

# National and Kapodistrian University of Athens

---

MSc Thesis

**Evaluation of three text-to-image models**

Supervisor: **Theocharis Theocharis**



# MSc Thesis Structure

- **Chapter 1: Introduction**

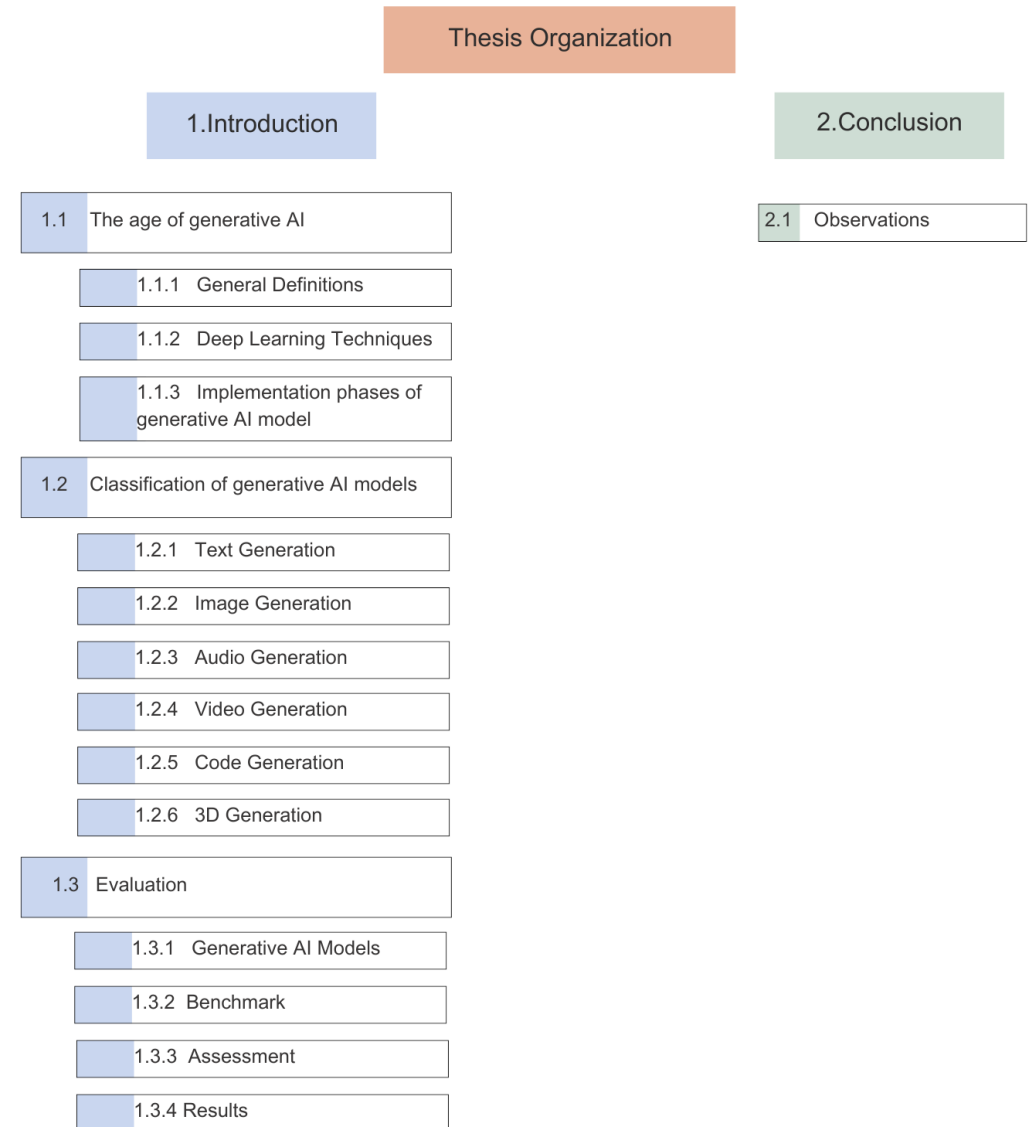
- We first provide all the necessary definitions commonly used around the AI, Generative AI, Generative AI model.
- Then we offer a classification of some of the advanced GAI algorithms. The categorization of the models has been done based on the generated content, output.

- **Chapter 2: Evaluation**

- We elucidate the three models under scrutiny, namely, the prevalent open-source frameworks (Stable Diffusion and Pix-Art- $\alpha$ ), alongside the commercial counterpart, DALL-E 2.
- We define the benchmark, which consists of six task types and 12 prompts for each task.
- Also, we explain the 2 types of assessment, quantitative and qualitative.

- **Chapter 3: Conclusion**

- This chapter displays the results of the two types of assessments
- Furthermore, we present some observations from the execution of the text-to-image models.



# General Definitions

- **Generative AI:** Generative AI refers to artificial intelligence that can generate novel content, rather than simply analyzing or acting on existing data like expert system.
- **ML:** A computer program is said to learn from experience E with respect to some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
- **Model:** A formal description of a system, process or equation used to simplify a complex subject.
- **Gen. Model:** Generative models directly predict a distribution and generate new data.
- **Foundation Model:** A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks. They are based on deep neural networks and self-supervised learning.
- **LLM:** A (large) language model (LLM) refers to neural networks for modeling and generating text data that typically combine three characteristics. First, the language model uses a large-scale, sequential neural network (e.g., transformer with an attention mechanism). Second, the neural network is pre-trained through self-supervision in which auxiliary tasks are designed to learn a representation of natural language without risk of overfitting (e.g., next-word prediction). Third, the pre-training makes use of large-scale datasets of text (e.g., Wikipedia, or even multi-language datasets).

GENERATION.

Gediral Sporrution

Dedatio!!

Gerdenct

Gosetigs

Gektartion

Gelectory

Portty

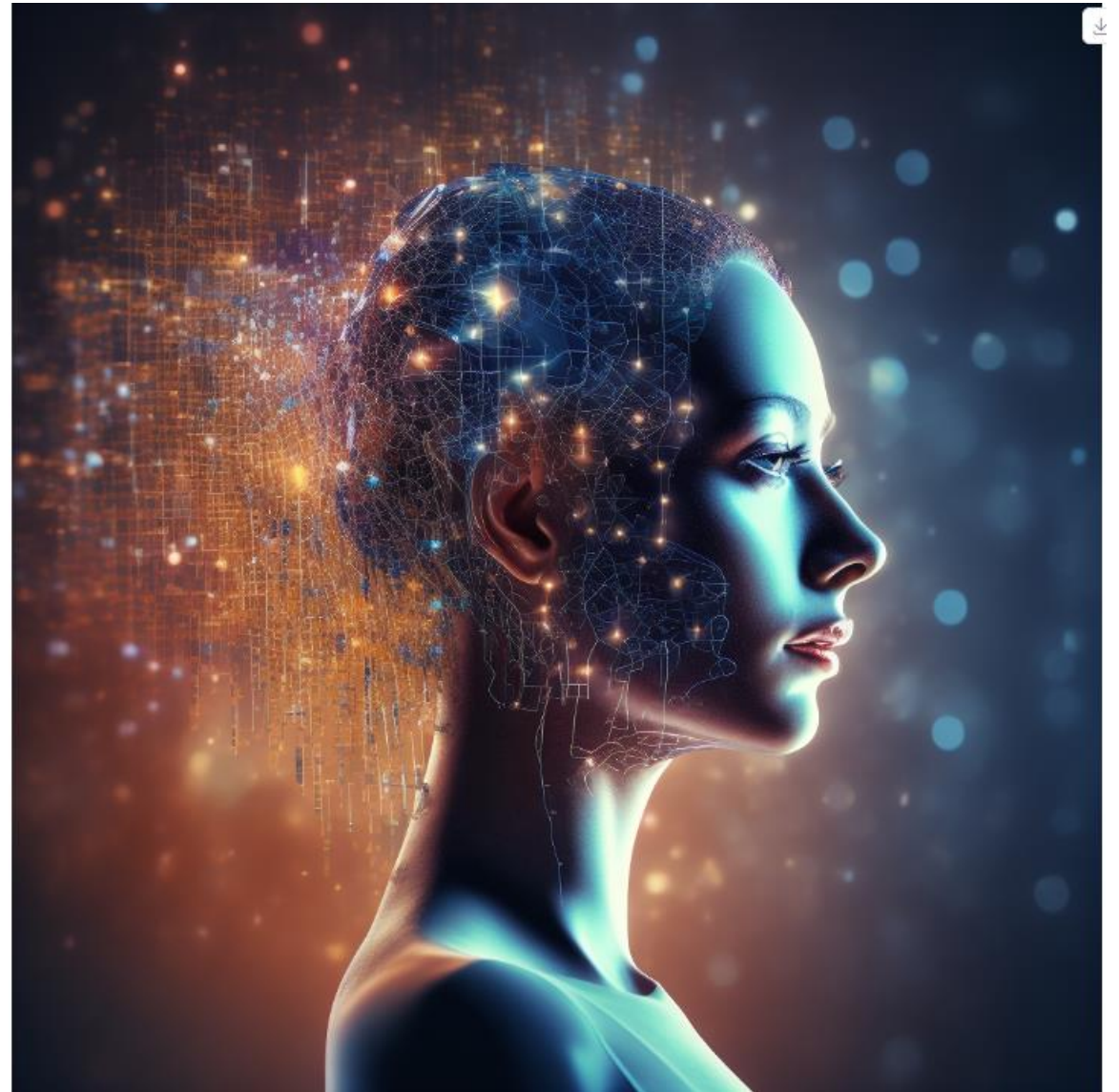
Denketioy

Apeials

DALL-E 2 prompt: An image of General Definitions, based on Simpson TV series style

# Deep Learning Techniques

---



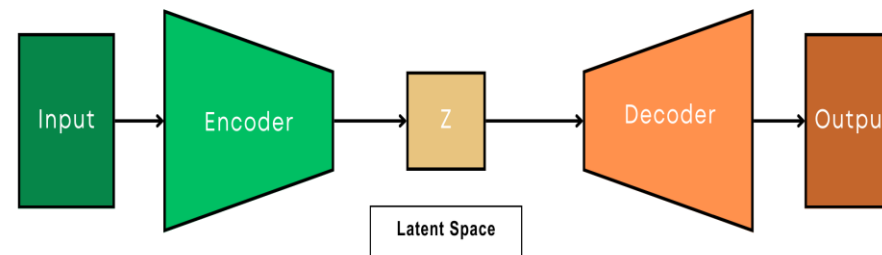
Pix-Art-α prompt: Give me an image about the Deep Learning Techniques

# VAEs

## Variational Auto-Encoders

---

- Introduced 2013
- Composed by E (encoder) & D (decoder)
- **Encoder:**
  - Maps the data point to a probability distribution that spans the latent space.
- **Latent Space:**
  - Low-dimensional representation that captures the core elements of the input data.
- **Decoder:** reconstructs input from the latent space in order to generate new data.

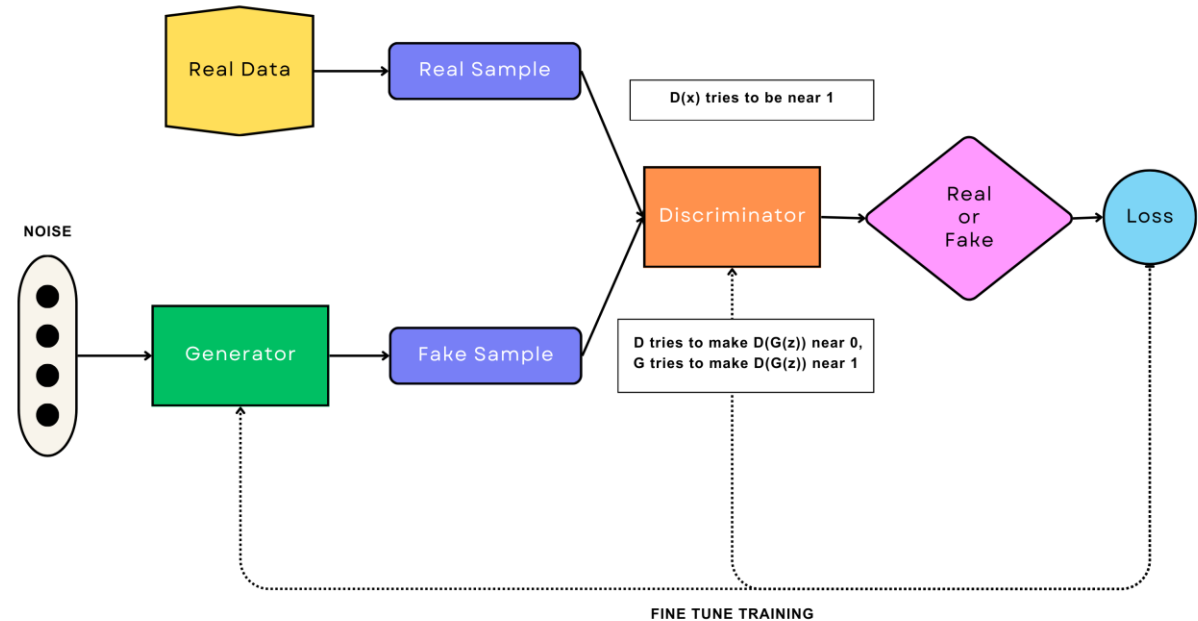


# GANs

## Generative Adversarial Networks

---

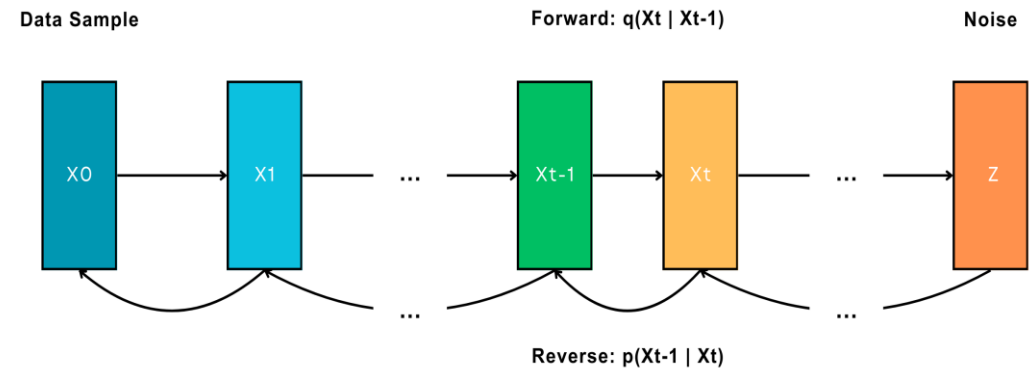
- Introduced 2014
- Composed by G (generator) & D (discriminator)
- It's a 2-player game, generator tries to trick the discriminator, while the discriminator goals is to identify whether a sample is from a true data distribution.
- **Generator:**
  - Creates output from sampled random noise.
- **Discriminator:**
  - Tries to understand if a sample is from a true distribution or from the generated distribution.



# Diffusion Models

---

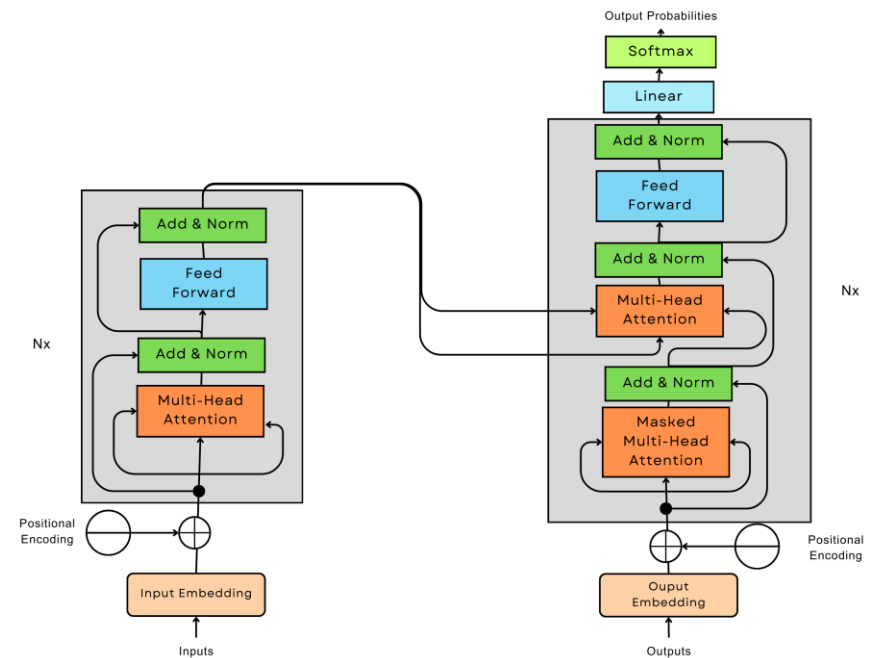
- Composed by 2 step process: Forward Diffusion Process & Reverse Diffusion Process.
- **Forward Diffusion Process**
  - Gradually add Gaussian noise and produce a sequence of noisy samples.
- **Reverse Diffusion Process**
  - Recreate true sample from a Gaussian noise input.



# Transformers

- Introduced in 2017
- Composed by E (encoder) & D (decoder)
- **Encoder:**
  - Consist of six layers.
  - Each layer apply linear transformations to all the words, but each layer apply different weights and bias.
- **Decoder**
  - Consist of six layers.
  - Each layer apply linear transformations.

Transformers called foundation models, because driving a paradigm shift in AI.



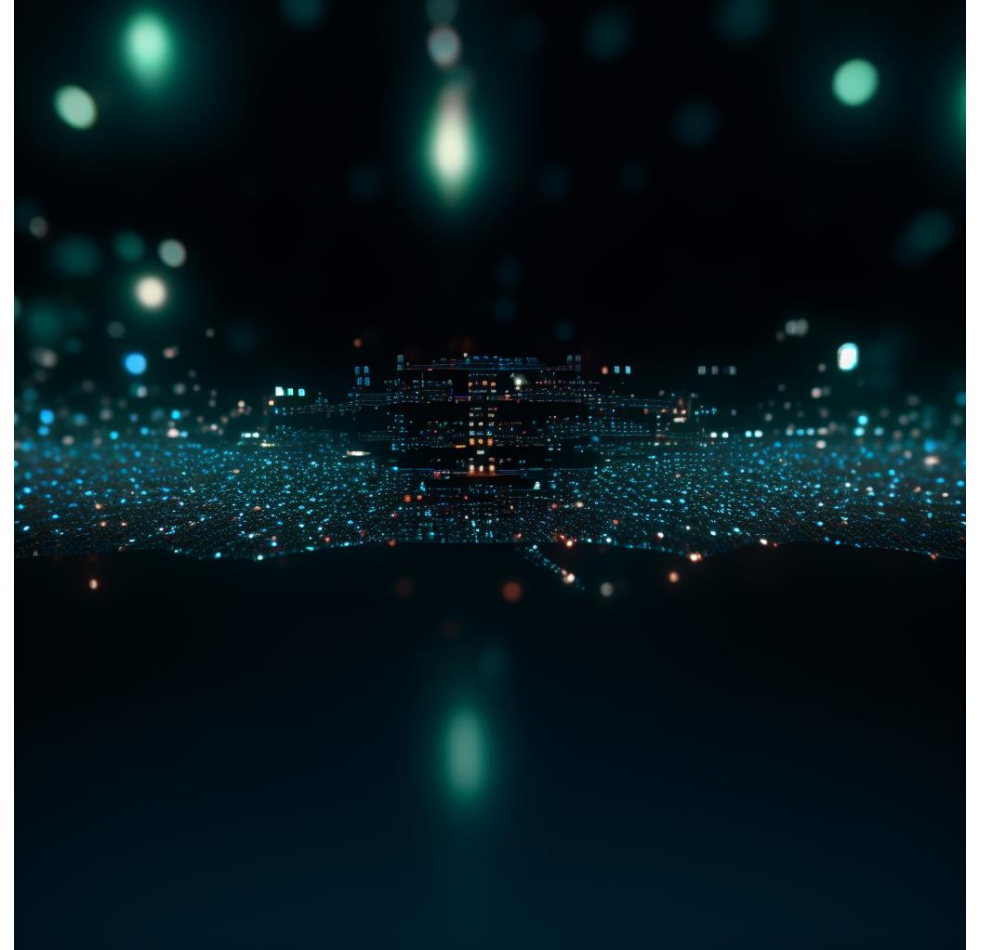
# Implementation phases of Generative AI

---



# Classification of Generative AI

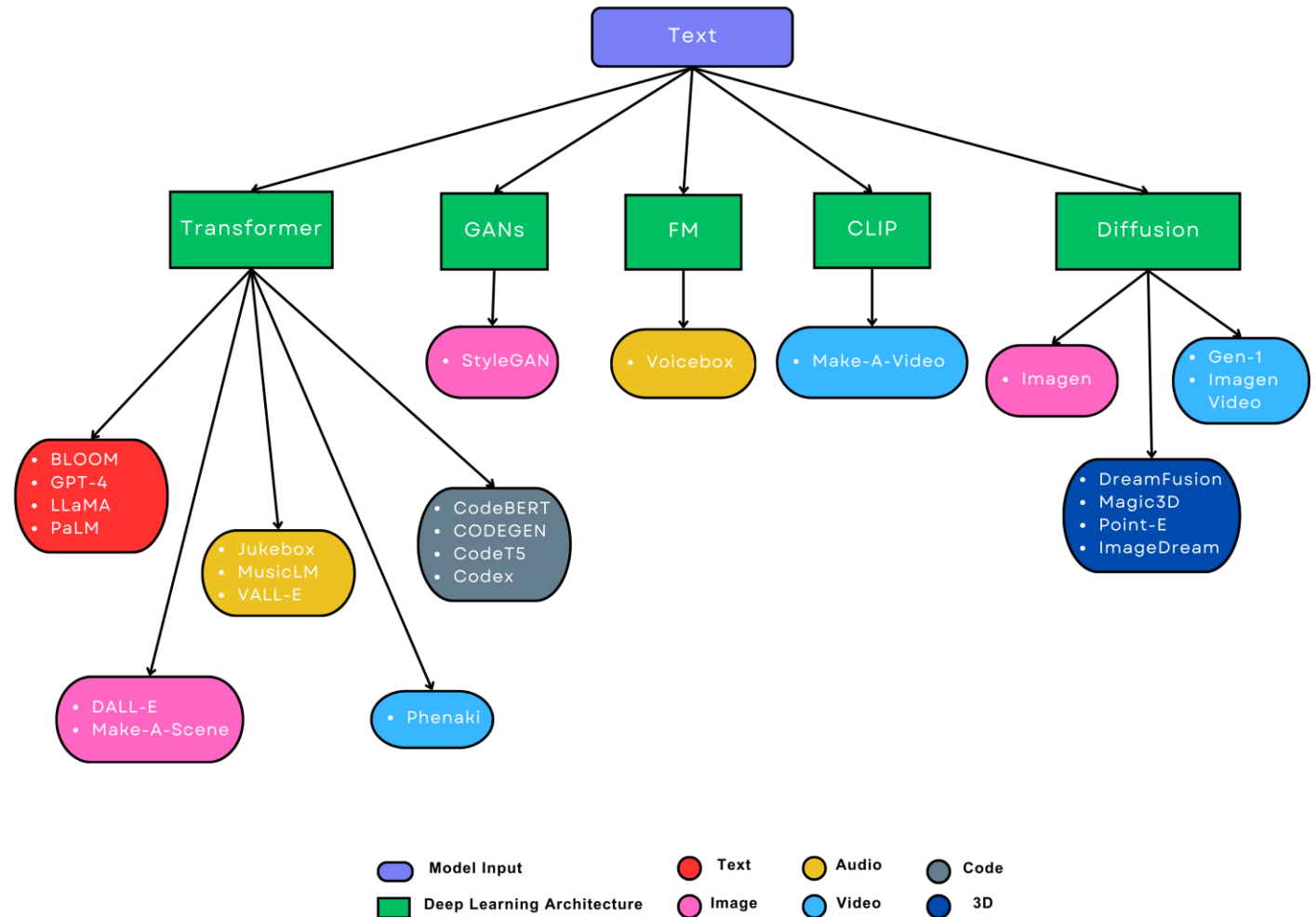
---



Pix-Art-a prompt: Give me an image of taxonomy of generative ai

# Taxonomy of Generative AI Models

---



# Text- to -~~XXX~~

## Text

Input	Output	Prescribed Task	Model
text	text	Generate text in multiple languages	BLOOM
		Multimodal model which can accept image and text inputs and produce text outputs	GPT-4
		Taking a sequence of words as an input and predicts a next word to recursively generate text	LLaMA
		It has strong capabilities in multilingual tasks and source code generation	PaLM

## Image

Input	Output	Prescribed Task	Model
text	image	Generate images from text descriptions	DALL-E
		Generate images based on textual instructions	Imagen
		Generate images based on text and optional a simple sketch input	Make-A-Scene
image	image	Generate highly realistic and diverse synthetic images	StyleGAN

# Text- to -~~XXX~~

## Audio

Input	Output	Prescribed Task	Model
text	audio	Generate music that can be conditioned on an artist, genres and lyrics	Jukebox
		Generate high-fidelity music pieces from text descriptions	MusicLM
		Generate highly realistic, human-like speech in a variety of languages and accent	VALL-E
		Produces high-quality audio clips	Voicebox

## Video

Input	Output	Prescribed Task	Model
text	video	Transforms existing videos into new ones using text prompts or reference images	Gen-1
		Generate text-guided videos	ImagenVideo
			Make-A-Video
text+image	video	Generating video from text prompt and input image	Phenaki

# Text- to -~~XXX~~

## Code

Input	Output	Prescribed Task	Model
text	code	Generate valid programming code using natural language query	CodeBERT
		Generate valid programming code using natural language description in a multi-step process.	CODEGEN
		Generate valid programming code using natural language descriptions	CodeT5
		Generate valid programming code using natural language prompts and is most capable in Python	Codex

## 3D

Input	Output	Prescribed Task	Model
text	3D	Generate 3D objects based on textual descriptions	Dreamfusion
		Generate 3D images using textual descriptions	Magic3D
		Generate 3D point clouds from text prompts	Point-E
image	3D	Generate high quality 3D model from any view-point given a single image	ImageDream

# Evaluation

---

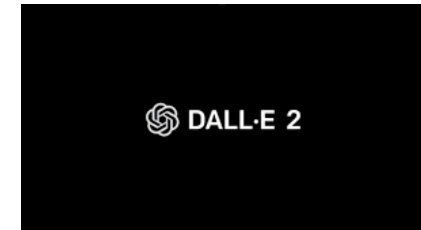


Pix-Art-a prompt: How do you imagine the Evaluation ?

# Models

---

- Three models were evaluated which share three fundamental hyperparameters: **prompt, height, and width.**
  - Stable Diffusion:
    - By CompVis, Stability AI
    - It was run locally through the gui AutomaticGUI111. (see [here](#))
    - Checkpoint: Stable-Diffusion-v1-5 was used
  - DALL-E 2:
    - By OpenAI
    - Python code was written and OpenAI API was used. (see [here](#))
    - Program asks for: prompt, size of image and number of images
  - Pix-Art- $\alpha$ :
    - By Huawei Noah's Ark Lab
    - It was used the web application (see [here](#))
    - Checkpoint: PixArt-XL-2-1024-MS



# Benchmark

---

- A multi-task Benchmark for evaluating text-to-image models was created based on [Drawbench](#)
- Six Task types (see [here](#))
  - Coloring
  - Counting
  - Conflicting
  - Text
  - Positional
  - Faces
- Each task type contains 12 prompts.
- In total we have 72 prompts.

Index	Prompt
1	A red colored car
2	A black colored car
3	A pink colored car
4	A navy blue colored car
5	A red car and a white sheep
6	A blue bird and a brown bear
7	A green apple and a black backpack
8	A green cup and a blue cell phone
9	A red pencil in a green cup on a blue table
10	An office with five desks and seven colorful chairs
11	An orange bird scaring a blue scarecrow with a red pirate hat
12	Two pink football balls and three green basketball balls on a bench

# Approach

---

- Two types of assessment for the evaluation of the models
  - **Qualitative**
    - Based on Human Evaluation
    - Raters asked 1 question and select: 1 (worst) to 5 (best)
      - The image represents the text caption: [TEXT CAPTION] ?
      - Question subjectively evaluates image-text alignment.
    - 2 questionnaires, one for [512](#) and the other for [1024](#)
  - **Quantitative**
    - Based on CLIP model
    - Python code was written (see [here](#))
    - CLIP is a family of models
    - We use the model with the best performance ViT-L/14@336px
    - CLIP score is the cosine similarity between the given prompt and the produced image

# Conclusion

---

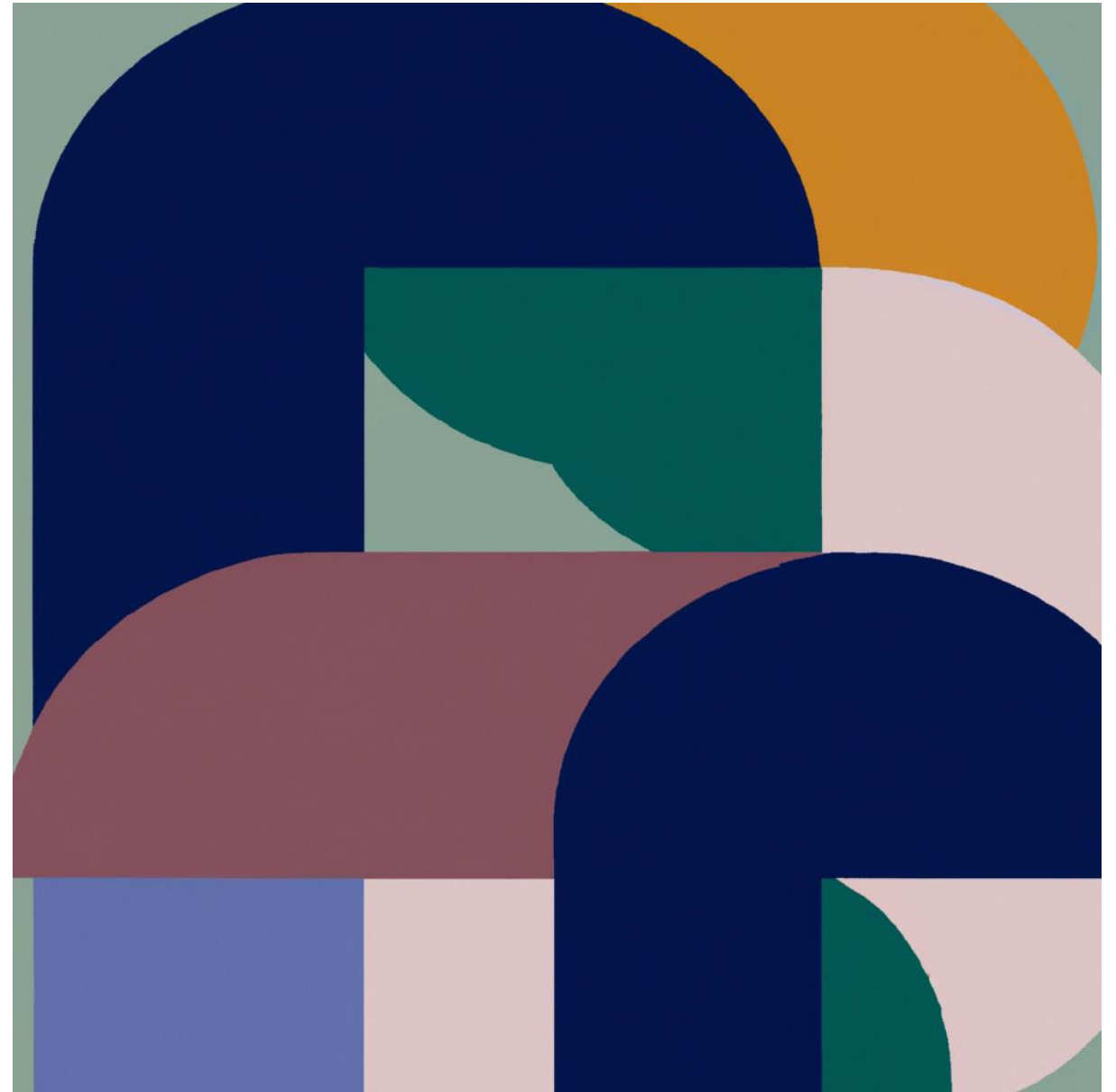


DALL-E 2 prompt: An image which describe the “Conclusion”

# Results Qualitative

---

We surveyed 17 people (employees from the Data and Analytics department of an IT company) who evaluated the performance of three image generation models: Stable Diffusion, DALL-E 2, and Pix-Art- $\alpha$ . Each person evaluated 432 images, providing ratings between 1 (worst) and 5 (best). We collected a total of 7344 scores from 2 questionnaires, 512 and 1024.



DALL-E 2 prompt: An image which describe the Qualitative Results, on Matisse style

# Results Qualitative 512


Normalized ratings (%) for human evaluation by Task and Model			
Task Type	DALL-E 2	Pix-Art- $\alpha$	Stable Diffusion
Colouring	74,51	67,28	60,17
Conflicting	49,75	61,76	34,44
Counting	74,02	51,47	41,91
Faces	74,02	69,12	52,08
Positional	41,30	48,53	40,44
Text	38,60	13,24	13,73

The normalized scores for images of size (width, height) = (512, 512)

If we examine the chart depicting the score values for each model and for each task across the models, specifically for images sized 512x512, we can make several observations:

- DALL-E 2: Raters identified four task categories where this model performed well.
  - DALL-E 2 achieved the highest rater scores for image-text alignment in coloring, counting, faces, and text tasks. This suggests that images generated by DALL-E 2 best match the descriptions provided in the prompts for these aspects (coloring, counting, faces, and text).
- Pix-Art- $\alpha$ : Raters identified two task categories where this model performed well.
  - Pix-Art- $\alpha$  achieved the highest rater scores for image-text alignment in conflicting and positional tasks. This suggests that images generated by Pix-Art- $\alpha$  best match the descriptions provided in the prompts for these aspects (conflicting, positional).
- Stable Diffusion: Raters did not identify any single task category where this model performed well.

# Results Qualitative 1024



Normalized ratings (%) for human evaluation by Task and Model			
Task Type	DALL-E 2	Pix-Art- $\alpha$	Stable Diffusion
Colouring	72,06	64,22	31,74
Conflicting	50,00	58,46	27,33
Counting	71,08	56,37	31,37
Faces	75,49	72,06	27,08
Positional	46,94	42,28	25,37
Text	49,02	15,44	13,11

The normalized scores for images of size (width, height) = (1024, 1024)

If we examine the chart depicting the score values for each model and for each task across the models, specifically for images sized 1024x1024, we can make several observations:

- DALL-E 2: Raters identified five task categories where this model performed well.
  - DALL-E 2 achieved the highest rater scores for image-text alignment in coloring, counting, faces, positional and text tasks. This suggests that images generated by DALL-E 2 best match the descriptions provided in the prompts for these aspects (coloring, counting, faces, positional and text).
- Pix-Art- $\alpha$ : Raters identified one task category where this model performed well.
  - Pix-Art- $\alpha$  achieved the highest rater scores for image-text alignment in conflicting task. This suggests that images generated by Pix-Art- $\alpha$  best match the descriptions provided in the prompts for these aspects (conflicting).
- Stable Diffusion: Raters did not identify any single task category where this model performed well.
  - The model's ratings are not only low, but also show a significant gap compared to the scores of the previous model.

# Results

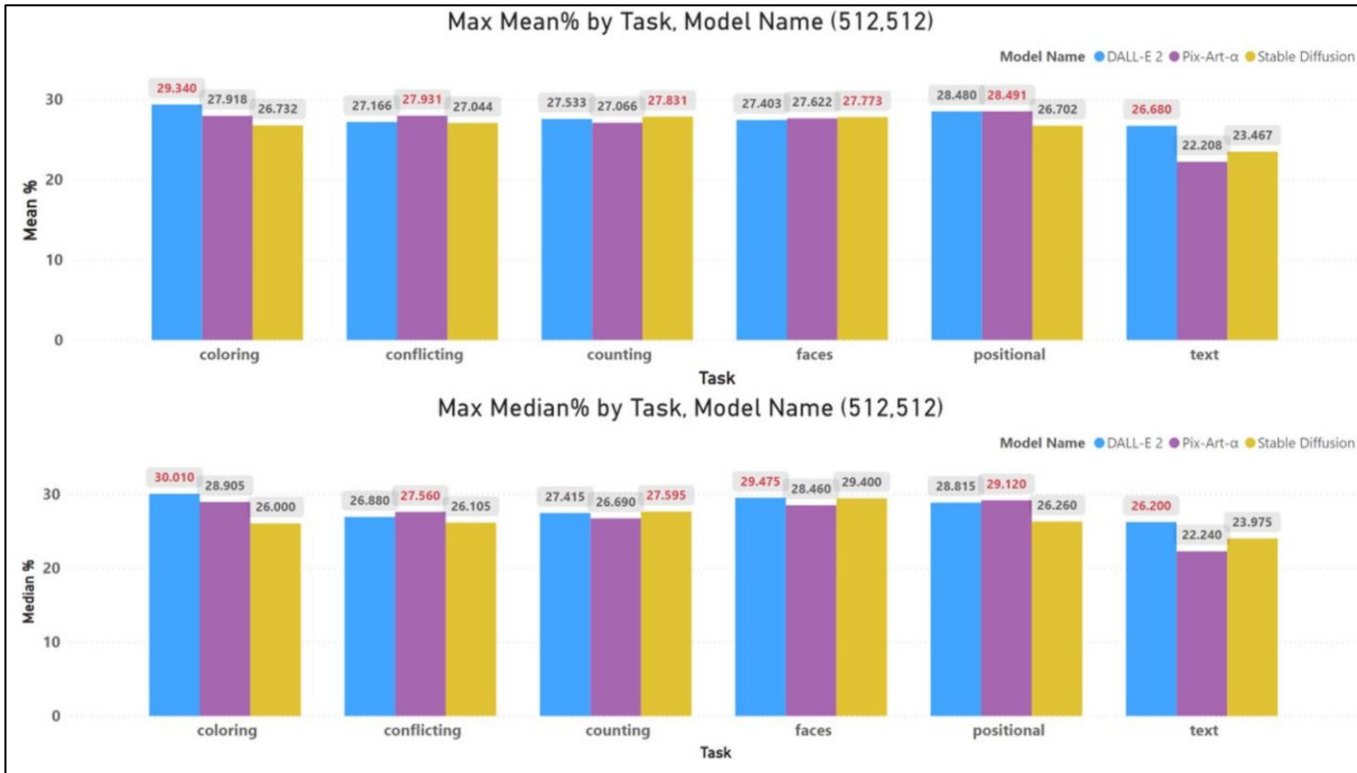
## Quantitative

- Detailed clip scores for all the pairs (prompt, image) is [here](#).
- Mean and Median of the scores per category is [here](#).
- We ran the ViT-L/14@336px model, which demonstrates superior performance compared to all other models within the CLIP family. This model utilizes the Vision Transformer as its image encoder



DALL-E 2 prompt: An image which describe the Quantitative Results

# Results Quantitative 512



If we examine the chart depicting the mean and median values for each task across the models, specifically for images sized 512x512, we can make several observations. The optimal alignment of image-text for each model based on CLIP score is:

## 1. Stable Diffusion: counting

- Stable Diffusion effectively translates the count descriptions into visually accurate images.

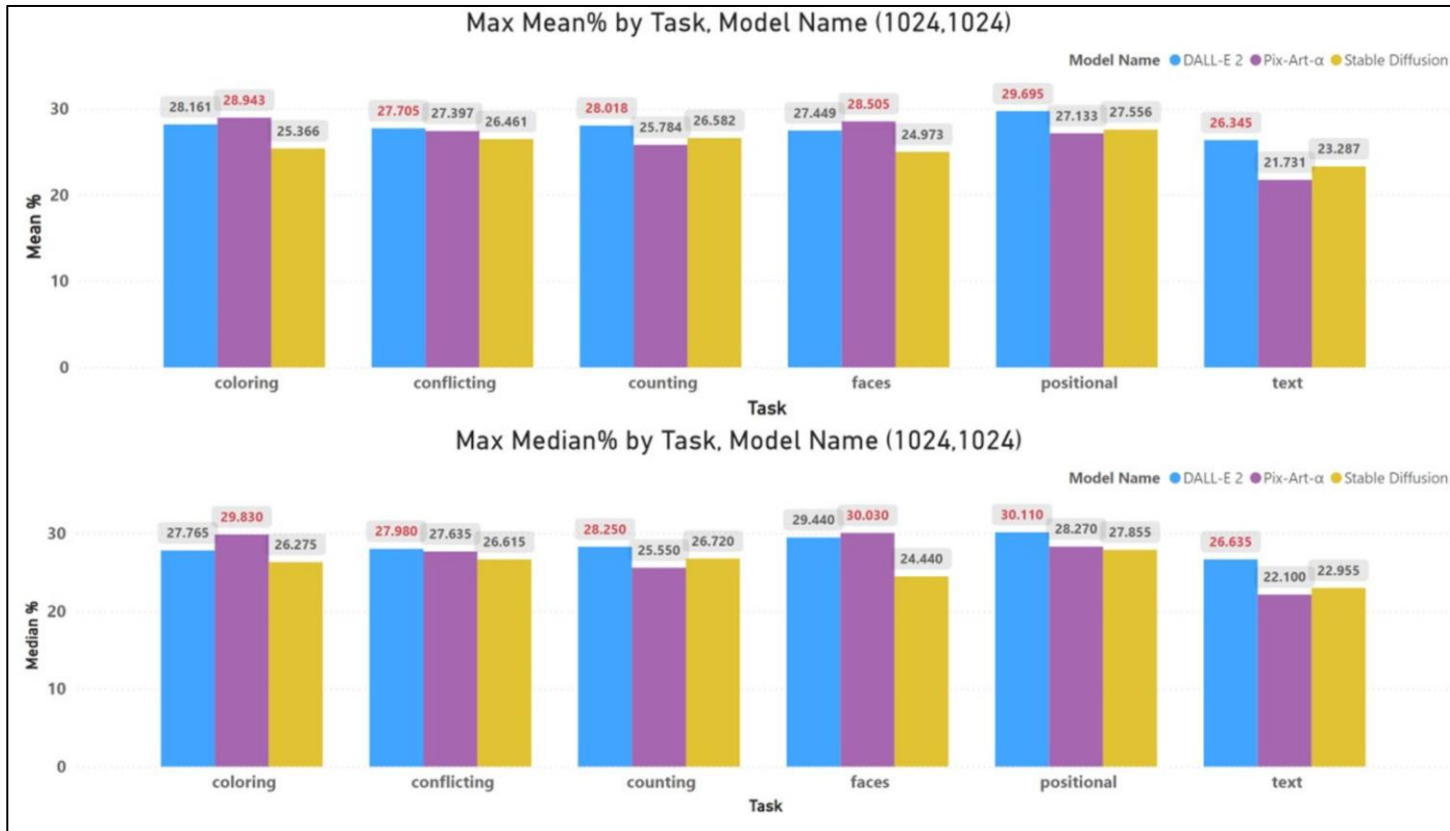
## 2. DALL-E 2: coloring, text, faces

- DALL-E 2 has the best image-text alignment CLIP score on coloring and text tasks, suggesting that its generated images closely align with the given prompts in terms of color and demonstrating its proficiency in translating textual prompts into visually coherent images.
- On the faces task, DALL-E 2's higher median suggests it performs more consistently well across the dataset, while Stable Diffusion's higher mean indicates that it has some very high-performing instances that raise its average score. This indicates that DALL-E 2 is better in generating images with facial features.

## 3. Pix-Art-α: conflicting, positional

- Pix-Art-α is particularly proficient in generating images based on positional prompts and conflicting scenarios

# Results Quantitative 1024



If we examine the chart depicting the mean and median values for each task across the models, specifically for images sized 1024x1024, we can make several observations. The optimal alignment of image-text for each model based on CLIP score is:

## 1. **Stable Diffusion: N/A**

- The stable diffusion does not achieve a maximum value on any task. This is to be expected, since the images are not suitable for any category. (see findings)

## 2. **DALL-E 2: conflicting, counting, positional, text**

- DALL-E 2 achieved the highest scores for conflicting scenarios, count descriptions, positional prompts and textual prompts, indicating its proficiency in translating the prompts into visually coherent images.

## 3. **Pix-Art-α: coloring, faces**

- Pix-Art-α has the best image-text alignment clip score on coloring and faces task, suggests that its generated images closely align with the given prompts in terms of color and in terms of facial characteristics.

# Observations

---

- There are some observations after the execution of the generative AI models 'Stable Diffusion, DALL-E 2, Pix-Art- $\alpha$ .
- These observations show the limitations of the 3 generative AI models.



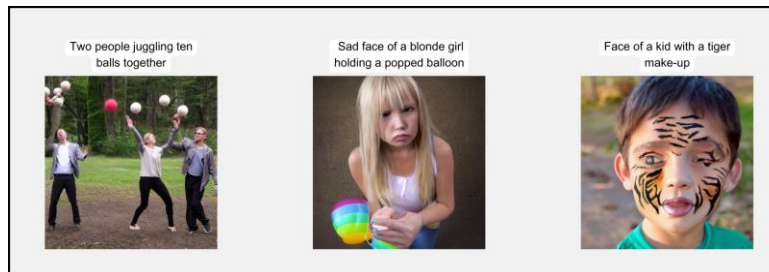
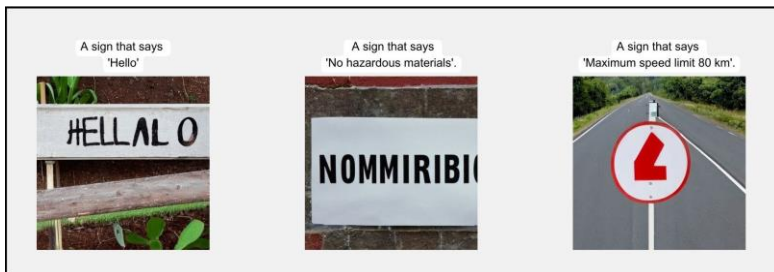
DALL-E 2 prompt: Give me an image in Pablo Picasso style for describing the term "Findings"

# Observations

## Stable Diffusion 512

---

- Legible Text cannot be rendered by the model.
- People characteristics are not generated properly.
- Difficult tasks involving compositionality are not generated properly.

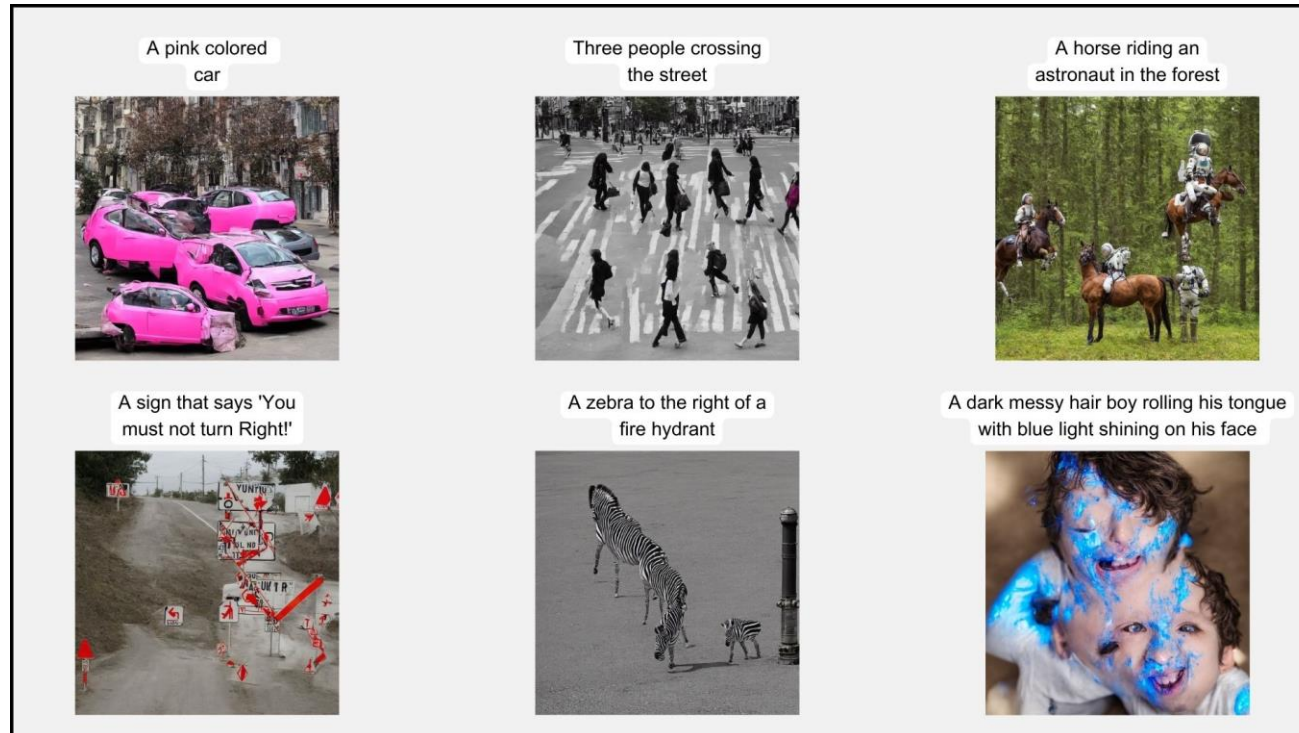


# Observations

## Stable Diffusion 1024

---

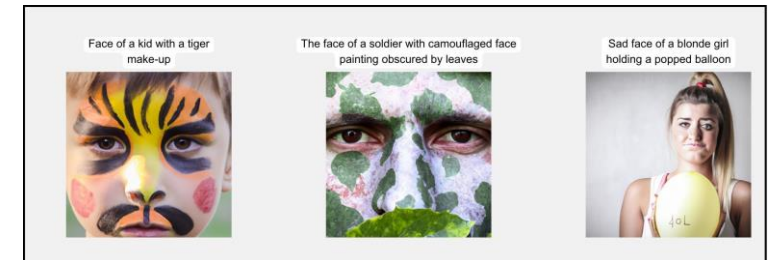
- The model's performance falls short across all six task types of the benchmark.



# Observations

## DALL-E 2 512

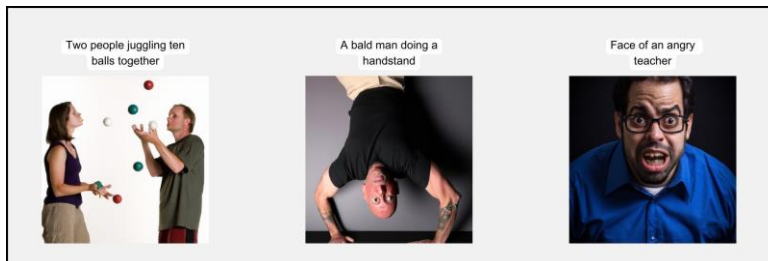
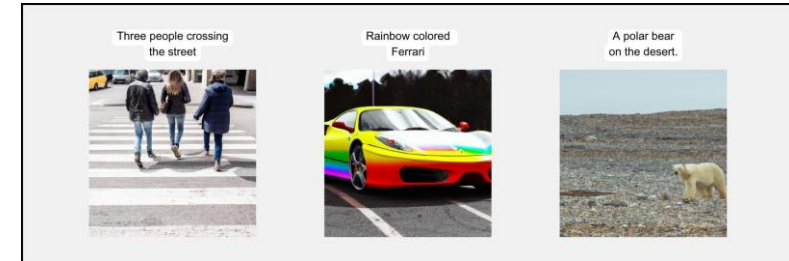
- DALL-E has difficulties on spelling.
- DALL-E can do scenes with generic backgrounds (a city, a landscape, etc) but even then, if that's not the main focus of the image then the fine details tend to get pretty scrambled.
- People characteristics are not generated properly when they are not defined properly on the prompt. It forgets how faces and fingers work.



# Observations

## DALL-E 2 1024

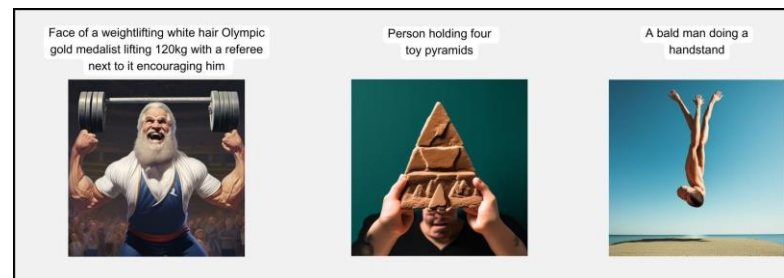
- DALL-E has difficulties on spelling.
- DALL-E is capable of creating scenes with generic backgrounds, like a city or landscape. However, when these backgrounds are not the primary focus, the fine details often become quite jumbled.
- People characteristics are not generated properly. It forgets how faces and fingers work.



# Observations

## Pix-Art-α 512

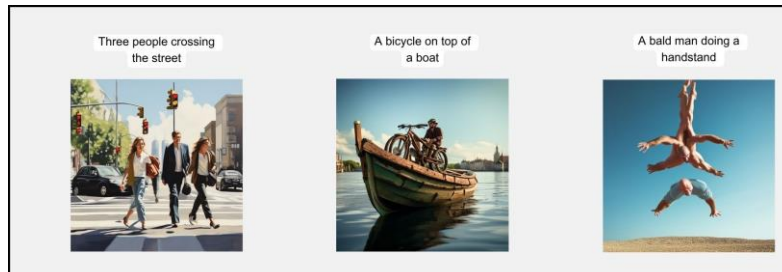
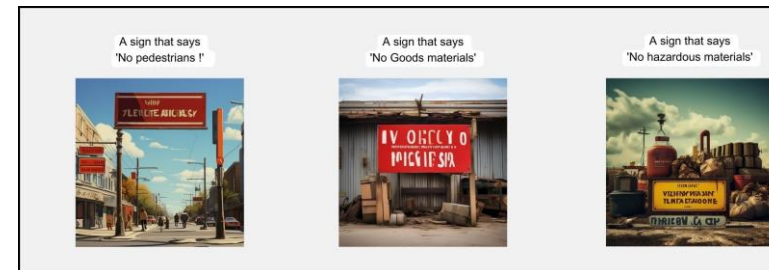
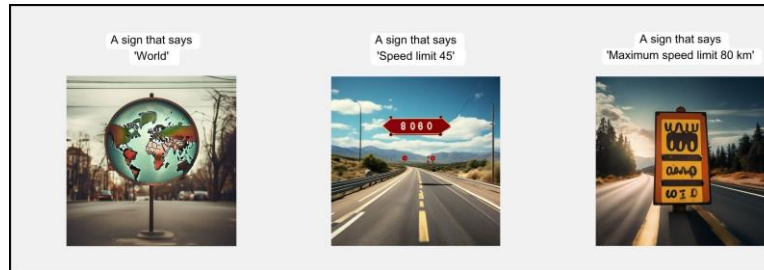
- Pix-Art-a cannot render legible text.
- The model struggles with counting tasks. It has weaknesses in accurately controlling the number of objects in an image.
- People characteristics are not generated properly. It forgets how faces, hands and fingers work.



# Observations

## Pix-Art- $\alpha$ 1024

- Pix art-a cannot render legible text.
- People characteristics are not generated properly.
- The model cannot count the objects in an image and struggles to accurately manage their number.

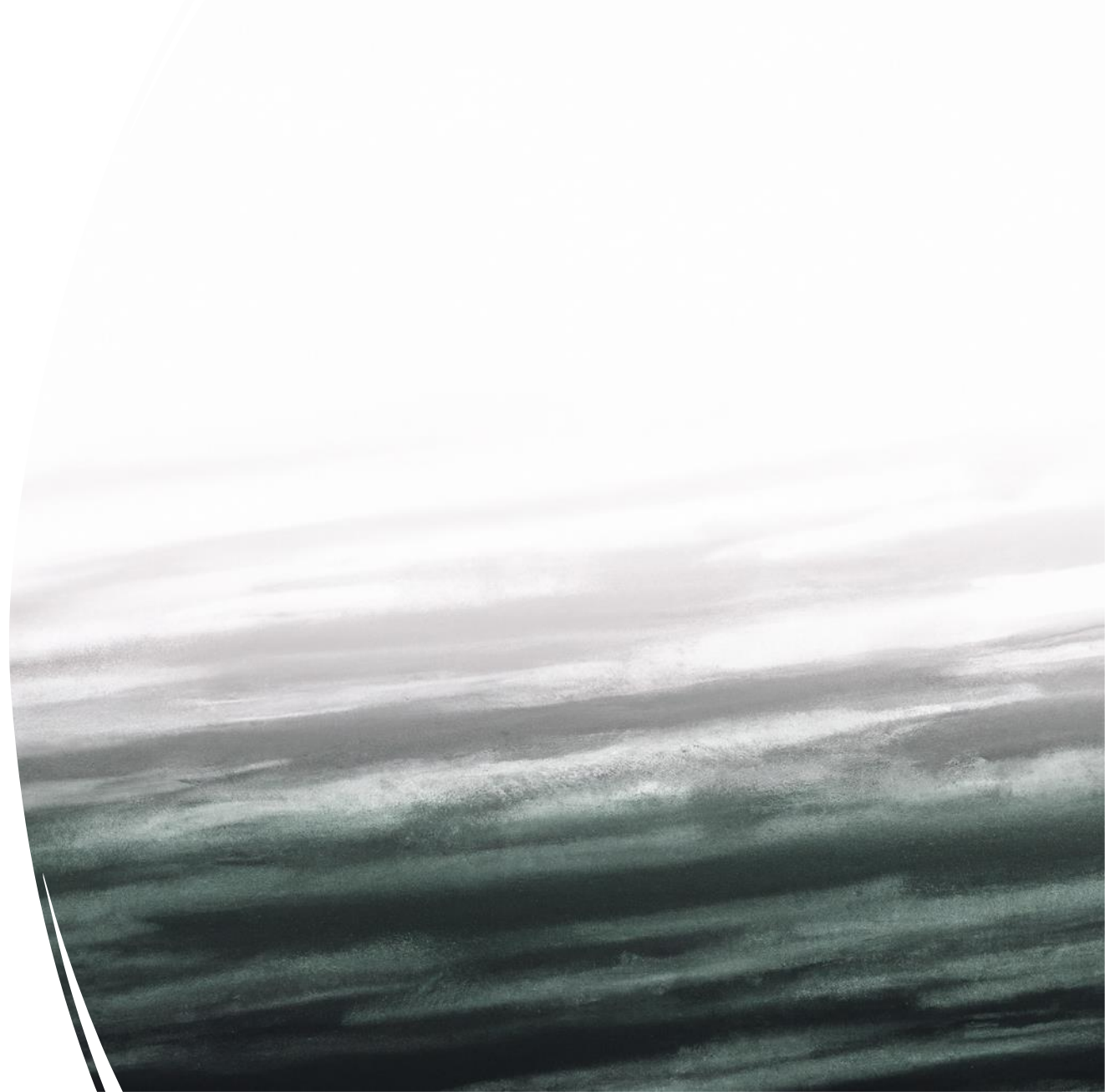


# THE END

MSc Thesis

**Evaluation of three text-to-image models**

Supervisor: **Theocharis Theocharis**



DALL-E 2 prompt: Give me an image in Rothko painter style which describe the "End"