National and Kapodistrian University of Athens

MSc Thesis

Generative AI in Graphics

Supervisor: **Theocharis Theocharis**

MSc Thesis Structure
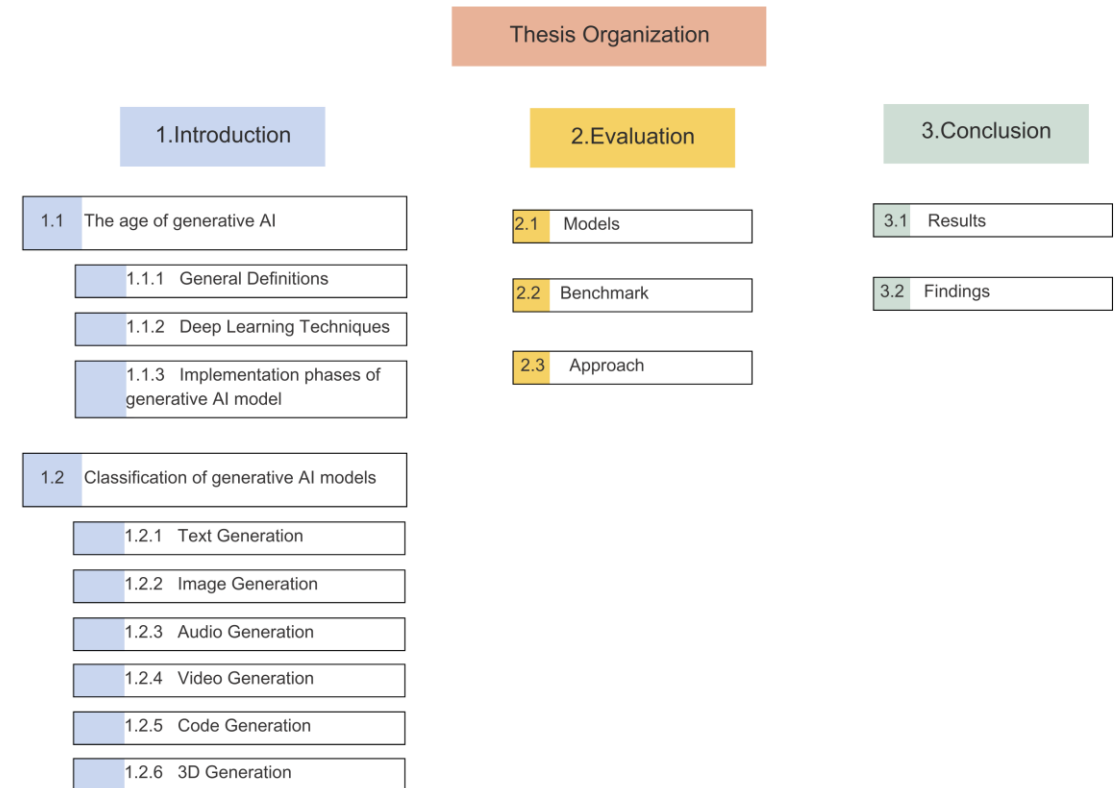
- **Chapter 1: Introduction**
    - We first provide all the necessary definitions commonly used around the AI, Generative AI, Generative AI model.
    - Then we offer a classification of some of the advanced GAI algorithms. The categorization of the models has been done based on the generated content, output.

- **Chapter 2: Evaluation**
    - We elucidate the three models under scrutiny, namely, the prevalent open-source frameworks (Stable Diffusion and Pix-Art-α), 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.

Thesis Organization

1.Introduction

1.1	The age of generative AI

	1.1.1	General Definitions
	1.1.2	Deep Learning Techniques
	1.1.3	Implementation phases of generative AI model

1.2	Classification of generative AI models

	1.2.1	Text Generation
	1.2.2	Image Generation
	1.2.3	Audio Generation
	1.2.4	Video Generation
	1.2.5	Code Generation
	1.2.6	3D Generation

2.Evaluation

2.1	Models
2.2	Benchmark
2.3	Approach

3.Conclusion

| 3.1 | Results |
| 3.2 | Findings |

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).

GENRATION.

Gediral   Sporrution
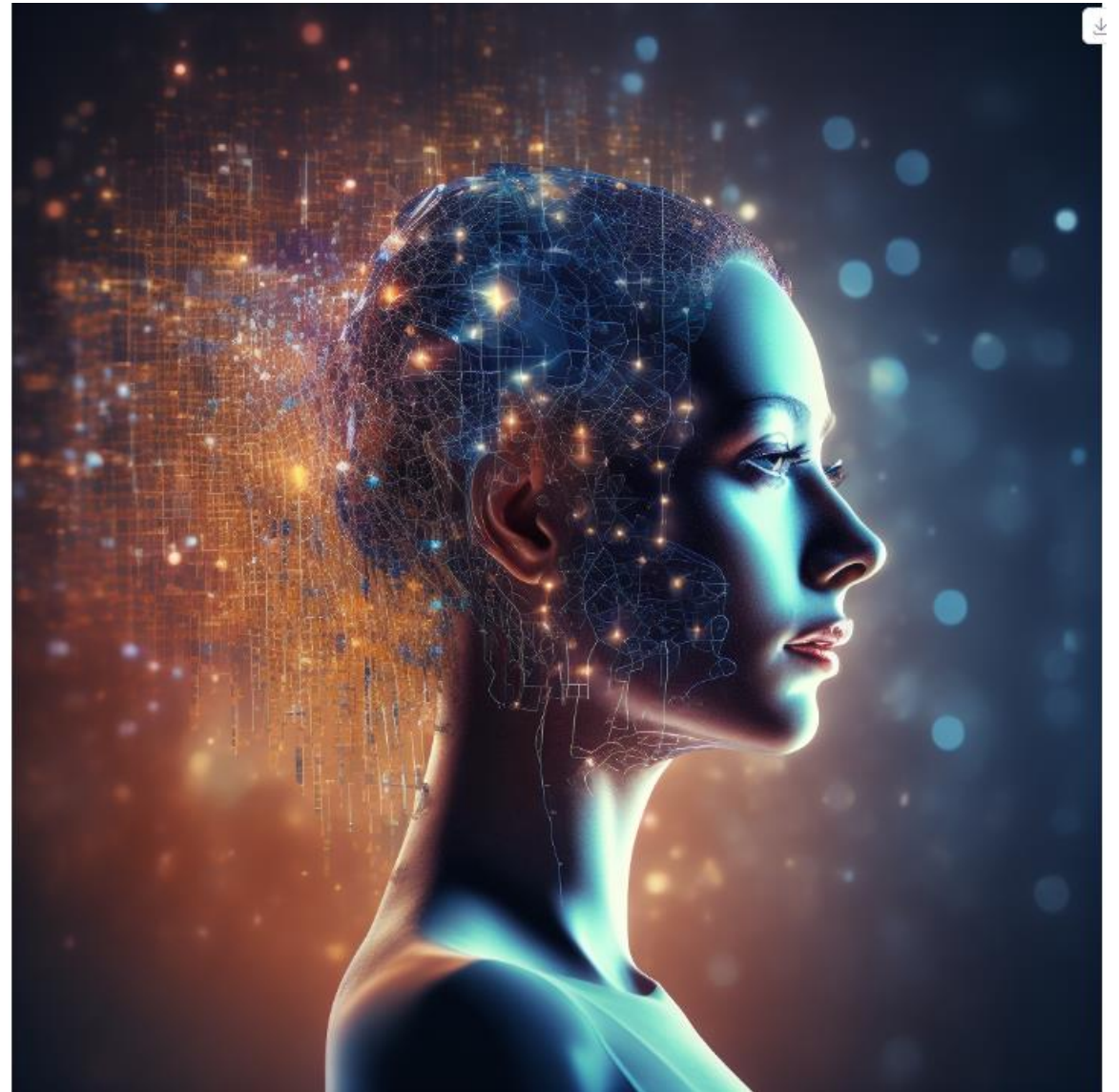
Dedatio!!   Gerdehct

Gosutigs   Getertion

Gelectory   Portty

Derretiov   Apeials

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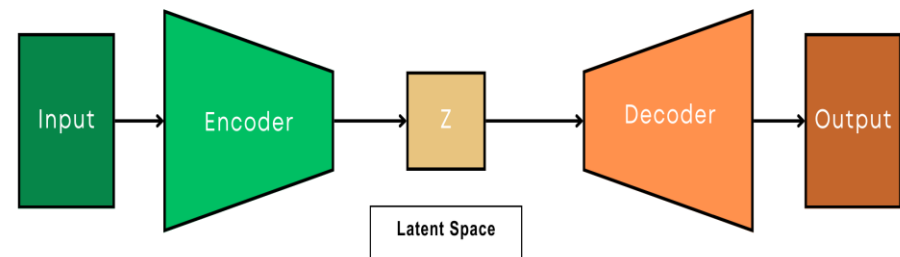
Deep Learning Techniques

Pix-Art-α prompt: Give me an image about the Deep Learning Techniques
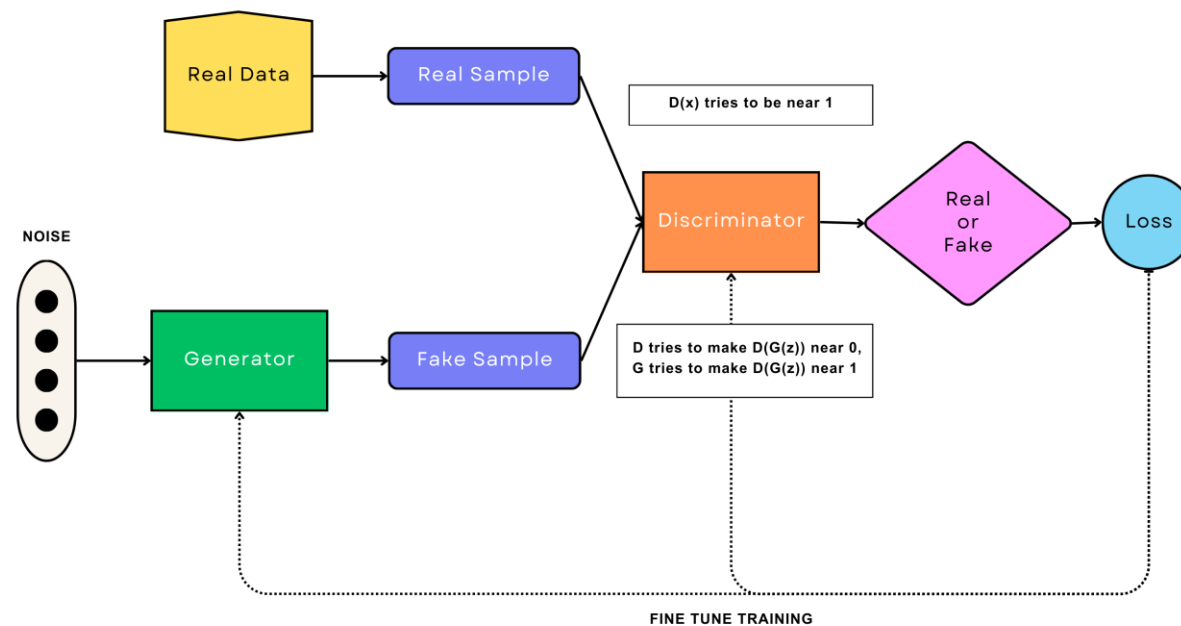
VAEs
Variational Auto-Encoders
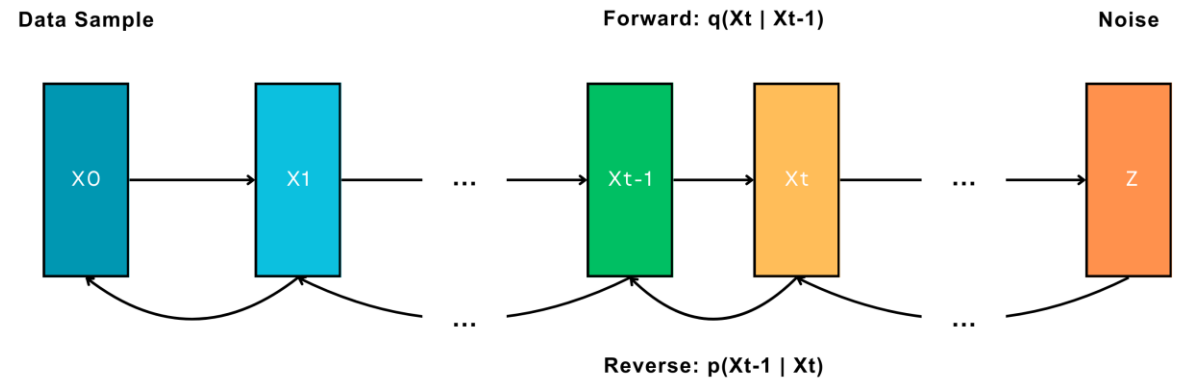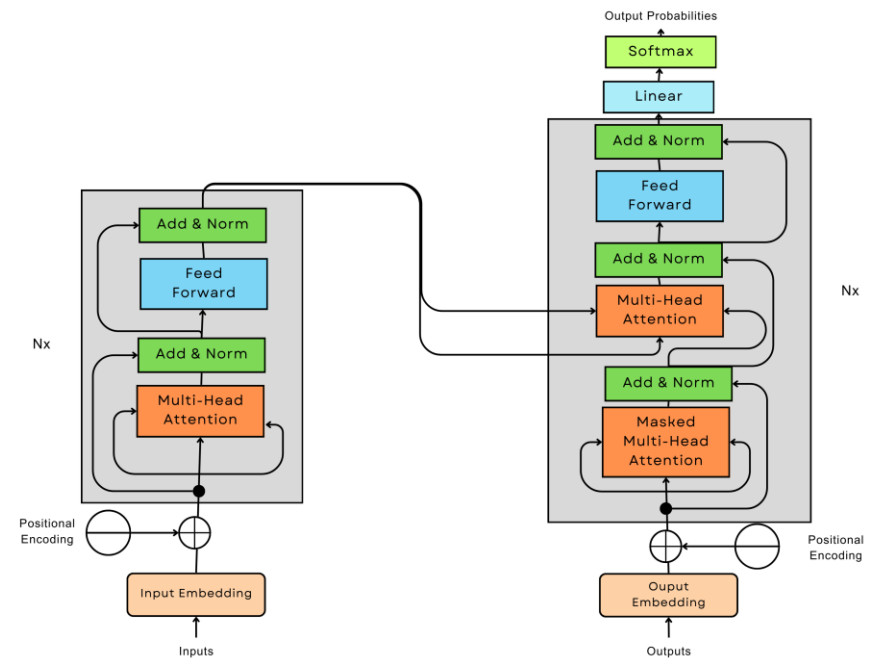
- **TODO**

GANs
Generative Adversarial Networks

- **TODO**

Diffusion Models

- **TODO**

Data Sample   Forward: q(Xt | Xt-1)   Noise

X0 → X1 → ... → Xt-1 → Xt → ... → Z

Reverse: p(Xt-1 | Xt)

Transformers

- **TODO**

Implementation phases of Generative AI

1 PROBLEM DEFINITION

Clarify what the problem is and the desired outcomes.

2 DATA COLLECTION & PROCESSING

Gather a large and representative dataset and preparing it for the training.

3 MODEL SELECTION

Pick out the appropriate deep learning architecture.

4 MODEL TRAINING

Train the chosen deep learning architecture using the collected dataset.

5 MODEL EVALUATION

Esteem the implementation based on evaluation metrics that are tailored to the specific task.

6 MODEL FINE-TUNING

Improve the performance of the implementation.

7 DEPLOYMENT

Deploy the implementation for real use in order to generate new samples.

8 MONITORING & MAINTENANCE

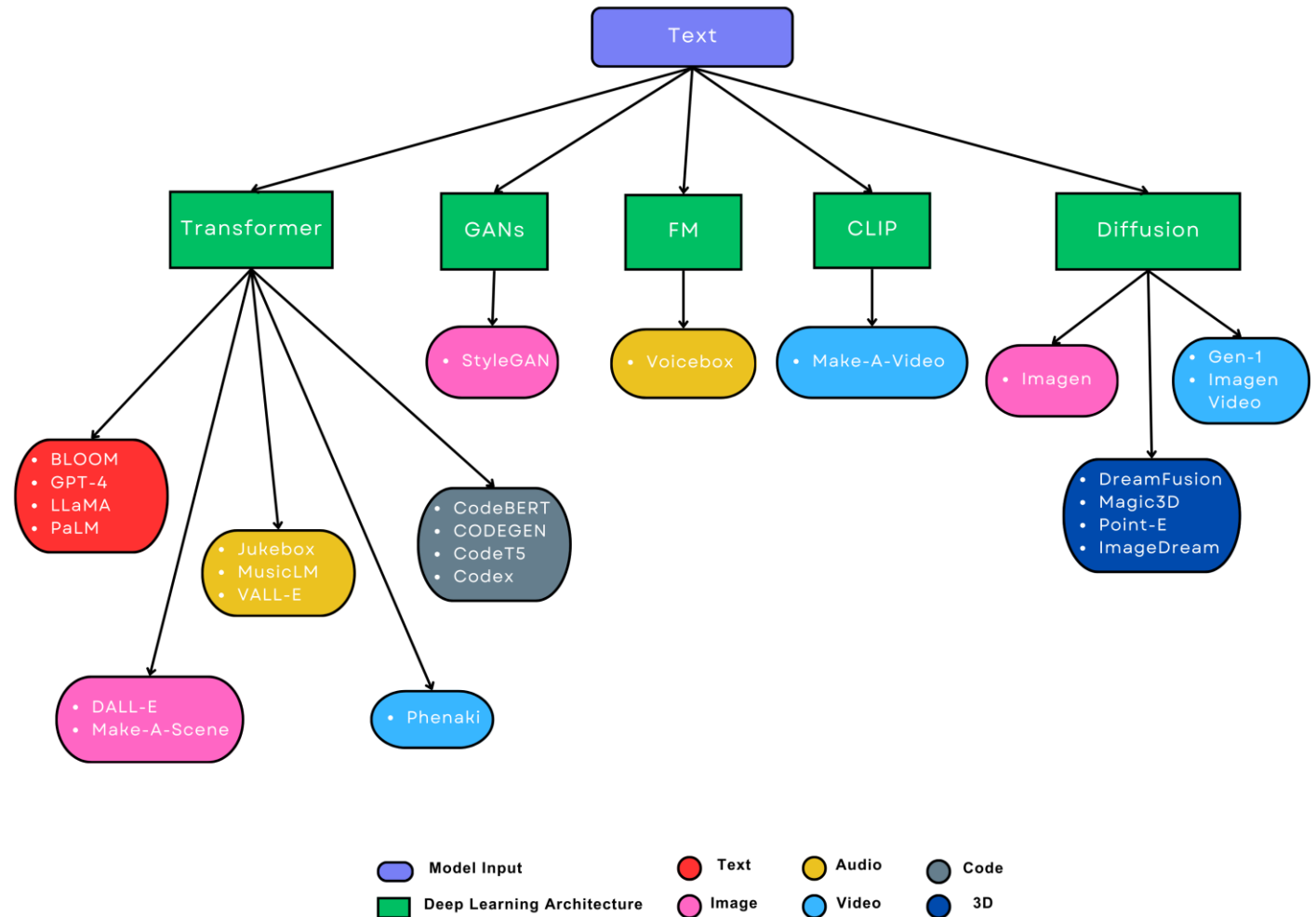Keep close track of how it's working and monitor the performance of the process.

Classification of Generative AI

Pix-Art-α 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 viewpoint given a single image	ImageDream

Evaluation

Pix-Art-α 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: promp, size of image and number of images
  - Pix-Art-α:
    - By Huawei Noah's Ark Lab
    - It was used the web application (see here)
    - Checkpoint: PixArt-XL-2-1024-MS
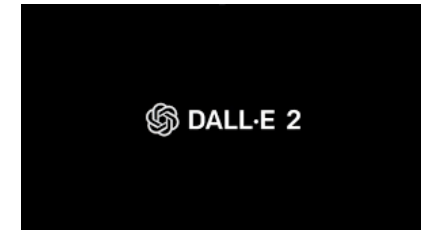
Benchmark

- 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
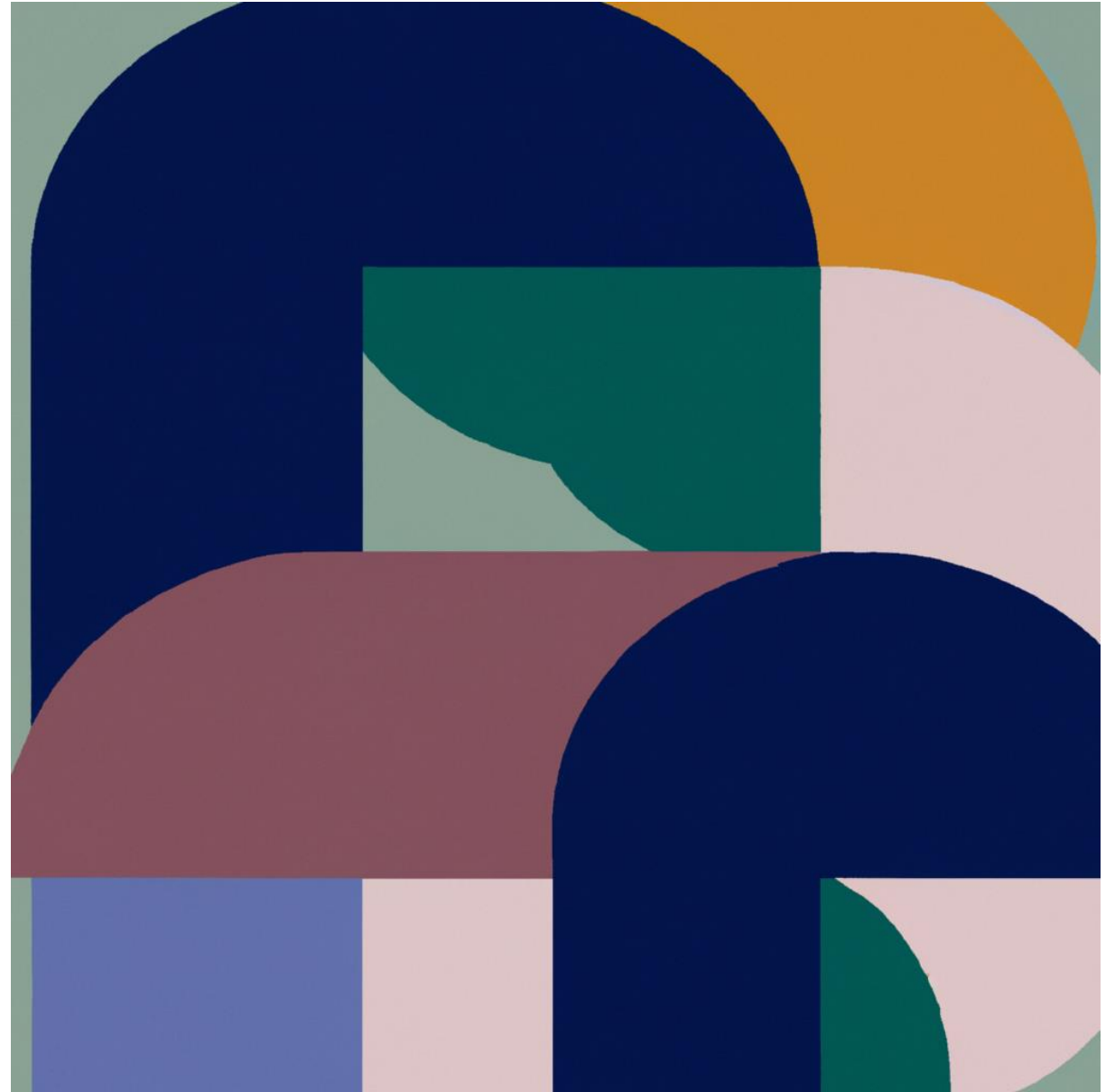   - CLIP score is the cosine similarity between the given prompt and the produced image

Conclusion

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

Results
Qualitative

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

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
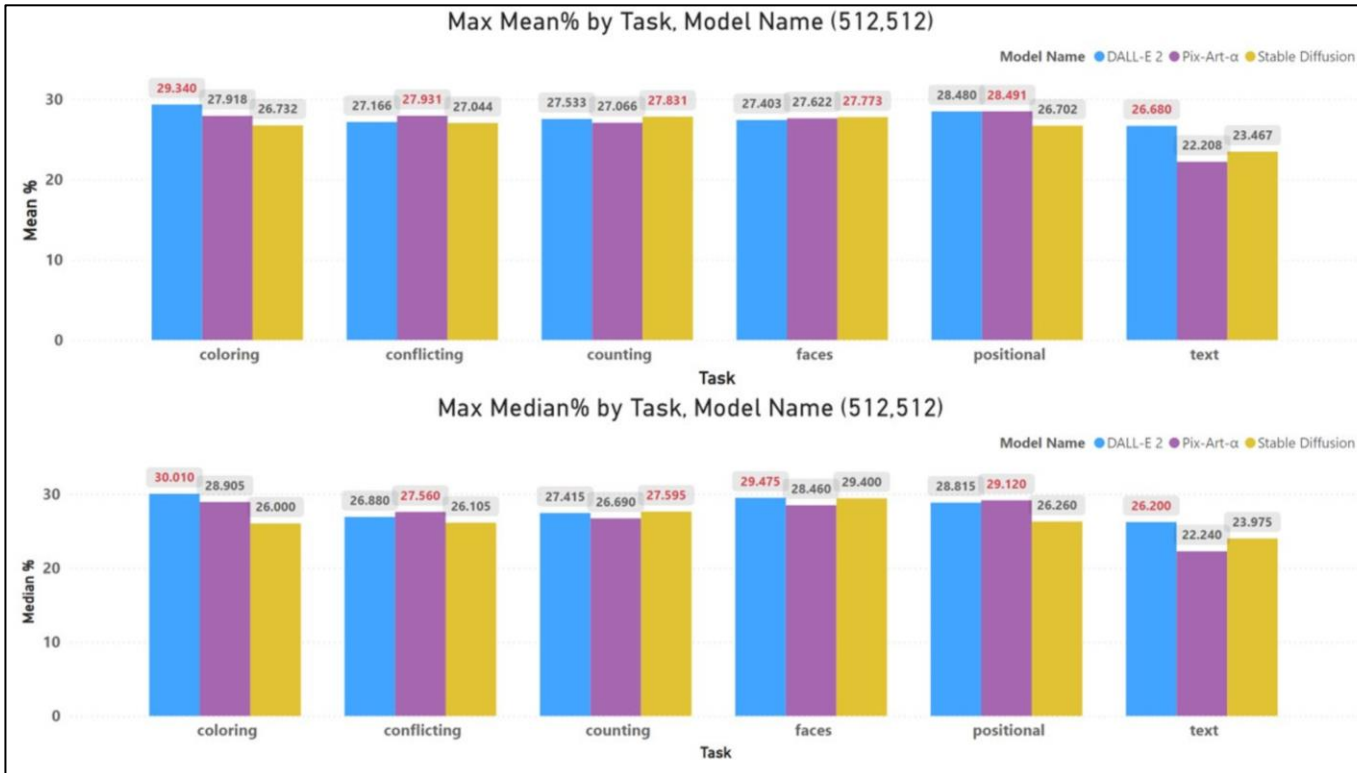
Max Mean% by Task, Model Name (512,512)

Max Median% by Task, Model Name (512,512)

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
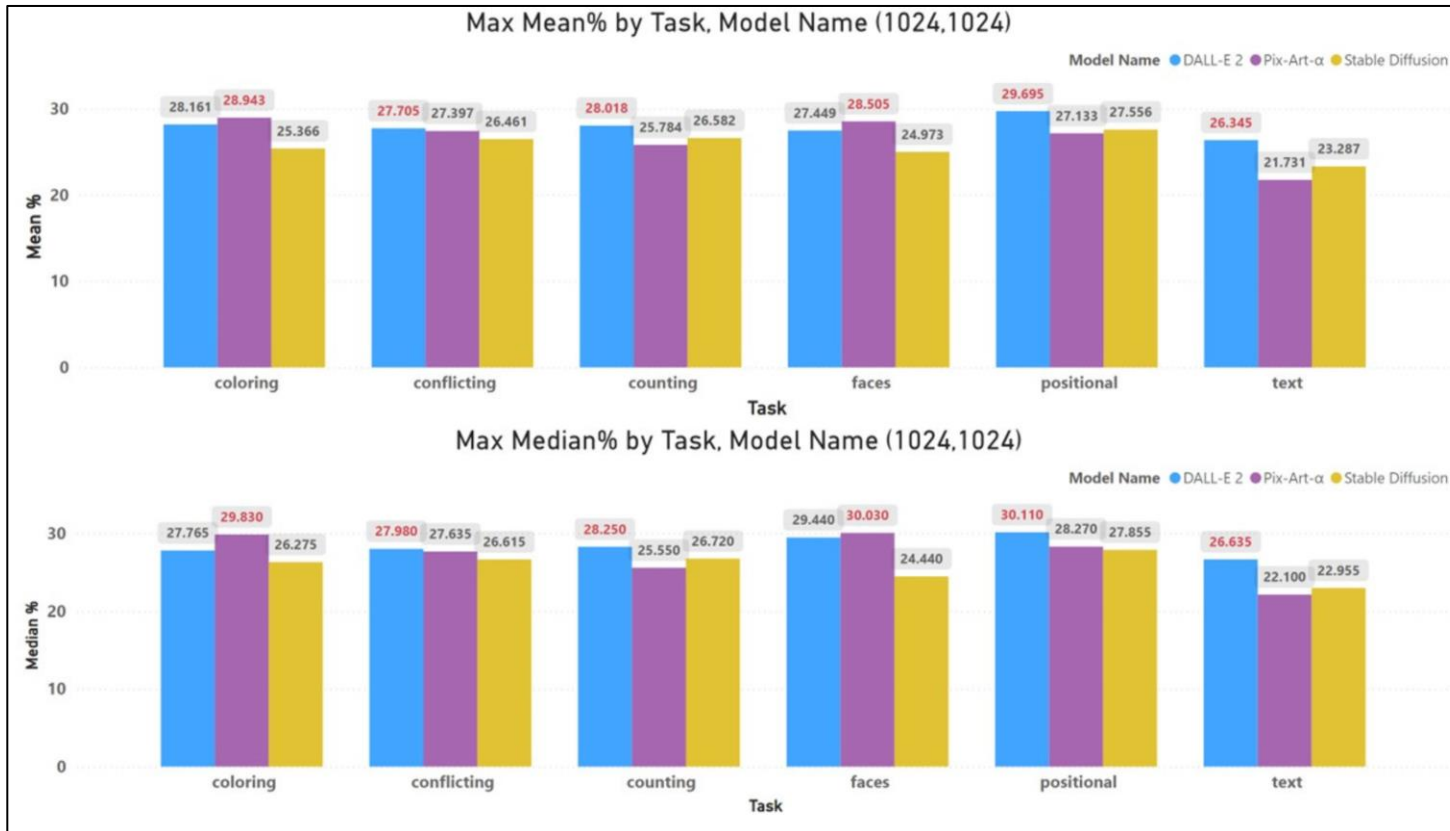
Max Mean% by Task, Model Name (1024,1024)

Max Median% by Task, Model Name (1024,1024)

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.
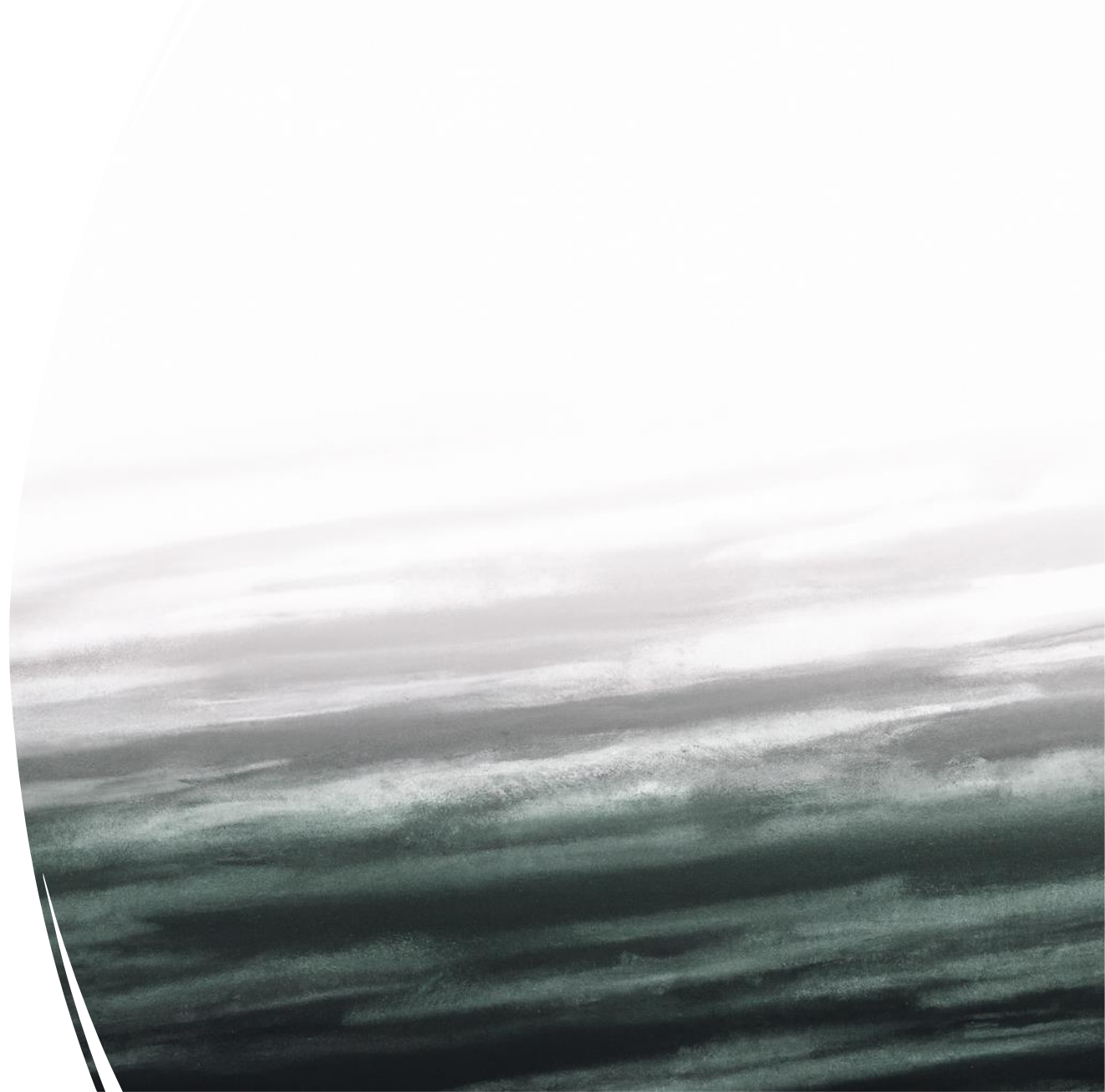
Findings

- **TODO**

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

THE END

MSc Thesis

Generative AI in Graphics

Supervisor: **Theocharis Theocharis**

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