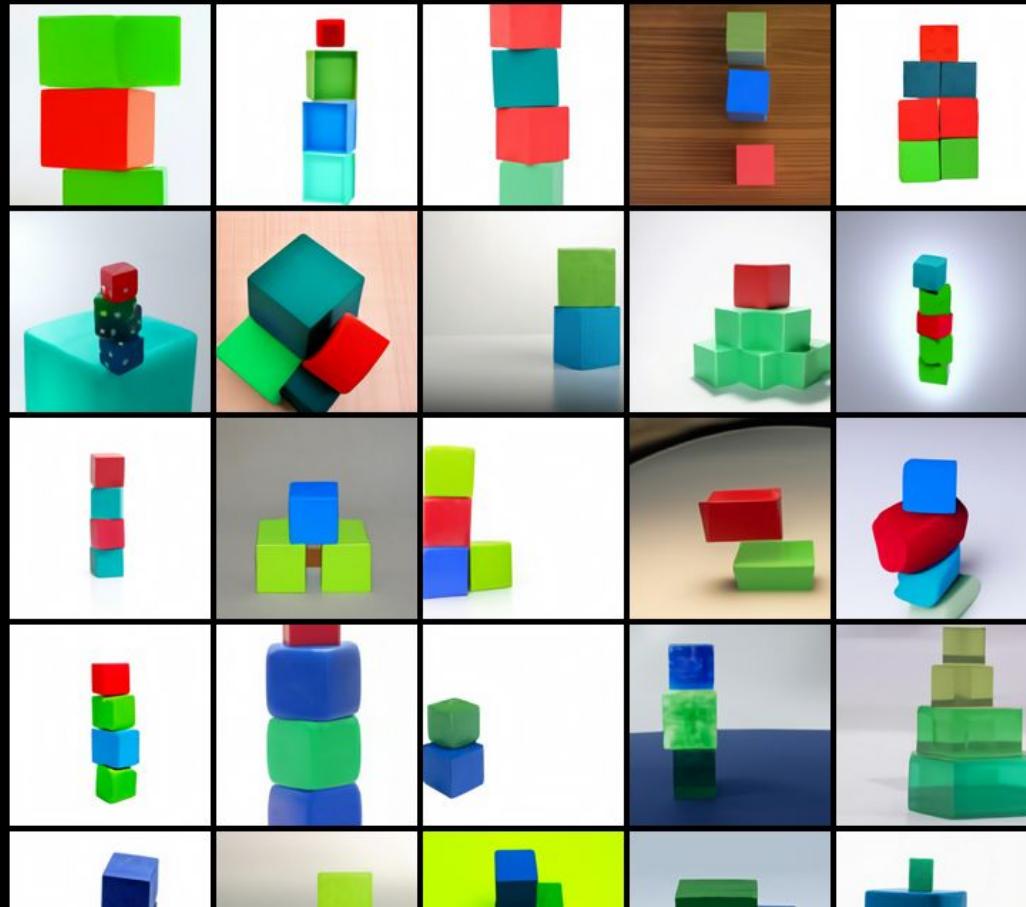


We find that DALL-E correctly identifies the types of relative position choices "sitting on" and "standing on" and sometimes appear to work well, while "standing behind," "standing right of" do not have a lower success rate when the object sitting on top of another compared to the other way around.

TEXT PROMPT

a stack of 3 cubes. a red cube is on the top, sitting on a green cube. the green cube is in the middle, sitting on a blue cube. the blue cube is on the bottom.

AI-GENERATED IMAGES



We find that DALL-E type image with one or two of correct colors. However, each setting tend to have colored precisely as spe

TEXT PROMPT

an emoji of a baby penguin wearing a blue hat, red gloves, green shirt, and yellow pants

AI-GENERATED
IMAGES



We find that DALL-E typically generates an image with two or three items having the correct color, while the samples for each set of articles of clothing with

Capabilities

Controlling attributes

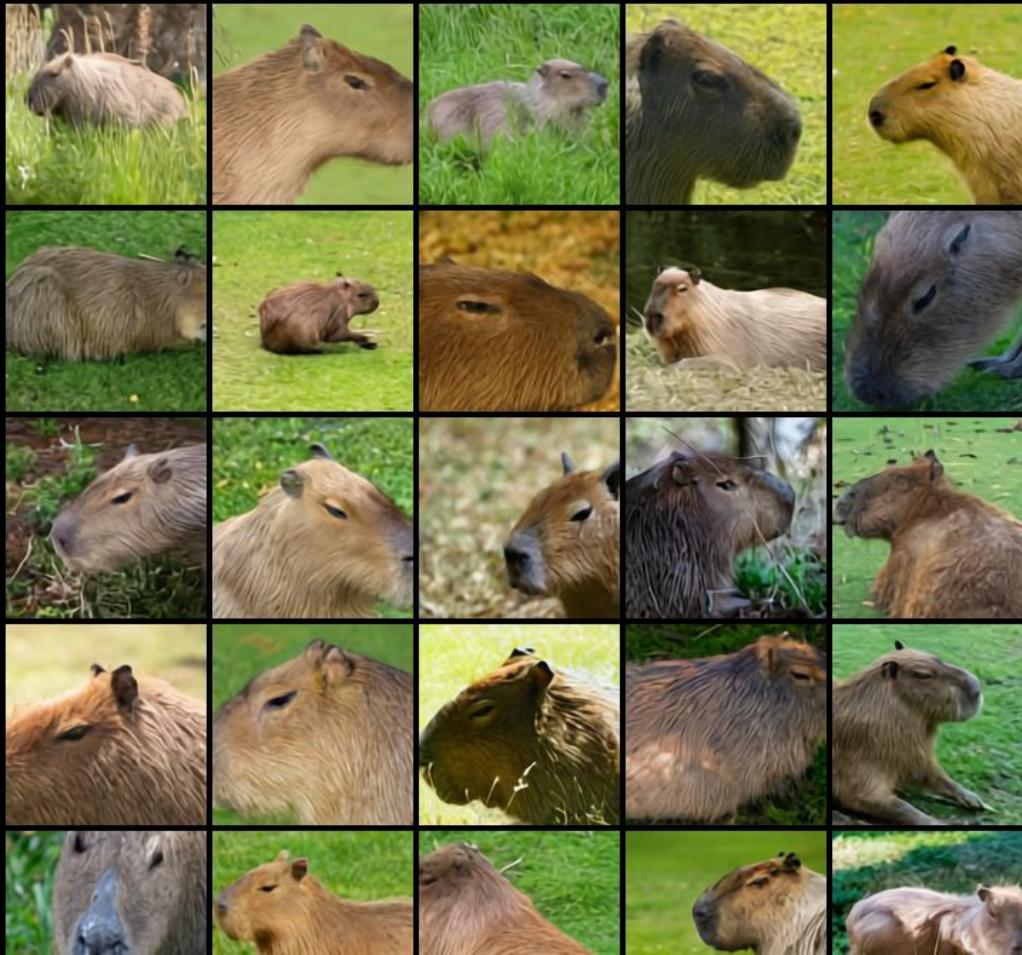
Drawing multiple objects

Perspective and 3 - dimensionality

TEXT PROMPT

an extreme close-up view of a capybara sitting in a field

AI-GENERATED
IMAGES

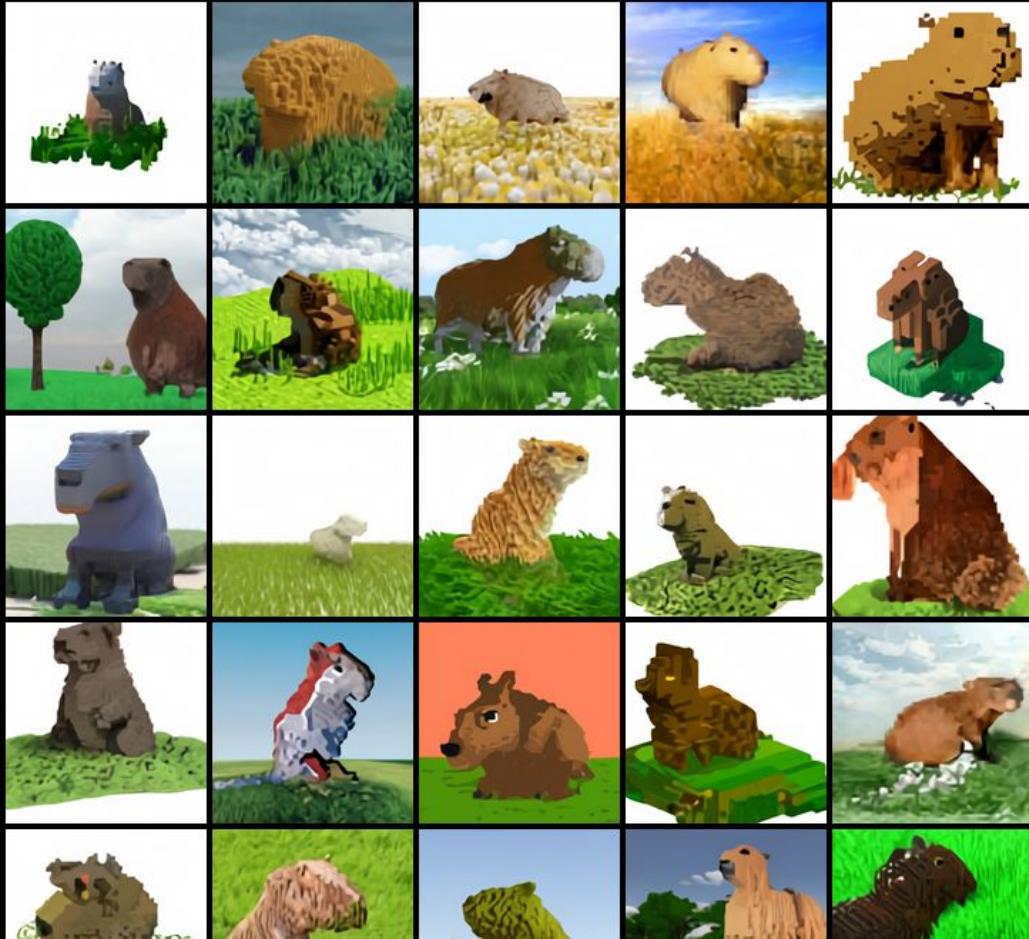


We find that DALL-E can generate images of animals in a variety of different views. These "extreme close-up" views, such as "an extreme close-up view of a capybara sitting in a field," require knowledge of the animal's appearance from unusual angles, as well as the fine-grained details of its fur.

TEXT PROMPT

a capybara made of voxels sitting in a field

AI-GENERATED
IMAGES



We find that DALL-E is able to generate a surface of each of the animals in the chosen 3D style, such as "made of voxels," and render plausible shading depending on the sun. The "x-ray" style does not work reliably, but it shows that DALL-E can orient the bones within the voxel volume (though not anatomical configurations).

TEXT PROMPT

a photograph of a bust of homer

IMAGE PROMPT



AI-GENERATED IMAGES

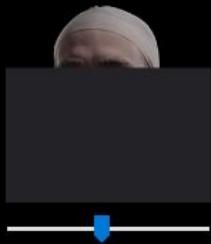


We prompt DALL-E with describing a well-known region of an image showing particular angle. Then, we complete the remaining this contextual information repeatedly, each time rotating more degrees, and find to recover smooth animation known figures, with each precise specification of lighting.

TEXT PROMPT

a photograph of a bust of homer

IMAGE PROMPT



AI-GENERATED IMAGES



We prompt DALL-E with describing a well-known region of an image shown particular angle. Then, we complete the remaining this contextual information repeatedly, each time rotating more degrees, and find recover smooth animation known figures, with each precise specification of lighting.

TEXT PROMPT

a plain white cube looking at its own reflection in a mirror. a plain white cube gazing at itself in a mirror.

IMAGE PROMPT



AI-GENERATED IMAGES



Similar to what was done with DALL-E to complete the sequence of frames, even though the mirror and reflective floor in the mirror usually resemble it, it often does not reflect it in a physically correct way. Below is a sequence of an object drawn on a surface, typically more plausible.

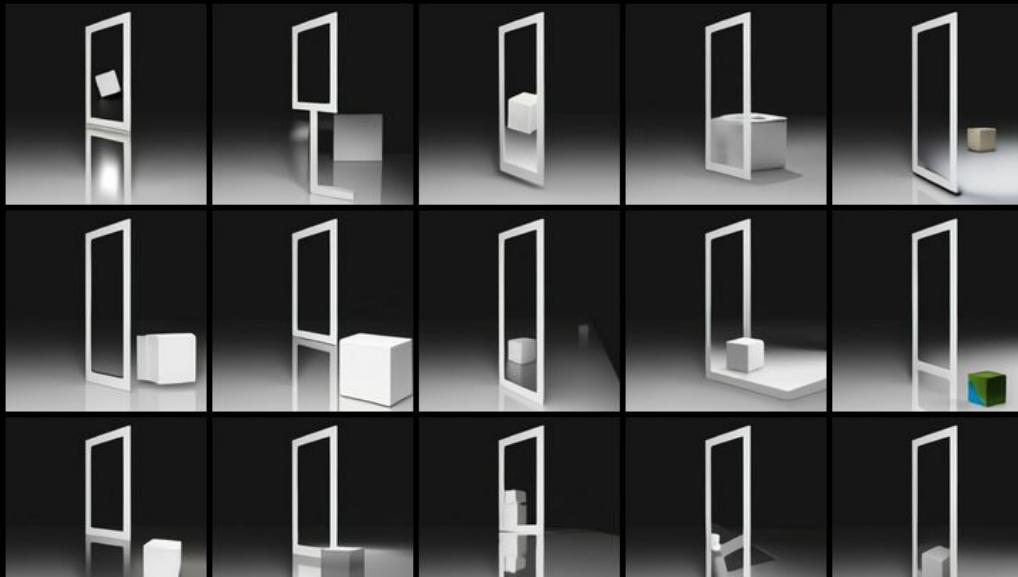
TEXT PROMPT

a plain white cube looking at its own reflection in a mirror. a plain white cube gazing at itself in a mirror.

IMAGE PROMPT



AI-GENERATED
IMAGES



Similar to what was done with DALL-E to complete the sequence of frames, even though the mirror and reflective floor are not physically correct, the mirror usually resembles it, it often does not reflect it in a physically correct way. Below is an example of an object drawn on a floor that looks more plausible than the one from DALL-E.

Capabilities

Controlling attributes

Drawing multiple objects

Perspective and 3 - dimensionality

Internal and external structures

TEXT PROMPT

a cross-section view of a walnut

AI-GENERATED
IMAGES



We find that DALL-E is a
of several different kind

TEXT PROMPT

a macro photograph of brain coral

AI-GENERATED
IMAGES



We find that DALL-E is able to generate external details of objects. These details are more pronounced as the object is viewed up close.

Capabilities

Controlling attributes

Drawing multiple objects

Perspective and 3 - dimensionality

Internal and external structures

Inferring contextual details

TEXT PROMPT

a ... of a capybara sitting in a field at sunrise

AI-GENERATED
IMAGES

painting



painting in pop art style



painting in cubist style



painting in surrealist style



painting in the style of van gogh



painting in the style
of claude monet



drawing



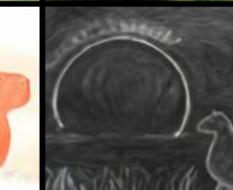
charcoal drawing



crayon drawing



chalk drawing



We find that DALL-E is able to generate a scene in a variety of different styles and adapt the lighting, shading, and composition based on the time of day.

TEXT PROMPT

a stained glass window with an image of a blue strawberry

AI-GENERATED
IMAGES



We find that DALL-E is a representation of the object, medium on which it is based, "a mural," "a soda can," and must change how it draws the angle and curvature. For "a stained glass window" it must alter the appearance of how it usually appears.

TEXT PROMPT

a store front that has the word 'openai' written on it. a store front that has the word 'openai' written on it. a store front that has the word 'openai' written on it.
openai store front.

AI-GENERATED
IMAGES



We find that DALL-E is able to generate images based on the text and adapt the writing style to match the context in which it appears. For example, "a car" and "a license plate" each require the application of different types of fonts, and "a nebula in the sky" require the application of a font that looks like it could be changed.

Generally, the longer the text prompt is, the more likely it is to be prompted to write, the longer it takes. We find that the success rate of generating parts of the caption are higher than the success rate of generating the whole caption. Sampling temperature for generating parts of the caption is decreased, although the success rate is still higher for simpler and less realistic prompts.

Capabilities

Controlling attributes

Drawing multiple objects

Perspective and 3 - dimensionality

Internal and external structures

Inferring contextual details

Applications of preceding capabilities

TEXT PROMPT

a male mannequin dressed in an orange and black flannel shirt and black jeans

IMAGE PROMPT



AI-GENERATED IMAGES



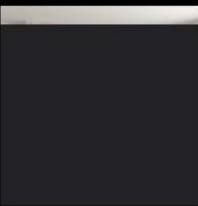
We explore DALL-E's ability to generate multiple variations of mannequins in a variety of styles. When prompted with two colors, such as orange and white, DALL-E generates a range of possibilities for how those colors can be used for the same article of clothing.

DALL-E also seems to often generate variations of items using common colors with other items. For example, when prompted with "orange" and "navy," DALL-E sometimes generates items that are blue, or shades very close to blue. DALL-E sometimes combines colors in ways that are not typical of brown or brighter shades.

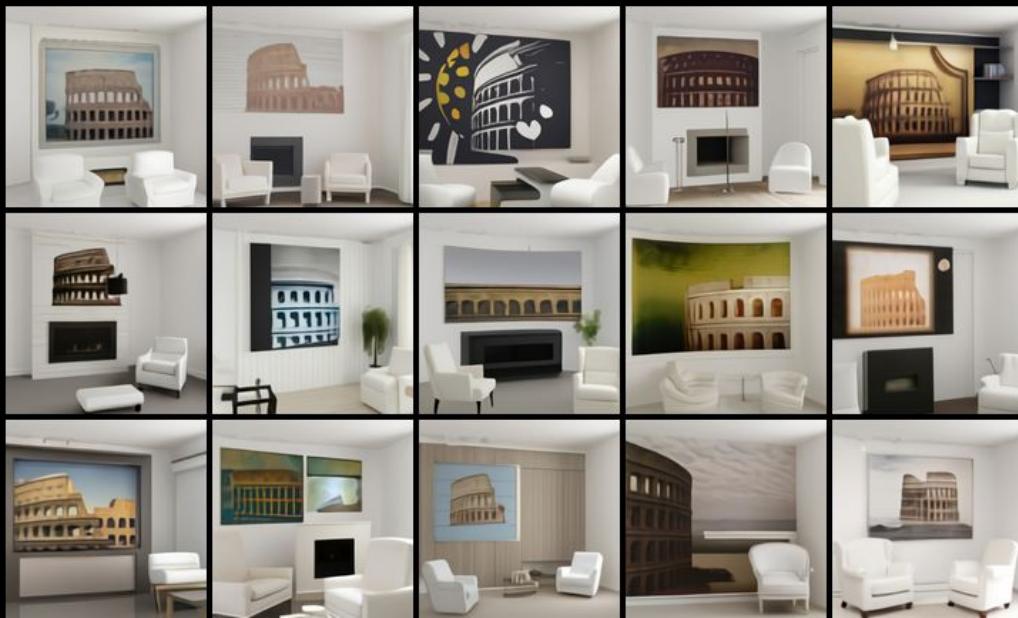
TEXT PROMPT

a living room with two white armchairs and a painting of the colosseum. the painting is mounted above a modern fireplace.

IMAGE PROMPT



AI-GENERATED IMAGES



We explore DALL-E's ability to generate images of rooms with several different subjects. It can generate paintings of many different subjects, including buildings like "the colosseum" and characters like "yoda." However, it exhibits a variety of interesting failures. Painting is almost always successful, but DALL-E sometimes fails to generate the correct number of objects.

Capabilities

Controlling attributes

Drawing multiple objects

Perspective and 3 - dimensionality

Internal and external structures

Inferring contextual details

Applications of preceding capabilities

Combining unrelated concepts

TEXT PROMPT

a snail made of harp. a snail with the texture of a harp.

AI-GENERATED
IMAGES



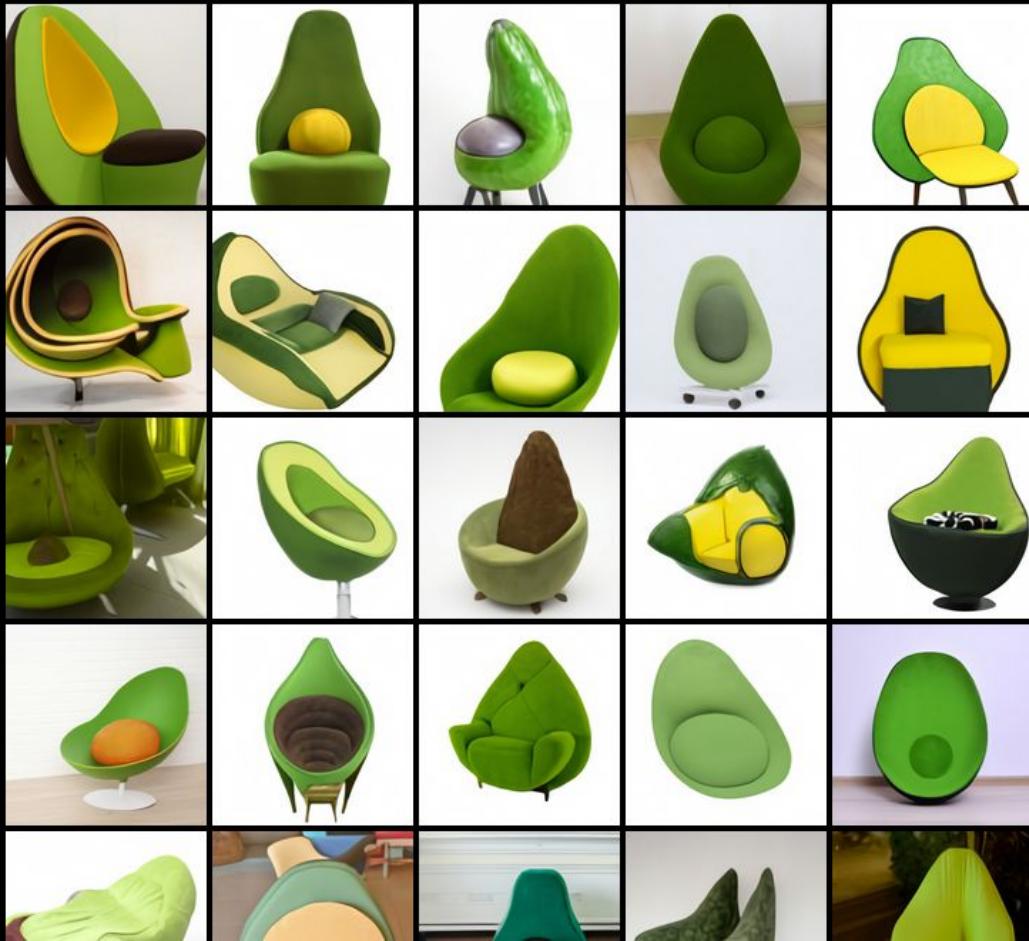
We find that DALL-E can synthesize from a variety of musical instruments, for items. While not always DALL-E sometimes takes objects into consideration how to combine them. For example, if we prompt to draw "a snail made of harp", DALL-E sometimes relates the prompt to the spiral of the snail's shell.

In a previous section, we saw how objects are introduced in a scene that are liable to confuse the algorithm about the objects and their specific roles. In this section, we see a different sort of failure, one where rather than binding some objects together to form a specified concept (say, a snail made of harp), DALL-E just draws two separate items.

TEXT PROMPT

an armchair in the shape of an avocado. an armchair imitating an avocado.

AI-GENERATED
IMAGES



In the preceding visual, DALL-E's ability to generate fantastical imagery by combining two unrelated concepts is evident. It shows its ability to take inspiration from a simple idea while respecting the constraints of the prompt. The generated images are designed, ideally produced, and appear to be practical. This demonstrates that prompting DALL-E with "in the shape of," "in the form of," or "imitating" a specific object gives it the ability to do so.

When generating some variations of the prompt "an armchair in the shape of an avocado," DALL-E appears to relate the concept of the avocado to the back of the chair. This means that the avocado is susceptible to the same constraints mentioned in the previous section.

Capabilities

...

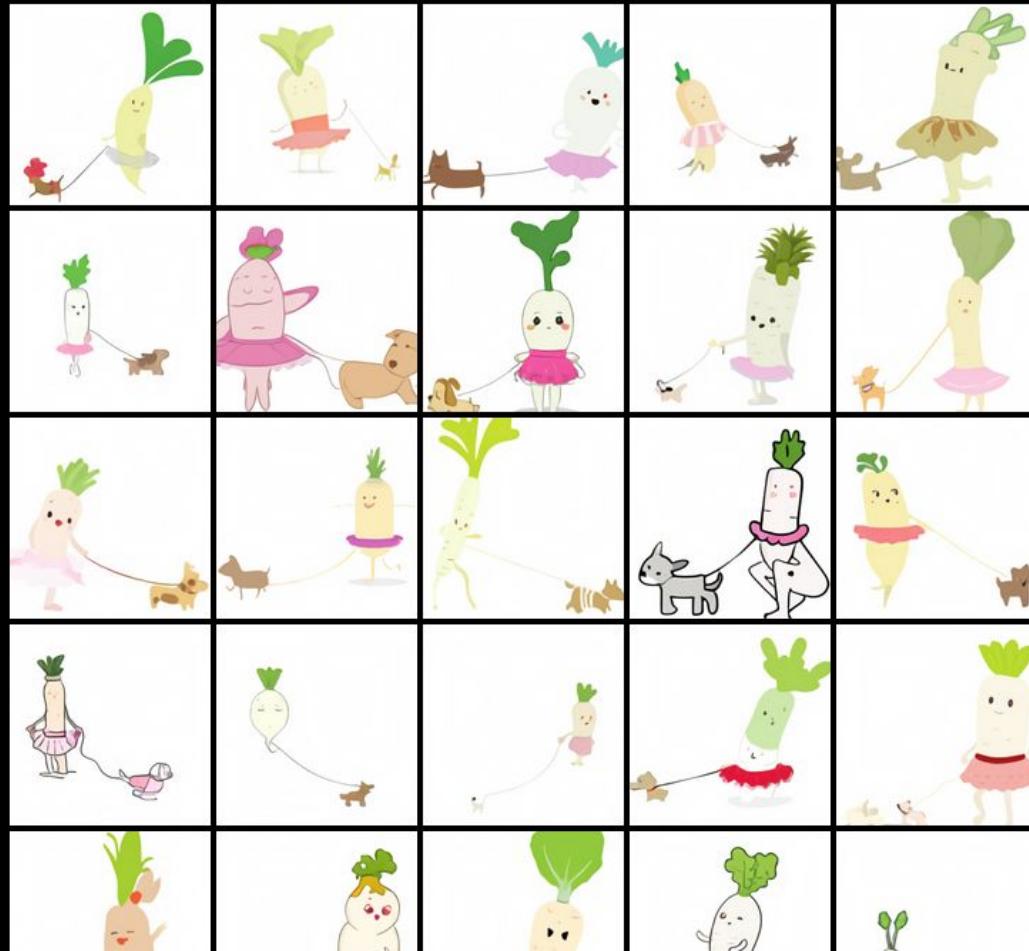
Combining unrelated concepts

Animal illustrations

TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED
IMAGES



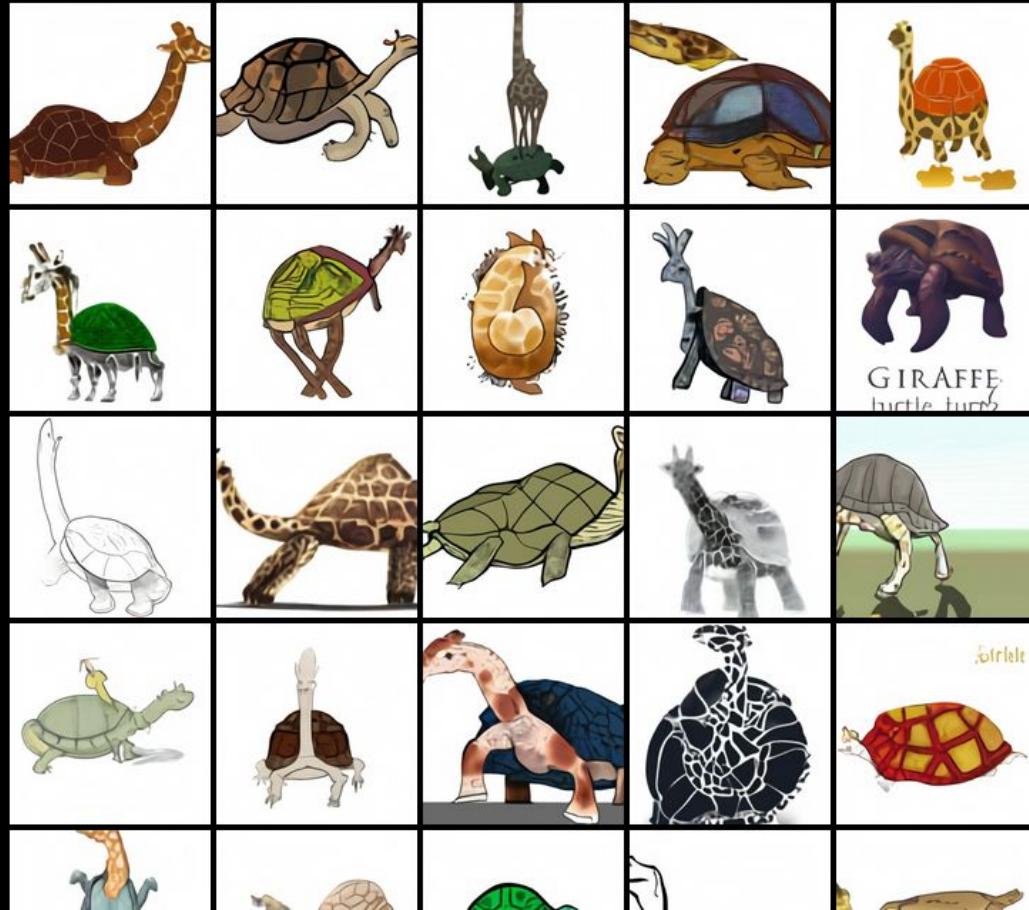
We find that DALL-E is able to transfer some human attributes such as clothing to animals and food items. We include "wielding a blue lightsaber" to demonstrate its ability to incorporate previously unseen concepts.

We find it interesting how DALL-E transfers body parts onto animals. For example, when asked to draw a daikon radish sipping a latte, or riding a motorcycle, DALL-E draws the kerchief, hand, and leg locations.

TEXT PROMPT

a professional high quality illustration of a giraffe turtle chimera. a giraffe imitating a turtle. a giraffe made of turtle.

AI-GENERATED IMAGES



We find that DALL-E is able to combine distinct animal features. For example, it can include "pikachu" to explore its ability to incorporate knowledge about "robot" to explore its ability to create cyborgs. Generally, the animal mentioned in the prompt is dominant.

We also find that inserting "professional high quality" and "emoji" sometimes improves the consistency of the results.

TEXT PROMPT

a professional high quality emoji of a lovestruck cup of boba

AI-GENERATED
IMAGES



We find that DALL-E is able to transfer some emojis to other objects, such as food items. In visual, we find that inserting "professional high quality" sometimes improves the consistency of the results.

Capabilities

...

Combining unrelated concepts

Animal illustrations

zero shot visual reasoning

TEXT PROMPT

the exact same cat on the top as a sketch on the bottom

IMAGE PROMPT



AI-GENERATED IMAGES



We find that DALL-E is able to generate a variety of image transformations, ranging from straightforward ones, such as "pink" and "photo reflect", to more complex ones like "animal view". These transformations tend to be the most reliable, as they are often not copied or recombined. The transformation "animal view" requires DALL-E to correctly identify the animal in the photo, and it does this with the appropriate degree of reliability, and for several categories of animals. DALL-E only generates plausible images for categories it has seen before.

TEXT PROMPT

the exact same teapot on the top with 'gpt' written on it on the bottom

IMAGE PROMPT



AI-GENERATED
IMAGES



We find that DALL-E can generate different kinds of images from the same text prompt. For example, given a prompt like "the exact same teapot on the top with 'gpt' written on it on the bottom", we find that DALL-E generates a grid of 10 images. The first three rows show red teapots with the word "gpt" written on them in different styles (cursive, bold, and block letters). The fourth row shows a variety of other teapots, including a white teapot, a green teapot, and a grey teapot, all with the word "gpt" written on them. This demonstrates that DALL-E can map the letters onto the teapot in a plausible way. While this is impressive, it can also lead to unreliable results. For example, if we ask DALL-E to generate a "tiny" teapot, it will often produce a small teapot, even though the prompt specifies a large teapot. Similarly, if we ask DALL-E to generate a "broken" teapot, it will often produce a teapot that is not actually broken, even though the prompt specifies a broken teapot.

TEXT PROMPT a sequence of geometric shapes. [Set D]

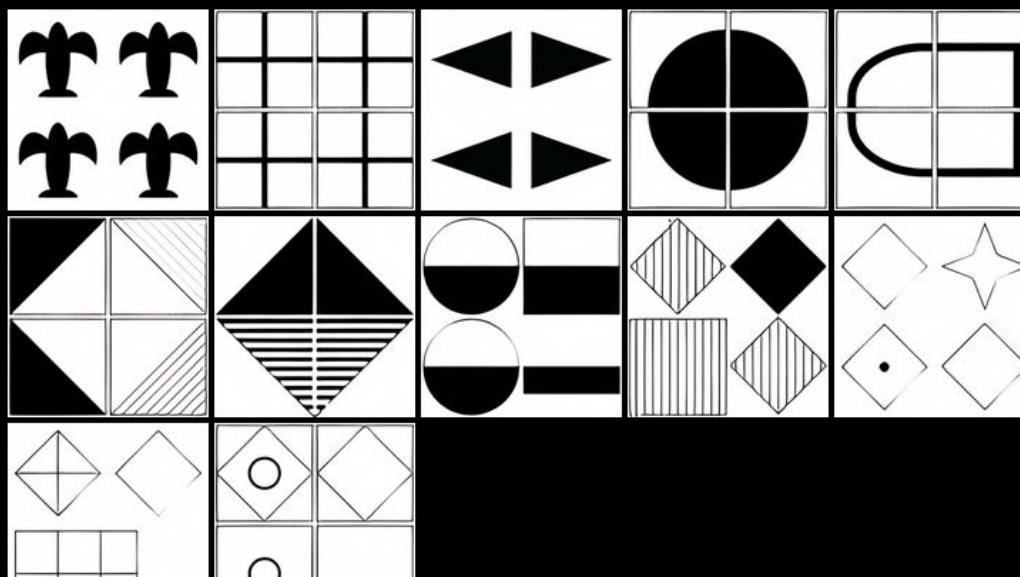
EXAMPLE IMAGE PROMPT



Set B

■ Inverted

AI-GENERATED IMAGES



Rather than treating the choice problem as original, DALL-E to complete the each image using arguments consider its completion's close visual match to the

DALL-E is often able to solve problems that involve continuing simple geometric reasoning, such as Set C. It is sometimes able to solve problems that involve recognizing permutations and boolean operations, such as Set E. In some instances in set E tend to be very similar to the images in Set D, and DALL-E gets almost all of them right.

For each of the sets, we can see that DALL-E's performance on both the completion task and the matching task is quite good. This is true for all three sets, and it suggests that DALL-E is able to learn a lot from the examples it is given.

Capabilities

...

Combining unrelated concepts

Animal illustrations

zero shot visual reasoning

Geographic and temporal knowledge

TEXT PROMPT

a photo of the food of china

AI-GENERATED
IMAGES



We test DALL-E's understanding of geographical facts, such as cuisines, and local wildlife. DALL-E successfully answers many questions, such as those involving "China," which reflects superficial stereotypes about "food" and "wildlife," as well as the full diversity encountered in China.

TEXT PROMPT

a photo of alamo square, san francisco, from a street at night

AI-GENERATED
IMAGES

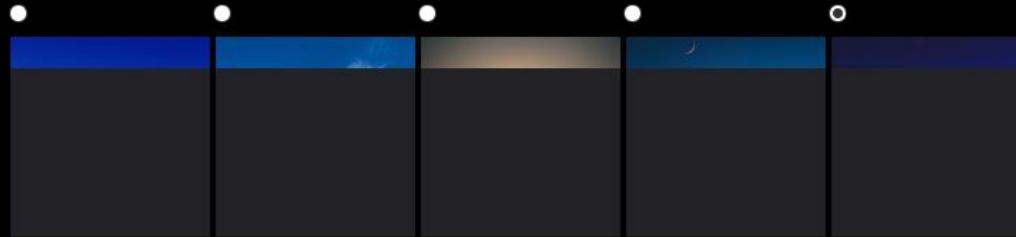


We find that DALL-E is rendering semblances of San Francisco. For locations such as San Francisco, déjà vu—eerie simulacra and cafes that remind us of locations that do not exist.

TEXT PROMPT

a photo of san francisco's golden gate bridge

IMAGE PROMPTS



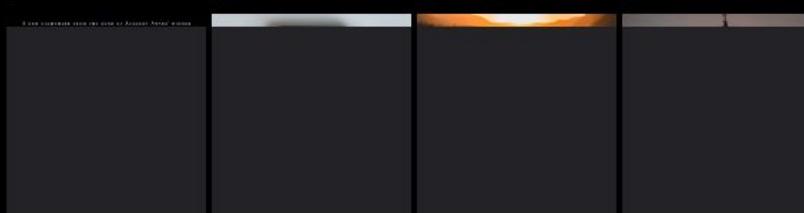
AI-GENERATED IMAGES



We can also prompt DALL-E to generate images of landmarks. In fact, we can even tell it exactly where the photo was taken by specifying the coordinates of the rows of the sky. When the prompt includes both a location and a landmark, DALL-E can correctly identify the landmark and its surroundings.

TEXT PROMPT a photo of a phone from the ...

IMAGE PROMPTS

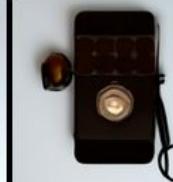


AI-GENERATED IMAGES

1900



1910s



20s



30s



40s



We find that DALL-E has stereotypical trends in output over the decades. Technologies tend to go through periods of stability, dramatically shifting form factor and becoming more incremental and streamlined.

50s



60s



70s



80s



90s



Source

<https://openai.com/blog/dall-e/>

Thanks

