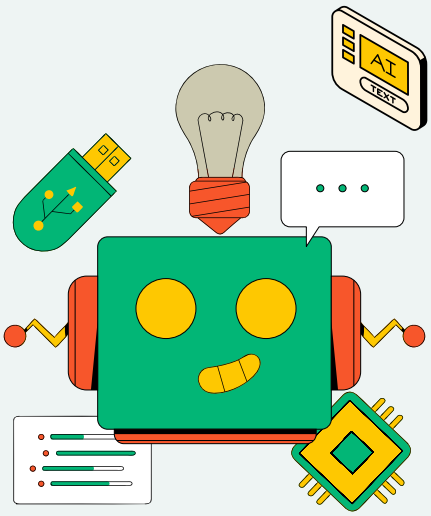
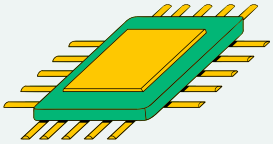


A TRIPLE OF IMAGE GENERATION TOOL

PRESENTATION

PRESENTED BY: KHTM1-08

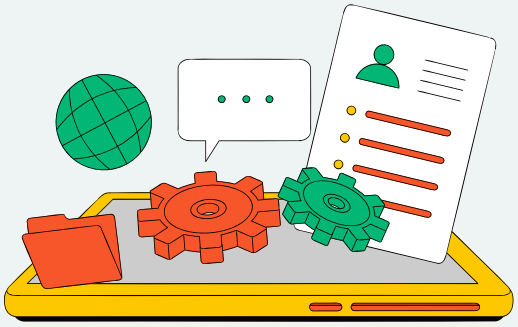


THE THREE IMAGE GENERATION TOOLS

Imagen

MidJourney

Dall-E



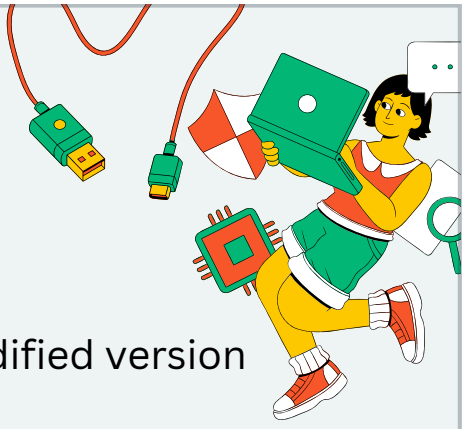
PRESENTATION OUTLINE

- Introduction
- What is image generation tool ?
- OpenAI DALL-E
- Google Imagen
- Midjourney
- Usage test
- Comparison
- Conclusion

INTRODUCTION

Imagine a world where machines can decipher text from all languages and then generate image from it.

We will embark on a journey through the ever-evolving landscape of Artificial Intelligence. From there, we can understand more about the image generation AI by learning how it work, what is used for and how it can help this generation.

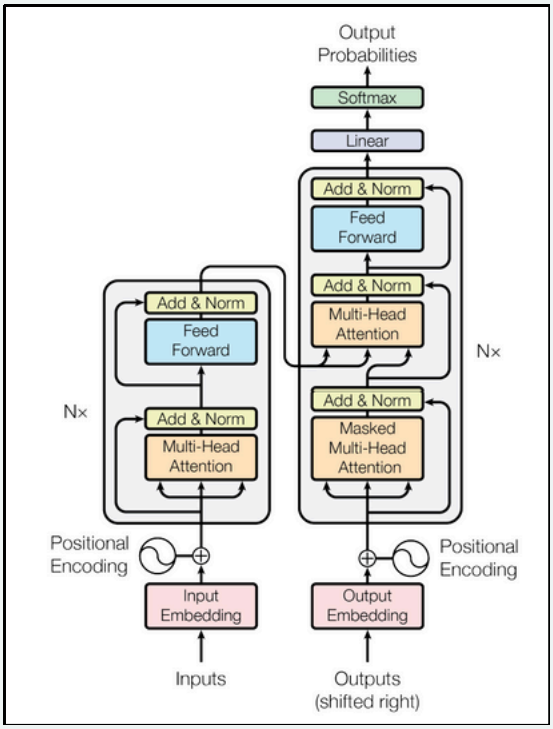


DALL-E

Developed by OpenAI in 2021 by using a modified version of GPT to generate images.

Latest: DALL-E 3 (Aug 2023, currently used for ChatGPT), implemented for Node.js and Python libraries.

Note: DALL-E 2 supports image edit and variation generation, but not DALL-E 3.

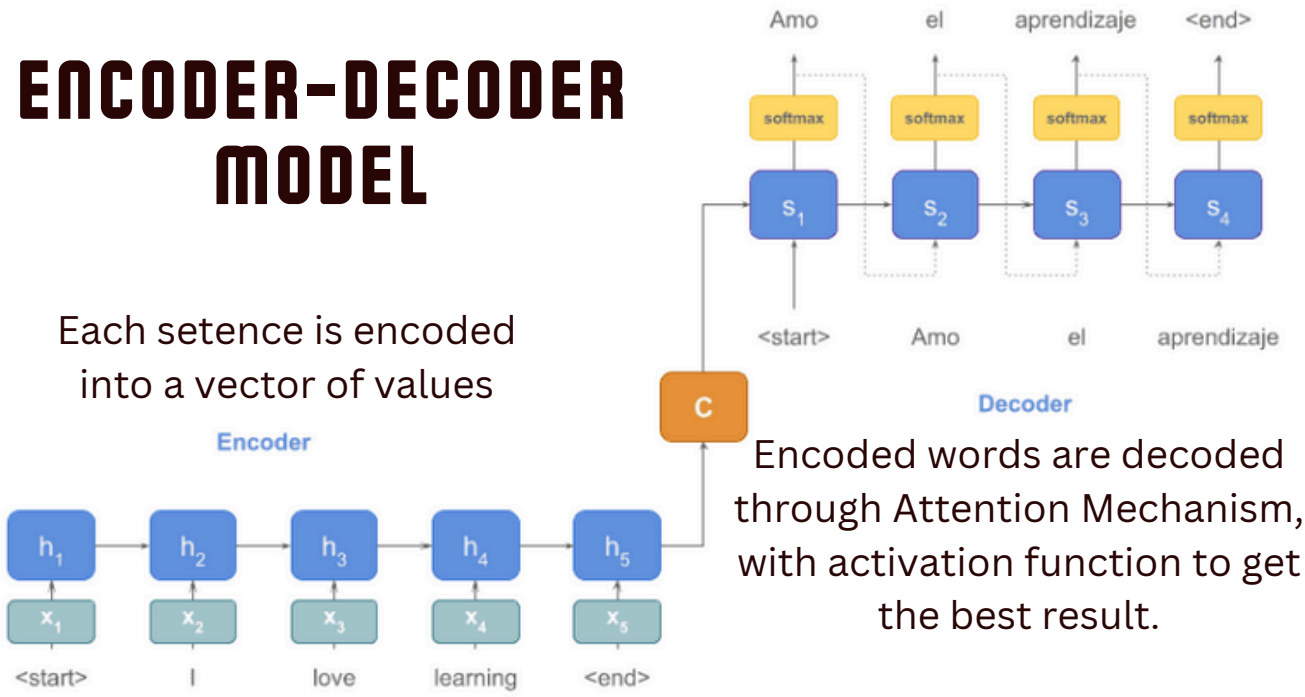


DALL-E 1: GPT MODEL

GPT: large language model (LLM) case of Transformer-based Encoder-Decoder model, which is based on Attention mechanism.

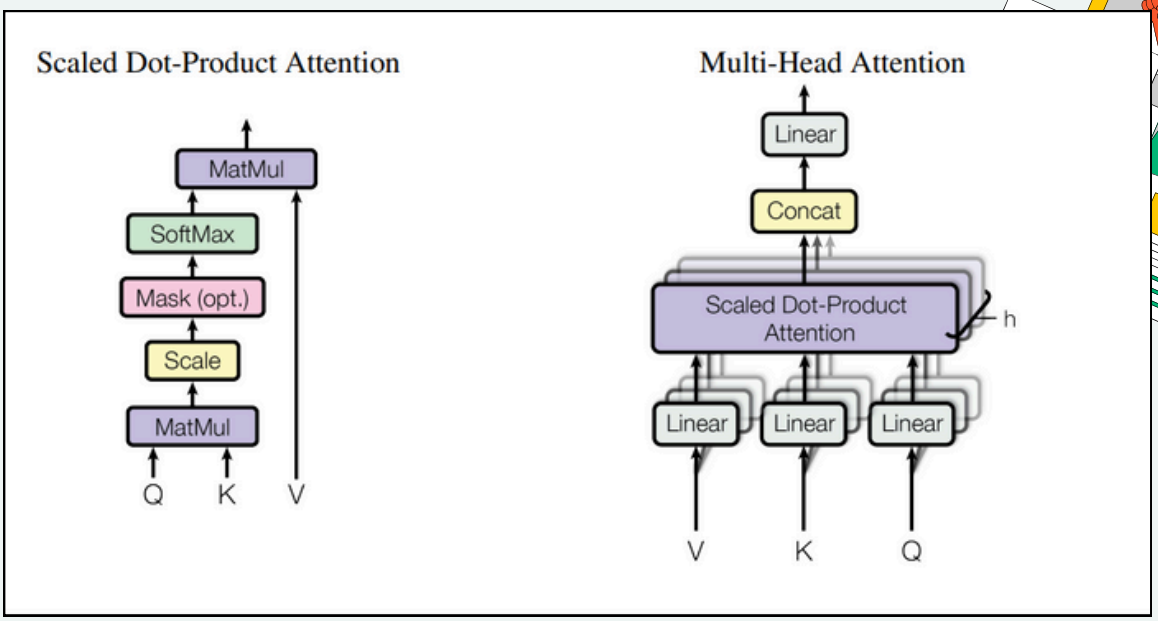
ENCODER-DECODER MODEL

Each setence is encoded into a vector of values



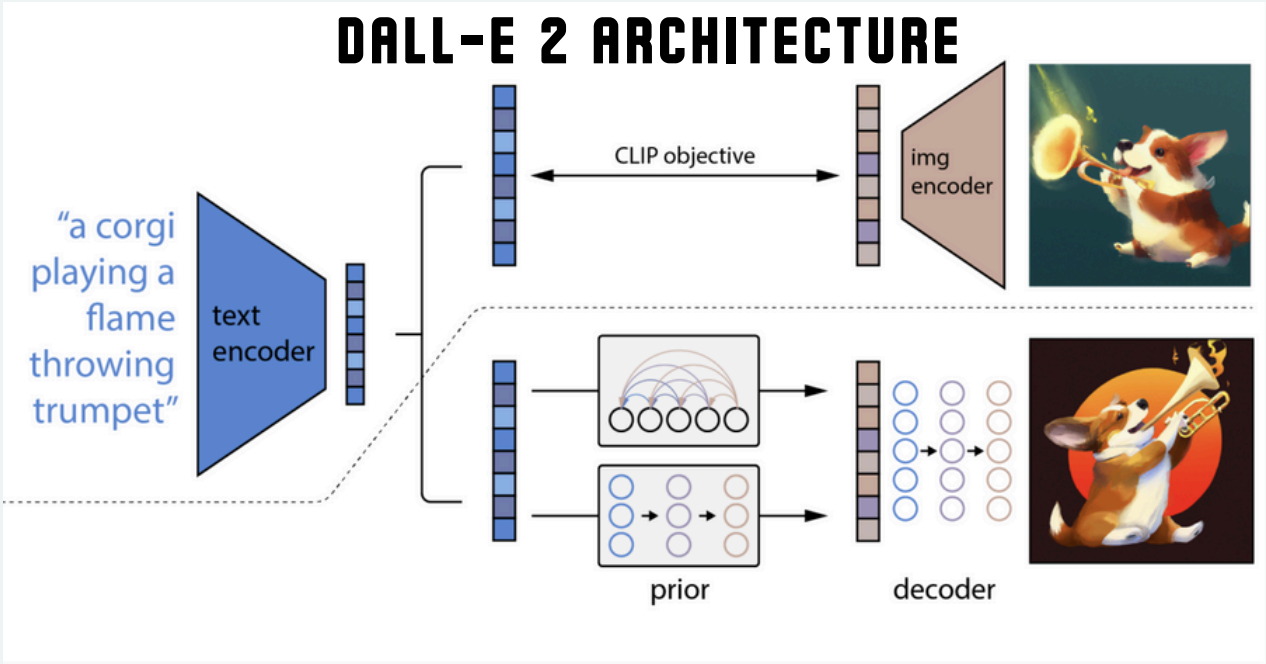
Encoded words are decoded through Attention Mechanism, with activation function to get the best result.

ATTENTION MECHANISM



Encoding: A sentence (input) generates a vector of queries (Q), keys (K) and values (V).

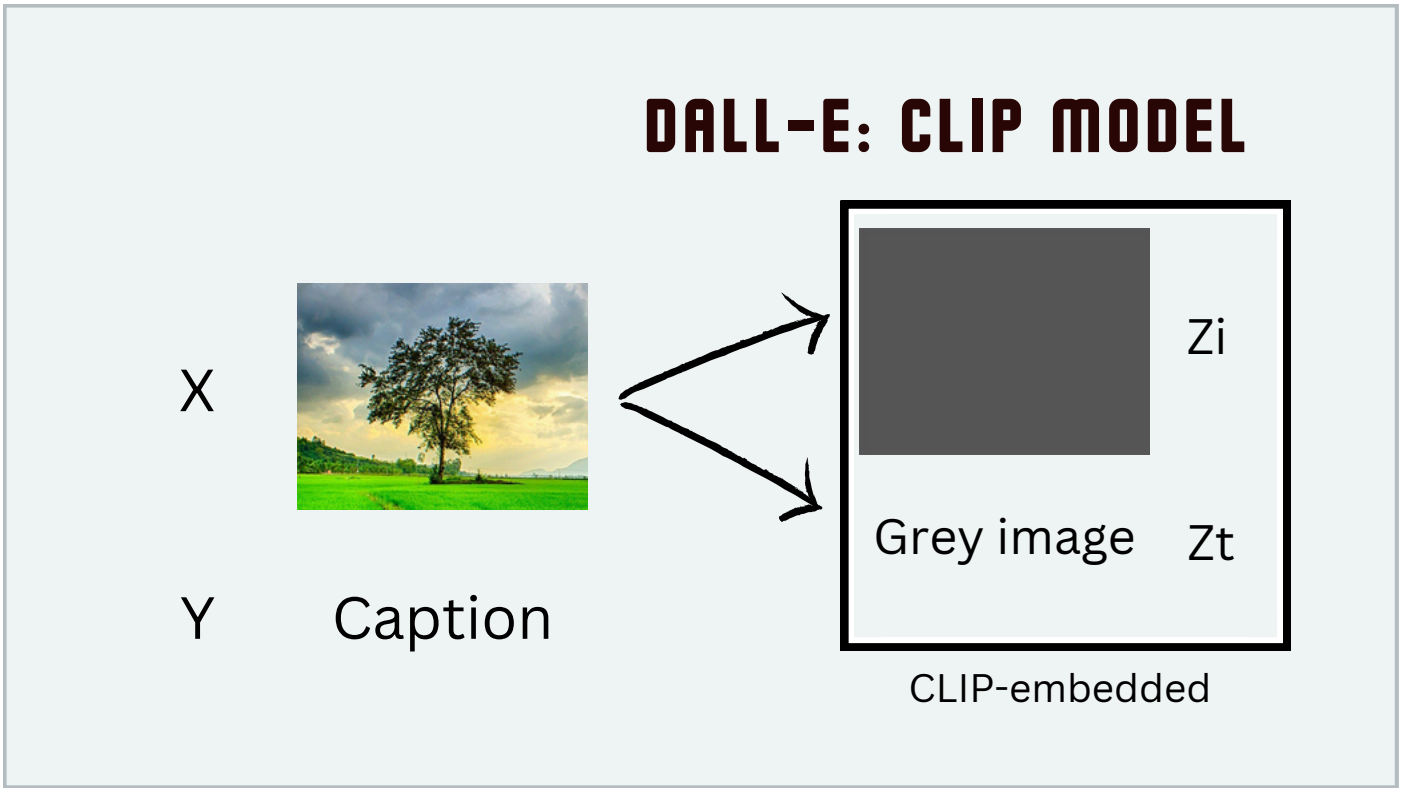
$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O$$
$$\text{where } \mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$


DALL-E 2 ARCHITECTURE
"a corgi playing a flame throwing trumpet"
text encoder
CLIP objective
img encoder
prior
decoder

HOW DO WE GUIDE THEM?

GUIDE IS IMPROVED BY INCORPORATING CLIP IMAGE EMBEDDINGS:

- A CLIP MODEL IS TRAINED TO UNDERSTAND THE RELATIONSHIP BETWEEN TEXT DESCRIPTIONS AND IMAGES.
- THIS MODEL GENERATES "IMAGE EMBEDDINGS" THAT CAPTURE THE ESSENCE OF AN IMAGE.
- THESE EMBEDDINGS, ALONG WITH THE ENCODED TEXT DESCRIPTION (OPTIONAL), ARE FED INTO THE DIFFUSION MODEL (DECODER) TO GUIDE THE IMAGE GENERATION PROCESS.
- THIS ALSO ALLOWS FOR INTERESTING FEATURES:
  - CREATING VARIATIONS OF EXISTING IMAGES.
  - GENERATING NEW IMAGES BASED ON BLENDS BETWEEN TWO IMAGES OR CAPTIONS.


DALL-E: CLIP MODEL
X
Y
Caption
Zi
Grey image
Zt
CLIP-embedded

DALL-E: CLIP MODEL

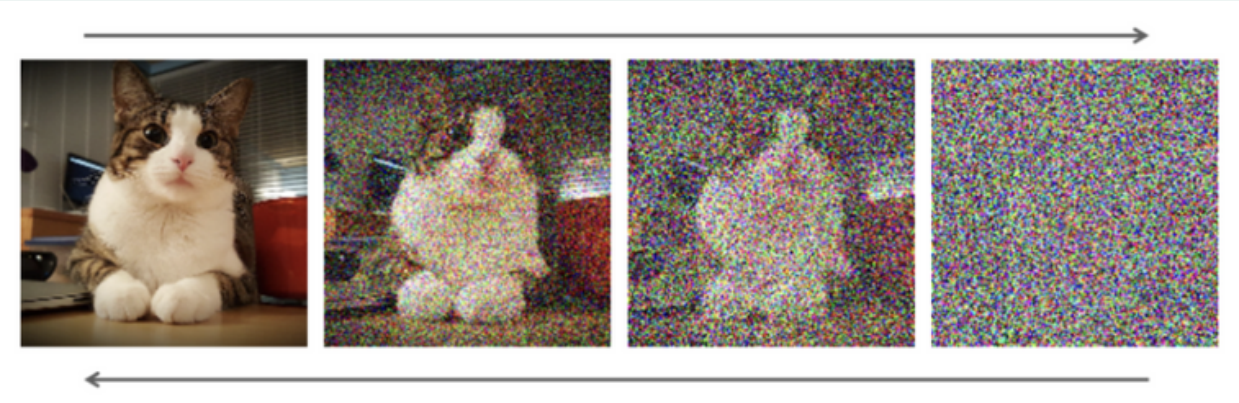
- A *prior* $P(z_i|y)$ that produces CLIP image embeddings z_i conditioned on captions y.
- A *decoder* $P(x|z_i, y)$ that produces images x conditioned on CLIP image embeddings z_i (and optionally text captions y).

$$P(x|y) = P(x, z_i|y) = P(x|z_i, y)P(z_i|y).$$

Diffusion model

HOW DO DIFFUSION MODELS WORK?

DIFFUSION MODELS WORK BY ADDING NOISE TO AN IMAGE IN SMALL STEPS, TRAINING A MODEL TO REVERSE THIS PROCESS. THIS ALLOWS THEM TO GENERATE ENTIRELY NEW IMAGES BY STARTING WITH RANDOM NOISE AND REVERSING THE NOISE ADDITION PROCESS.

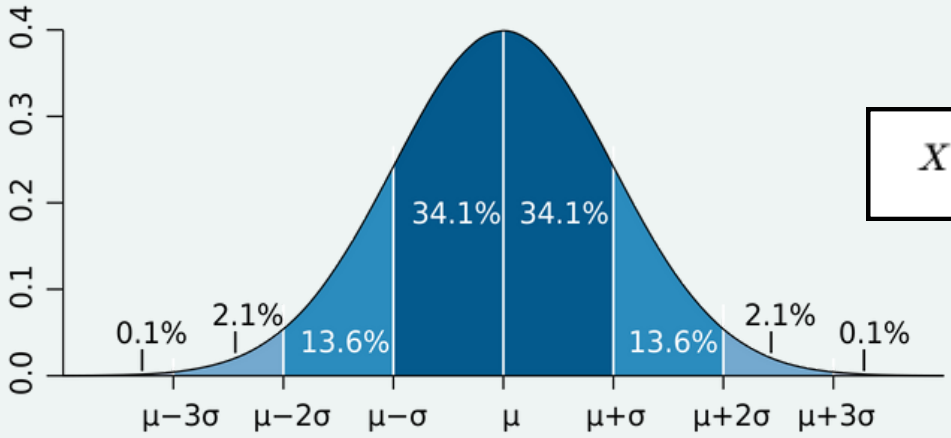


HOW DO WE GUIDE THEM?

CONDITIONED DIFFUSION: INJECTS ADDITIONAL INFORMATION LIKE TEXT DESCRIPTIONS OR CLASS LABELS INTO THE MODEL TO STEER THE IMAGE GENERATION TOWARDS SOMETHING RELEVANT.

CLASSIFIER GUIDANCE: USES A CLASSIFIER TO NUDGE THE NOISE ADDITION PROCESS IN EACH STEP, MAKING SURE THE FINAL IMAGE ALIGNS WITH THE DESIRED CLASS.

NORMAL DISTRIBUTION

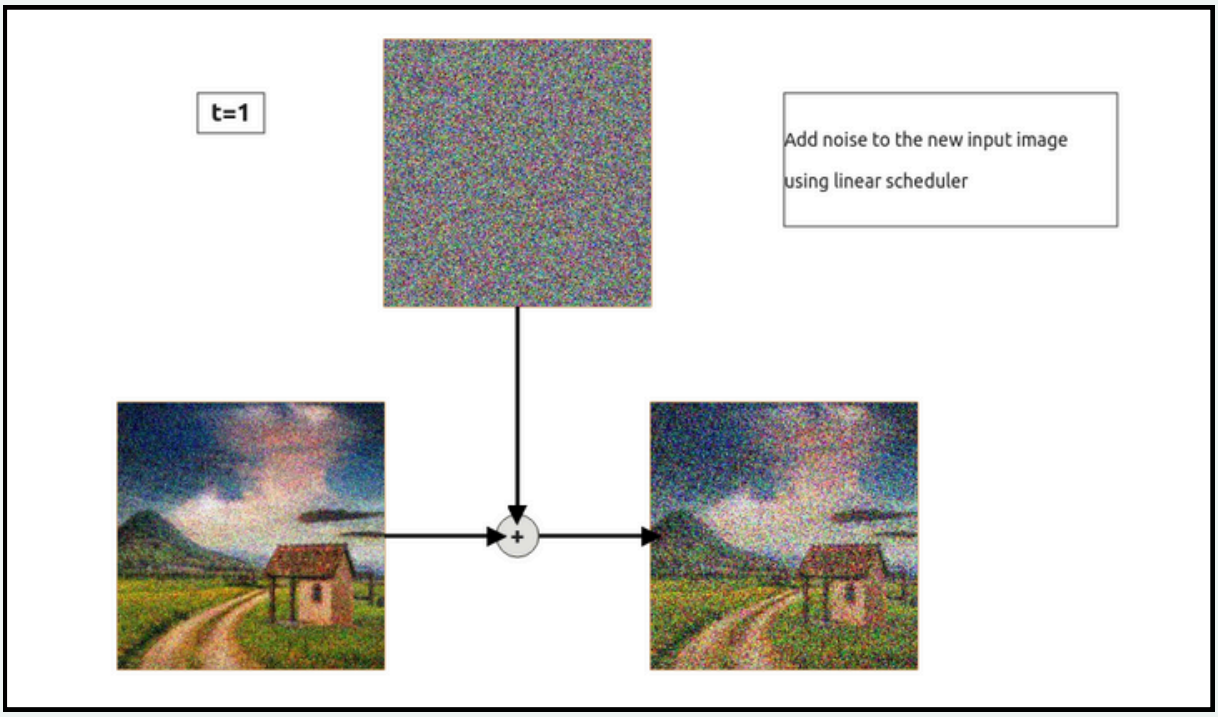


$$X \sim \mathcal{N}(\mu, \sigma^2).$$

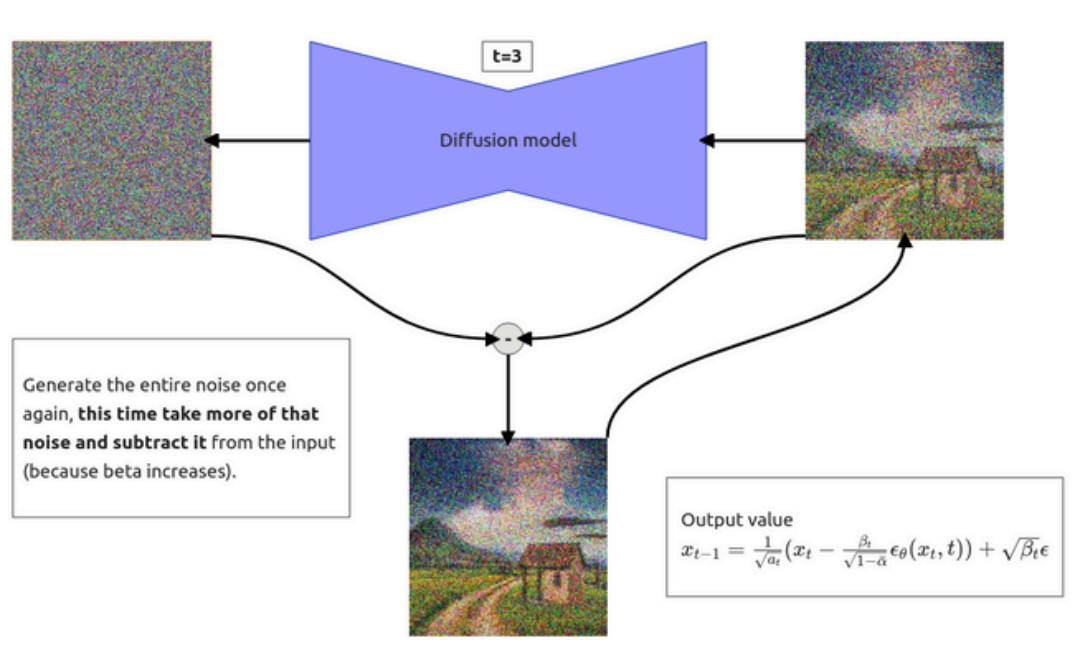
FORWARD DIFFUSION

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Normal distribution function helps (partially) keeping the approximate original texture, or information, even if the output looks random.


t=1
Add noise to the new input image using linear scheduler
+

REVERSE DIFFUSION



IMAGEN

Introduction

- Created by Google
- It is a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding.
- It builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation.

IMAGEN REVIEWS

IMAGE QUALITY

Can generate images with impressive realism, capture intricate details and textures

FLEXIBILITY

Users can input diverse sketches, from basic outlines to more complex compositions.

SPEED

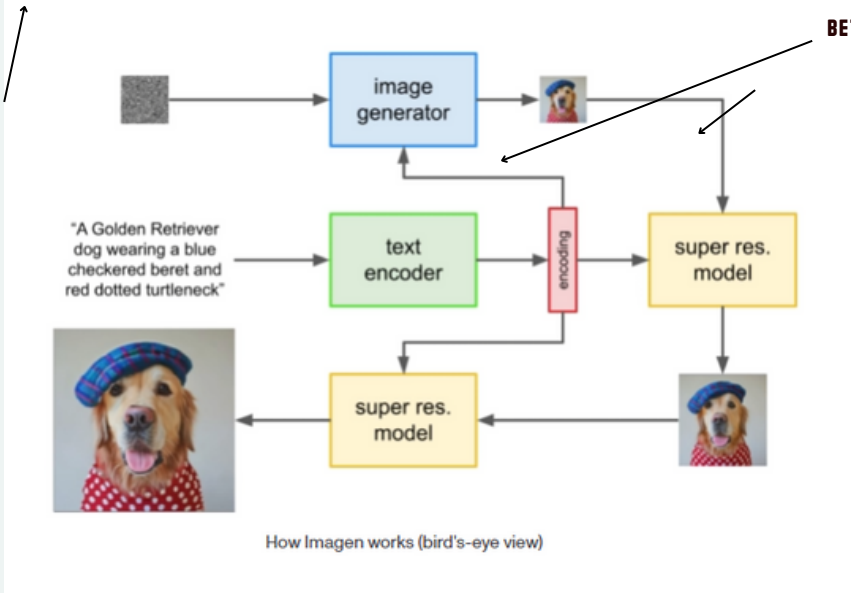
The system operates efficiently, producing high-quality images in a fraction of the time it would take an artist.

HOW DOES IMAGEN WORK ?

START WITH RANDOM NOISE

BETTER RES MODEL RESULT

GROWS THE IMAGE TO A HIGHER RESOLUTION



TEXT-TO-MODEL EXPLAINED

IMAGEN IMPROVES ON CLIP BY SIMPLY SWAPPING ITS CUSTOM-TRAINED TEXT ENCODER FOR A GENERIC LARGE LANGUAGE MODEL PRE-TRAINED ON TEXT-ONLY CORPORA.

IMPORTANT NOTE: **THE TEXT-ENCODER DOES NOT LEARN OR CHANGE THE WAY IT CREATES THE ENCODINGS.**

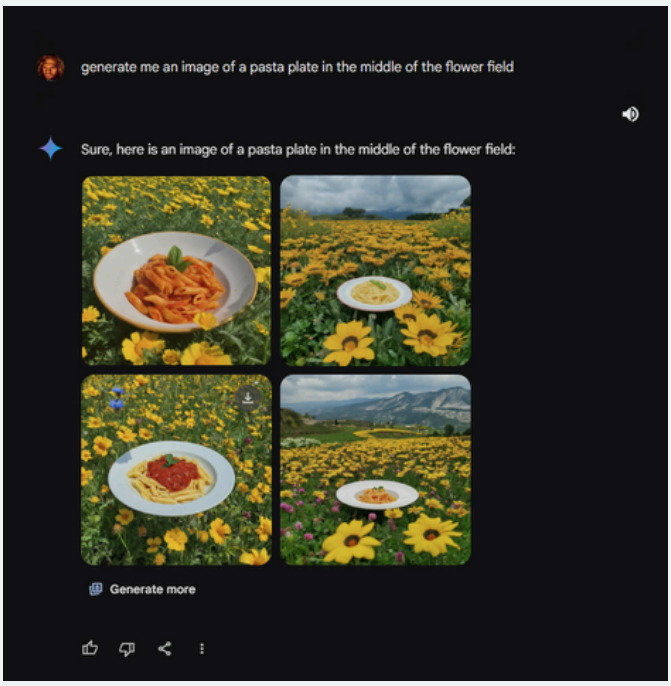

generate me an image of a pasta plate in the middle of the flower field
Sure, here is an image of a pasta plate in the middle of the flower field:
Generate more

IMAGEN UI

DALL·E results for the caption "An armchair in the shape of an avocado". Source: OpenAI's DALL·E blogpost.

DALL·E 2 results for the caption "An armchair in the shape of an avocado". Generated with OpenAI's DALL·E 2 beta.

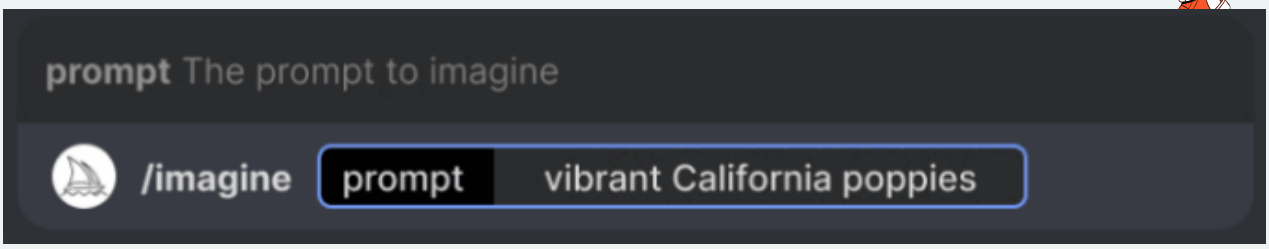
Stable Diffusion results for the caption "An armchair in the shape of an avocado". Generated locally in Tryolabs' hardware.

MIDJOURNEY

- Midjourney is created by Midjourney, Inc. in July 2022.
- Same as DALL-E, it's based on large transformer language models to understand prompts and diffusion model to generate images.
- Currently in April 2024, Midjourney is closed-source and requires paid subscription to generate images.

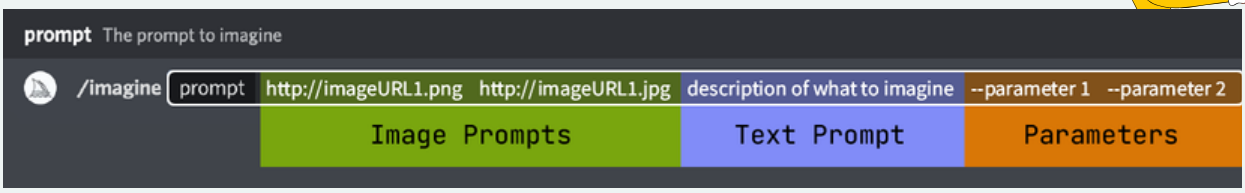
MIDJOURNEY

- Availiable via Discord
- Use /imagine command to begin creating images
-



MIDJOURNEY

- Advanced prompting

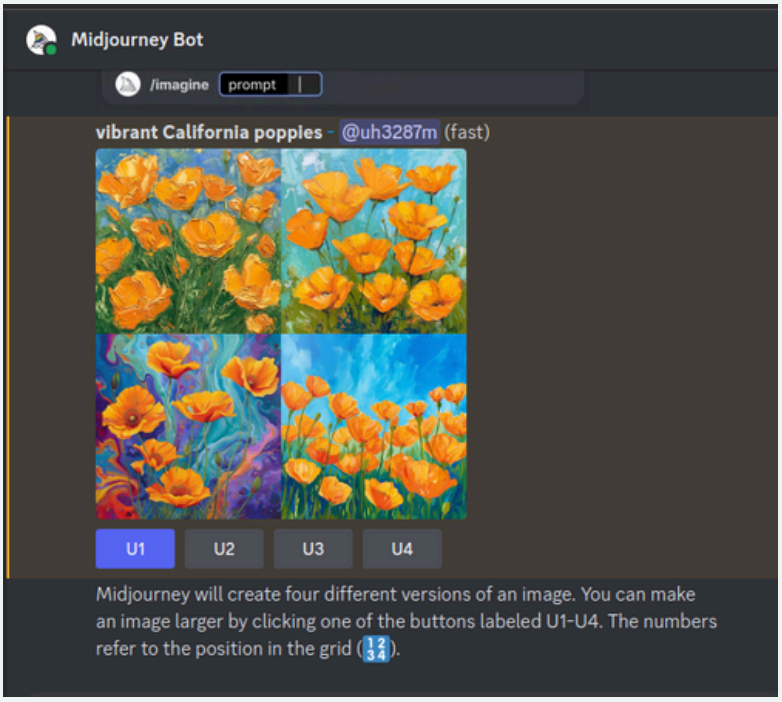


coding stuff > something > Python > temp.py > ...

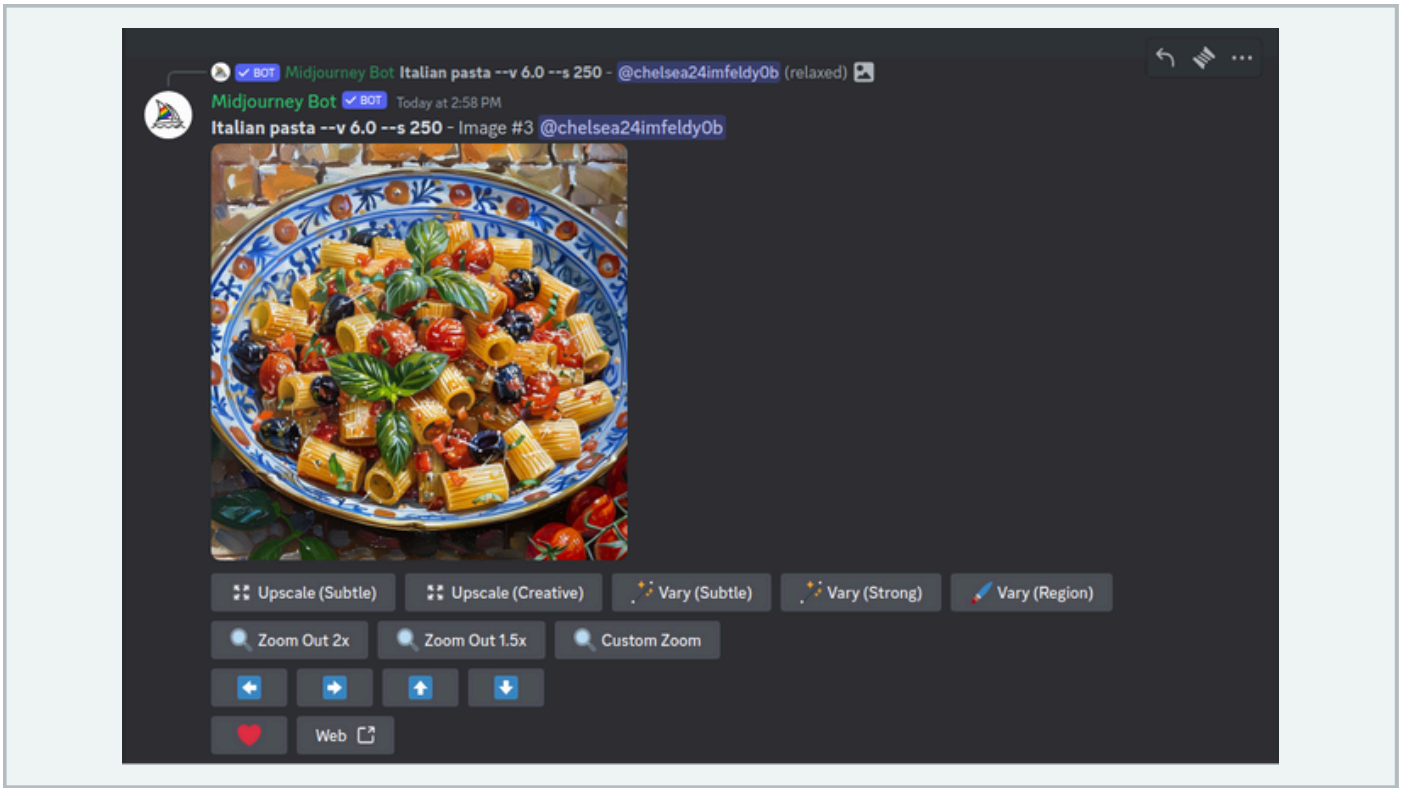
```python
from openai import OpenAI
import os
import requests
from PIL import Image

# setup credentials
client = OpenAI(
    api_key=os.environ.get("MY_OPENAI_APIKEY"),
)
# request
response = client.images.generate(
    model="dall-e-2",
    prompt="Pacman chasing Superman",
    size="1024x1024",
    quality="standard",
    n=5,
)
# output
for i in range(0, 5):
    img_url = response.data[i].url
    img = requests.get(img_url).content
    with open("image_" + str(i) + ".png", "wb") as image_file:
        image_file.write(img)
```

DALL-E


Midjourney Bot
/imagine
prompt
vibrant California poppies - @uh3287m (fast)
U1
U2
U3
U4
Midjourney will create four different versions of an image. You can make an image larger by clicking one of the buttons labeled U1-U4. The numbers refer to the position in the grid (1 2 3 4).

MIDJOURNEY


BOT Midjourney Bot Italian pasta --v 6.0 --s 250 - @chelsea24imfeldy0b (relaxed)
Midjourney Bot BOT Today at 2:58 PM
Italian pasta --v 6.0 --s 250 - Image #3 @chelsea24imfeldy0b
Upscale (Subtle)
Upscale (Creative)
Vary (Subtle)
Vary (Strong)
Vary (Region)
Zoom Out 2x
Zoom Out 1.5x
Custom Zoom
Web

coding stuff > something > Python > midjourney.py > ...

```python
import requests

X_API_KEY = "MjY3MjE4NTIyMzU3MjM1NzEy.GjSZHJ.u_2GGDRJU-8gR_Z1IF_-foKE6buRktEa1eQ6QM"

endpoint = "https://api.midjourneyapi.xyz/mj/v2/imagine"

headers = {
    "X-API-KEY": X_API_KEY
}

data = {
    "prompt": "a cute cat",
    "aspect_ratio": "4:3",
    "process_mode": "fast",
    "webhook_endpoint": endpoint,
    "webhook_secret": X_API_KEY
}

response = requests.post(endpoint, headers=headers, json=data)

print(response.status_code)
print(response.json())
```

	AVAILABILITY	PERFORMANCE	RESULT	APPLICATION
IMAGEN	Not yet available	Excels in detail and realism, potentially the best among the three.	Hyper-realistic images with exceptional focus on accuracy.	Ideal for projects requiring extreme photorealism, like product mockups or architectural visualizations.
MIDJOURNEY	Starts around $10/month for casual users.	Strong overall performance focus on artistic styles and details.	Wide range of artistic styles.	Well-suited for creative projects like concept art, illustrations, and exploring artistic styles.
DALL-E	Roughly starts at $15 for limited credits.	Good performance.	Tends to generate more stylized images.	Good for beginners due to its user-friendly interface and for projects where stylized visuals are desirable.

1. How to evaluate the results?

Inception Score (IS):

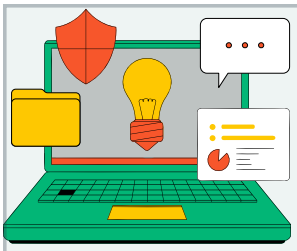
- **Focuses on: Diversity and realness of generated images.**
- **Calculation: IS considers two factors:**
  - **Average classifier confidence: Measures how confident a pre-trained image classifier is in assigning labels to generated images. High confidence suggests realistic features.**
  - **Class entropy: Evaluates the diversity of generated images across different classes.**
- **Higher is better**

Fréchet Inception Distance (FID):

- **Focuses on: Similarity of generated images to real images.**
- **Calculation: The distance between the distributions of features in both real and generated images in a high-dimensional space.**
- **Lower is better**

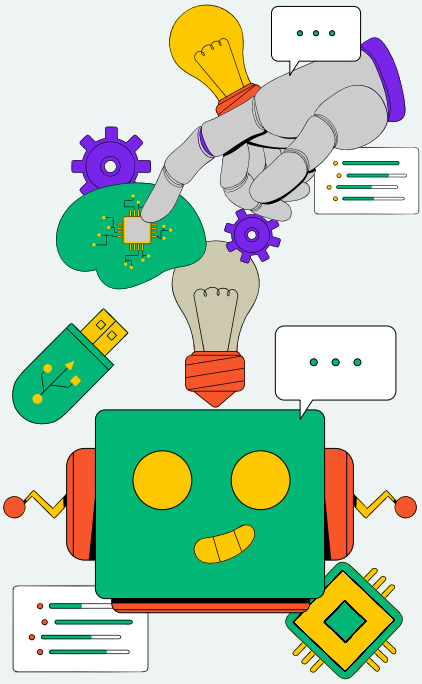
2.Good prompting tips?

- **Clarity is key: Be specific. Describe the subject, scene, and details.**
- **Style it up: Mention artistic styles (e.g., Van Gogh) or composition techniques (e.g., rule of thirds).**
- **Reference matters: Upload reference images or use keywords to guide the AI's style.**
- **Start simple, then refine: Gradually add details and use variations to get closer to your vision.**
- **Keep it concise & experiment: Avoid overly complex prompts.**



- Three tools all have different approach to the image generation process.
- They all have different image styles.
- Ultimately, the best choice depends on your needs.
- Keep in mind, this is a rapidly evolving field. As these tools continue to develop, the landscape might shift.

QUESTIONS AND ANSWERS



THANK YOU FOR LISTENING !