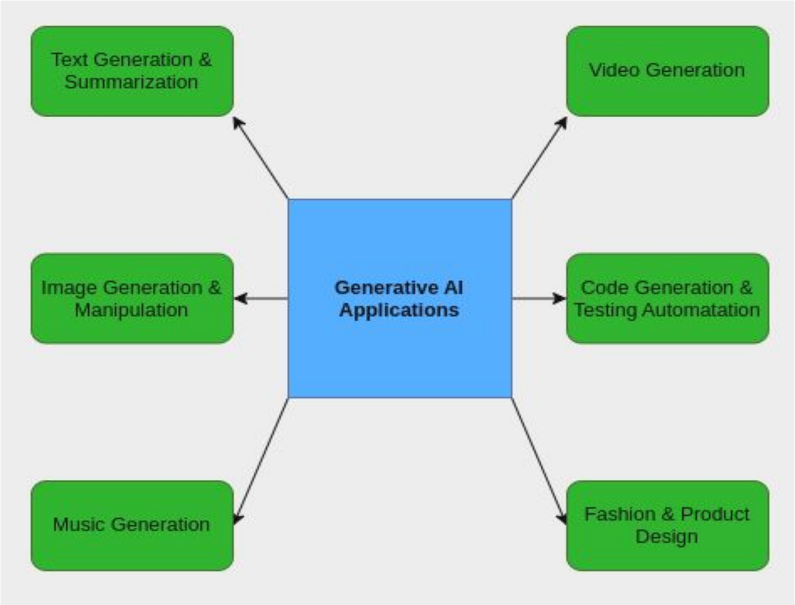
Insights from the Dataset

Approaching Interpretability in Generative AI Models by Estimating Training Data Influence

Theo **AI**valis
29/03/2024

Generative AI models

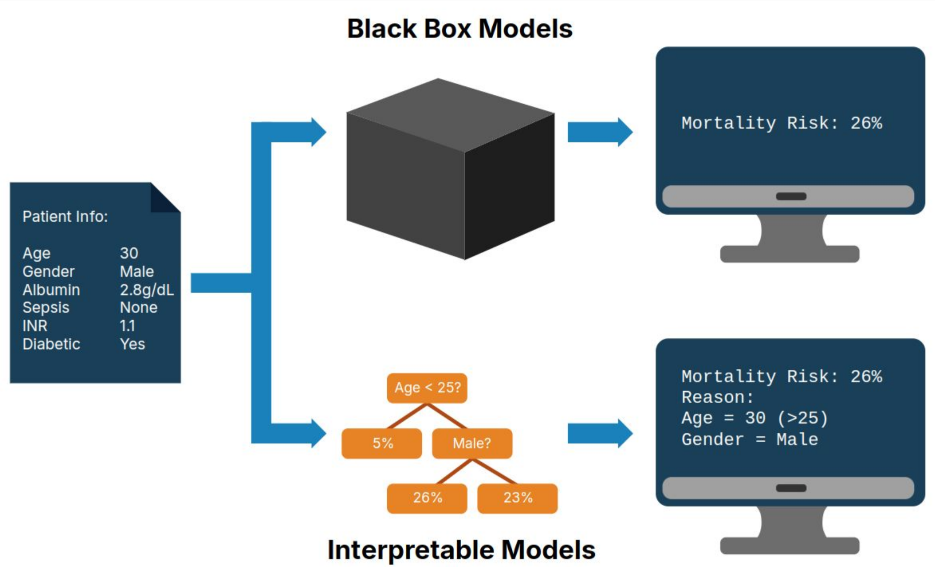
The key characteristic of generative AI is its ability to create something that does not exist in the training data explicitly. It captures the underlying complexity and diversity of the input and produces unique outputs that exhibit creativity and originality.


Text Generation & Summarization
Video Generation
Image Generation & Manipulation
Generative AI Applications
Code Generation & Testing Automatation
Music Generation
Fashion & Product Design

What is Interpretability?

Deep Learning methods has strong predictive ability but lack interpretability, while Traditional Machine Learning is not so powerful, but it is often easier to interpret.

Models are interpretable when humans can readily understand the reasoning behind predictions and decisions made by the model.



How others approach the problem?

- Local Explanation Methods: Employ LIME and SHAP method to explain model decisions at the local feature level.
- Integration of Natural Language Understanding and Visual Perception: Combine image classification with text explanations for enhanced interpretability.
- Importance of Individual Training Examples in VAEs: Understand how each example shapes the latent space and influences outputs for Variational Autoencoders (VAEs).
- Attention Mechanism and vectors of Transformers helps in the Interpretability of Deep Learning.
- Ablation Studies and Sensitivity Analysis: Systematically test model performance on different data parts to identify significant features impacting behavior.

Motivation

- Concerns Regarding Rights of Content Creators
  - AI Act (The proposed regulations on AI in the European Union)
    - AI technologies are developed and used in an ethical and responsible manner.
    - Manage the risks associated with AI technologies,like potential biases, discrimination, privacy violations, and safety concerns.
    - Include obligations for data quality, privacy, security, and access rights.
  - EGAIR (A group of artists, creatives, publishers and associations from all over Europe. They try to bring to the public attention how their data are being exploited without our consent.)
- Emergence of Interpretable AI
  - Complex inner mechanisms of these models
  - Vast amount of Training Data
- Human-AI Collaboration
  - Error Detection and Debugging
  - Crucial in certain domains like Medical and Education

Adobe’s View - Firefly

Firefly: Adobe’s Image-Generating Tool integrated in Photoshop.

- ➔ They trained the model using Adobe’s library of stock photos.
- ➔ Offers creators extra compensation when their material is used.
- ➔ Each generated output from Firefly comes with labels that indicate that it has been created using AI tools.

“Building Responsible Tech does not have to come at the cost of doing Business!”

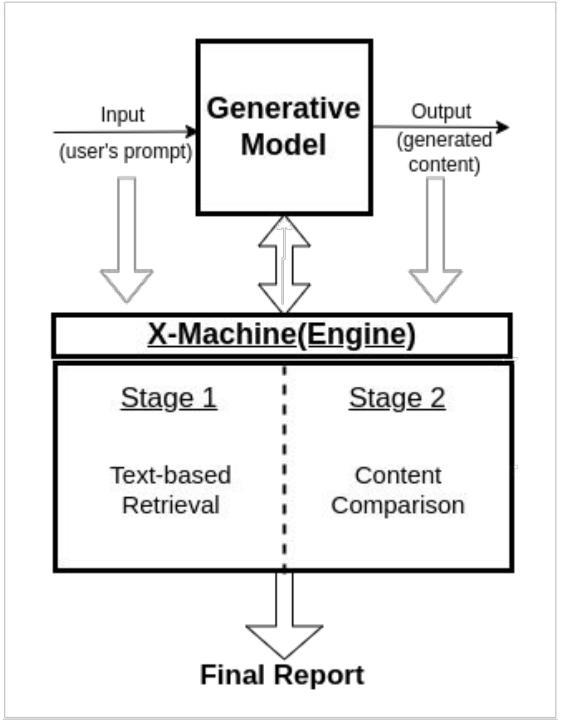
Adobe AI leaders.

Could I have an engine to understand the AI models?

Our Approach

Stage 1 : Try to retrieve most similar images based on the text explanation of the training data.

- ➔ Use traditional Retrieval methods(indexing etc.)
- ➔ Try to combine methods.



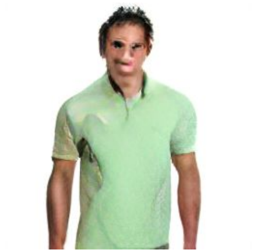
Stage 2 : After reducing search space, compare the remain content and find the final most crucial samples for the generation

- ➔ Raw Image Similarities
- ➔ Embeddings from Res50Net
- ➔ Combine them in order to finally rank the training images.

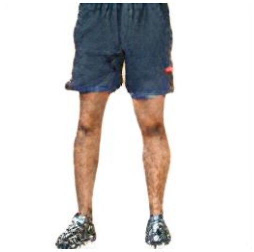
Local Generative Model

- Use of Pytorch-Dalle Package developed by OpenAI to replicate the functionality of the DALL·E model for text-to-image generation.
- NUWA extends DALL·E's capabilities to video and audio synthesis.
- Tested on popular small datasets (CIFAR - 10, CUB - 200, Fashion Items, Painting Dataset) and bigger ones (COCO - Dataset).
- A few of our models for fashion items generation are located on Colab as Demo to try it with your prompts.

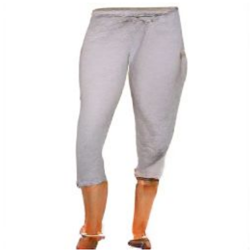
Examples of Local Generations



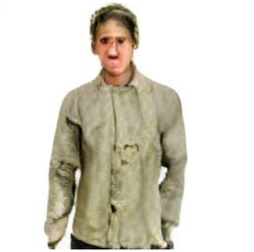
ADIDAS Mens Classic Green Polo T-shirt



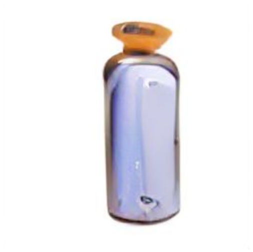
ADIDAS Mens Chelsea Striped Navy Shorts



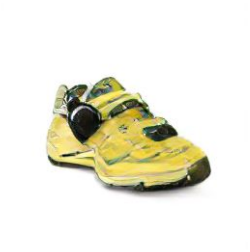
ADIDAS Womens Essential Cinder Grey Capris



Lee Men Olive Green Shirt



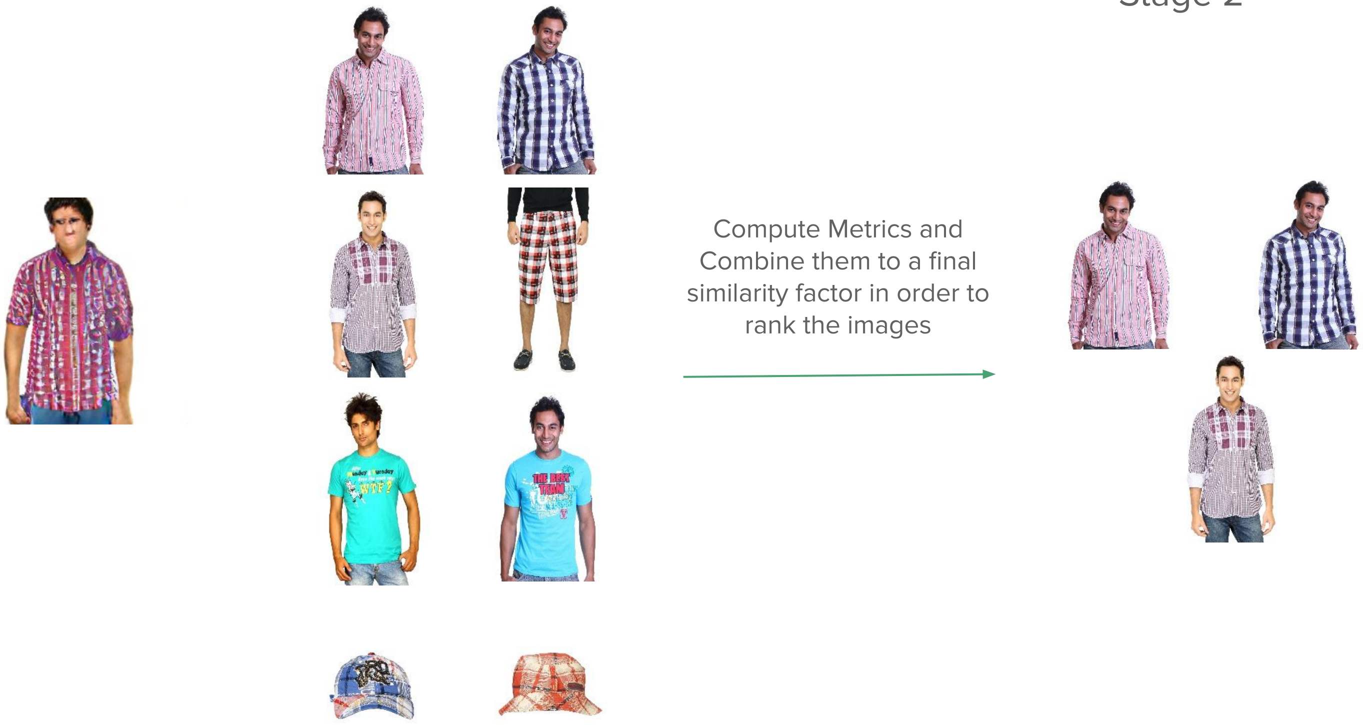
Quechua Easy-to-Carry Blue Water Bottle



Quechua Mens Arpenaz Flex Yellow Shoe

Searching in the training Dataset

Stage 1



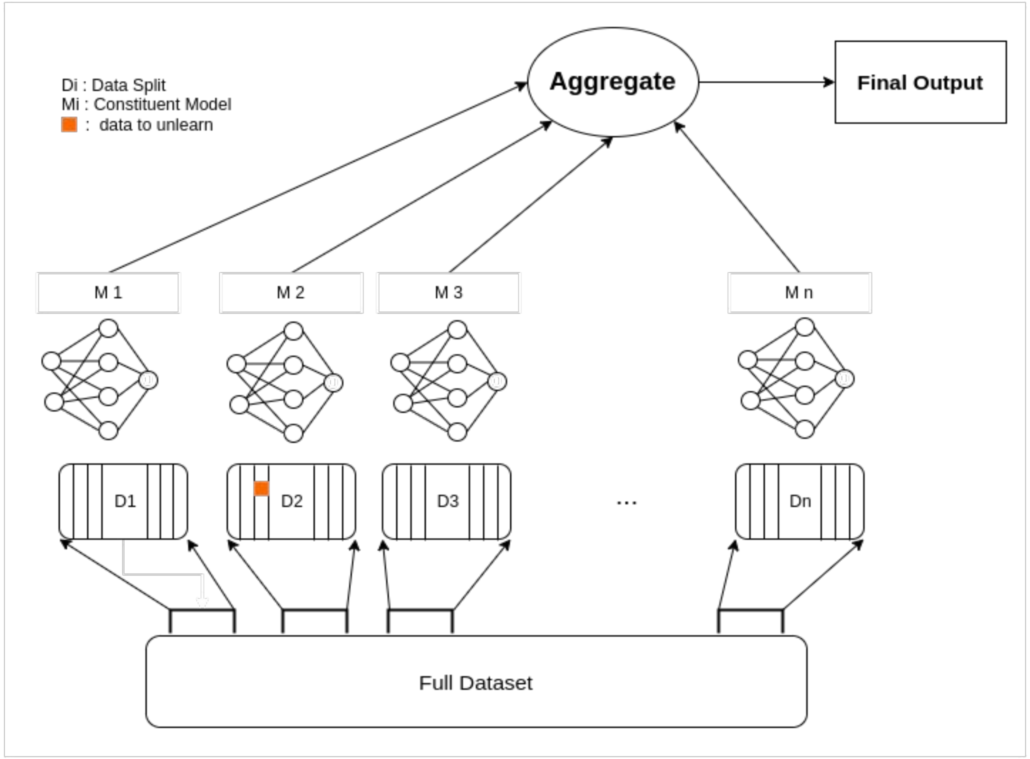
Compute Metrics and Combine them to a final similarity factor in order to rank the images

Stage 2

Now We have to "forget" them!

How to “Forget” them ?

- Retrain from Scratch.
- Probabilistic Methods.
- SISA Framework.



SISA Framework

Challenges

- Train a model with qualitative generations (Computer Resources, Large-Scale Training Dataset).

Trying to take advantage of the popular generative models we make the following **assumption**:

- Their training dataset is located in the well-known Search Engines.
  - We select **DuckDuckGo** search Engine.
  - Use **Midjourney** as a stable dataset of generated images with their text prompts (collected from several Discord channels where users asked for an AI generated image giving their text prompt.).

DuckDuckGo

Advantages of DuckDuckGo:

- Not tracking or storing personal information.
- Search without history or personal information being recorded.
- Does not engage in targeted advertising based on user data.
- Does not track users' locations or use location data to personalize search results.

All the searches were made via DuckDuckGo's API.

Midjourney Dataset

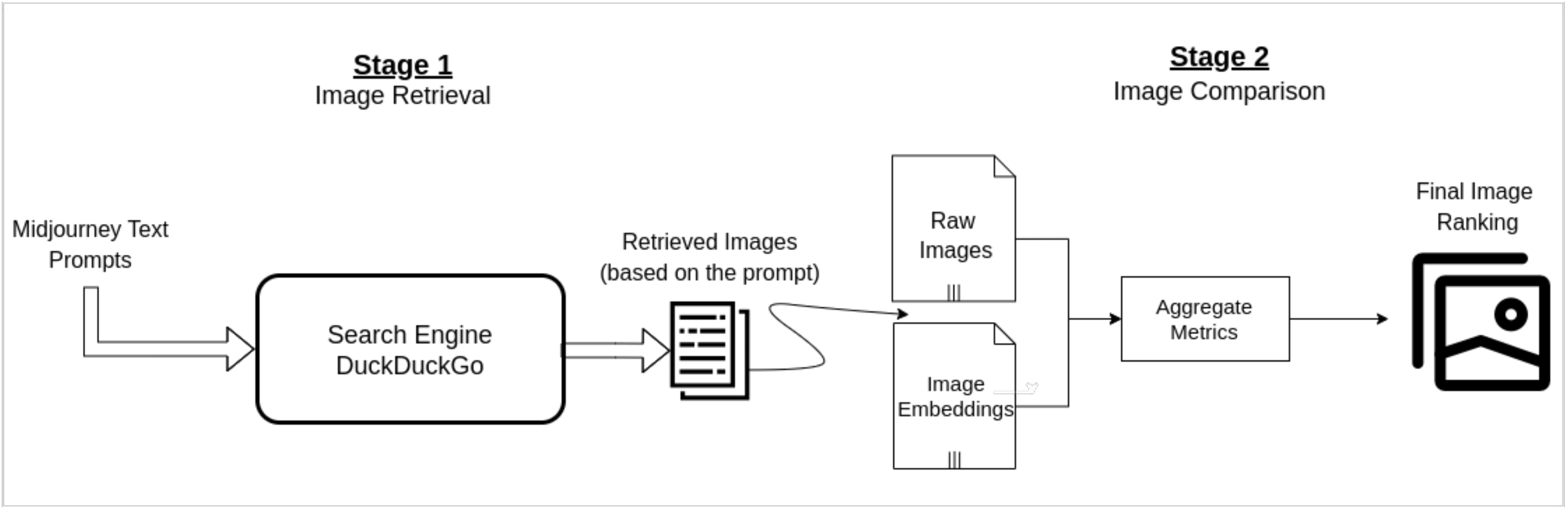
a fisherman village in indonesia by Ralph Steadman, dark velvet, yellow and teal, coconut tree

woman in a red dress sitting on a white chaise lounge in front of a large window in an elegant futuristic apartment, evening light, high-tech neoclassical architecture, babylonian hanging garden cityscape in the background

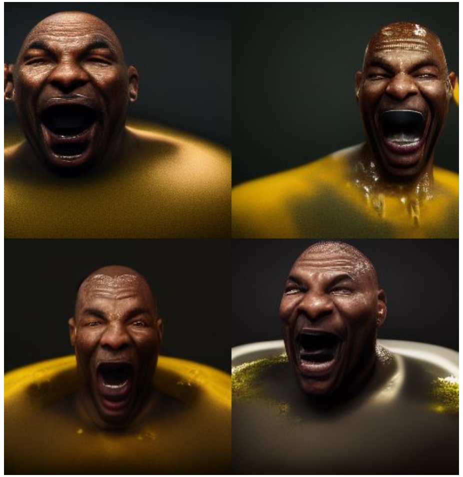
skinny white shark man puppet with cigarette, long black hair, big eyes, standing in studio, direct wide angle view

Robotic Mushroom

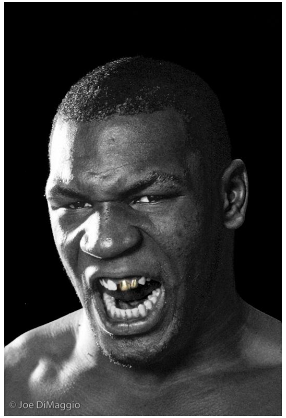
Experiments with DuckDuckGo

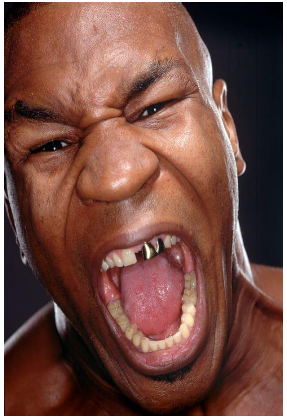


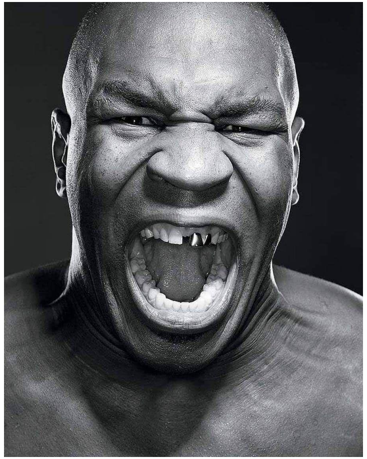
Results of Ranking



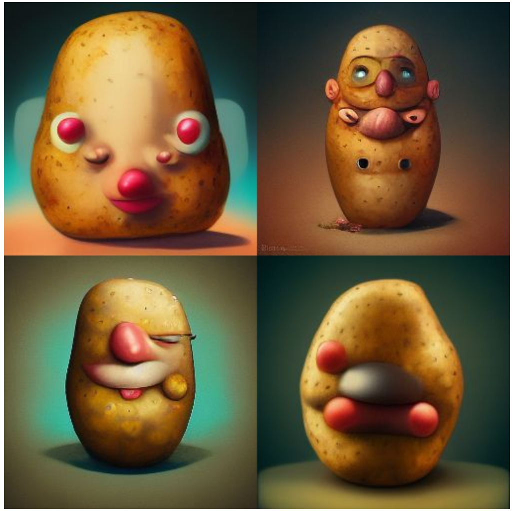
Mike Tyson screaming while covered in olive oil , realistic, 4k, shiny, render with octane



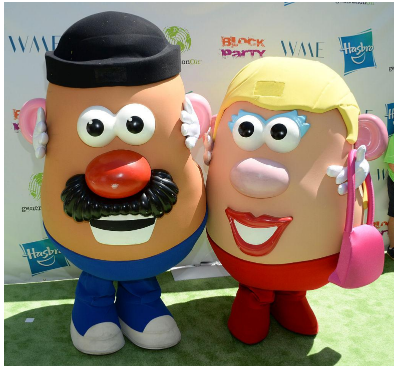




More Results …



potato head



Future Steps

- Local Generations
  - Train a larger-scale model with bigger or more general datasets.
  - Try to improve the quality of the generated images.
- Efficient Retrieval
  - Improve text retrieval methods.
  - Compare more efficiently the most similar images.
- User’s Interaction
  - Conduct Surveys to gain knowledge from experts

References

https://www.neebal.com/blog/generative-ai-vs.-predictive-ai-unraveling-the-distinctions-and-applications

https://www.turing.com/resources/generative-ai-applications

https://redblink.com/generative-ai-applications-use-cases/

https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48

https://www.xcally.com/news/interpretability-vs-explainability-understanding-the-importance-in-artificial-intelligence/

https://www.technologyreview.com/2024/03/26/1090129/how-adobes-bet-on-non-exploitative-ai-is-paying-off/?utm_source=LinkedIn&utm_medium=tr_social&utm_campaign=site_visitor.unpaid.engagement

https://ambiata.com/blog/2021-04-12-xai-part-1/

https://github.com/lucidrains/DALLE-pytorch

https://arxiv.org/abs/1912.03817

https://duckduckgo.com/

https://paperswithcode.com/paper/generated-faces-in-the-wild-quantitative