# TEXT TO IMAGE GENERATIVE WEB APPLICATION USING DALL-E 2

**MINI PROJECT REPORT**

Submitted to the Department of Computer Applications, Bharathiar University in partial fulfilment of the requirements for the award of the degree of

**MASTER OF SCIENCE IN DATA ANALYTICS**

Submitted by

**P R YASHWANTH RAJAN (22CSEG34)**

Under the guidance of

## Mr. K. MOORTHY, MCA.,

Department of Computer Applications

|  |
| --- |
|  |
|  |  |



**DEPARTMENT OF COMPUTER APPLICATIONS**

**BHARATHIAR UNIVERSITY**

**COIMBATORE – 641 046**

**DECEMBER – 2023**

**DECLARATION**

I hereby declare that this project, titled **“TEXT TO IMAGE GENERATIVE WEB APPLICATION USING DALL-E 2 2”** submitted to the Department of Computer Applications, Bharathiar University, Coimbatore is a record of original project work done by **P R YASHWANTH RAJAN (22CSEG34)** under the supervision and guidance of **Mr. K. MOORTHY** Department of Computer Applications, Bharathiar University, Coimbatore and that this project work has not previously formed the basis of the award of the Degree / Diploma / Associate Ship / Fellowship or similar title to any candidate of any university.

**Place:** Coimbatore Signature of Candidate

**Date: (P R YASHWANTH RAJAN)**

Countersigned by

## Mr. K. MOORTHY, MCA.,

Department of Computer Application

**CERTIFICATE**

This is to certify that the project work title **“TEXT TO IMAGE GENERATIVE WEB APPLICATION USING DALL-E 2 2”** submitted to the Department of Computer Applications, Bharathiar University in partial fulfilment of the requirement for the award of a Degree in **Master of Science in Data Analytics,** is a record of the original work done by **P R YASHWANTH RAJAN (22CSEG34)** under my supervision and guidance and this project work has not formed the basis of the award of any Degree / Diploma / Associate Ship / Fellowship or similar title to any candidate of any university.

**Place:** Coimbatore

## Date:

**Project Guide Head of the Department**

Submitted for the University Viva-Voce Examination held on ……………

**Internal Examiner External Examiner**

# ACKNOWLEDGEMENT

I express my respectful thanks to our Professor & Head of the Department **Dr. M. Punithavalli, B.Sc., M.Sc., M.Phil., Ph.D,** Department of Computer Applications, Bharathiar University, for permitting me to carry out my mini project report work in **“TEXT TO IMAGE GENERATIVE WEB APPLICATION USING DALL-E 2 2”**

I really deem it a special privilege to convey my prodigious and everlasting thanks to my guide **Mr. K. MOORTHY, MCA.,** Department of Computer Applications, Bharathiar University, for her valuable guidance and personal interest in this mini project report work. Last but not least, I also acknowledge the help done by my parents and acknowledge the encouraging support of my friends who were involved in this mini project, in one way or the other. I thank Almighty for showering the divine grace on me and offer prayers to the lord for everything that was given to me.

**ABSTRACT**

In an era where creativity meets artificial intelligence, this project introduces a cutting-edge web application leveraging OpenAI's DALL-E 2. The proposed platform is a Text-to-Image Generative Web Application designed to transform textual prompts into vivid AI-generated images. Users can input descriptive text, and the application harnesses the power of DALL-E 2 to create visually compelling Images. The core functionality enables users to witness their textual ideas materialize into unique and diverse visual compositions.

DALL-E 2's ability to understand and interpret textual input opens a gateway to limitless creative possibilities. Going beyond static outputs, the application offers the capability to generate multiple variations of the given input image. This feature allows users to explore diverse perspectives and artistic renditions of their initial concepts. Empowering users with creative control, the web application incorporates image editing functionalities. Users can refine and tailor the images, adding a personal touch to the AI-created visuals. The user-friendly interface is built on Streamlit, ensuring an intuitive and accessible experience. Streamlit's simplicity and versatility contribute to a seamless interaction between the user and the AI-powered image generation engine.

|  |  |  |
| --- | --- | --- |
| **S.No** | **TITLE** | **Page No** |
| 1. | INTRODUCTION |  |
| 2. | LITERATURE REVIEW |  |
| 3. | COMPONENTS AND MODELS USED   1. CLIP 2. PRIOR 3. Diffusion Model 4. GLIDE (decoder) |  |
| 4. | ARCHITECTURE AND WORKING |  |
| 5. | LIMITATIONS AND RISKS |  |
|  | MITIGATIONS |  |
| 6. | WEB APPLICATION   1. Streamlit 2. User Interface 3. Working Results |  |
| 8.  7. | CONCLUSTION |  |
|  | BIBLIOGRAPHY |  |

**TABLE OF CONTENTS**

**INTRODUCTION**

In the realm of artificial intelligence and creative innovation, OpenAI's DALL-E 2 emerges as a groundbreaking text-to-image generative model. DALL-E 2 showcases the transformative power of advanced neural networks, blending the capabilities of transformers with the ingenuity of generative models. At its core, DALL-E 2 redefines the boundaries of creative expression by seamlessly translating textual prompts into vivid and diverse visual compositions. This generative marvel is equipped with a dual architecture a fusion of a Variational Autoencoder (VAE) and Transformers. The VAE facilitates the creation of diverse and realistic image samples, while the Transformer's self-attention mechanisms enable the model to grasp intricate patterns and relationships within the data.

DALL-E 2's architecture involves taking input from textual prompts (text embedding), incorporating a learned prior distribution, and processing image information (image embedding). The decoders then use this combined information to generate images that align with the given textual descriptions.

DALL-E 2's training process involves exposure to a vast dataset of images and their corresponding textual descriptions, enabling the model to learn the intricate mapping between text and images. The result is an AI-powered engine capable of turning textual ideas into visually compelling art. It not only generates images from text but also produces variations for given images, allowing for creative exploration, and enables picture editing by seamlessly incorporating new objects.

In this project, we harness the creative potential of DALL-E 2 to develop a web application that not only transforms text into images but also empowers users with the ability to explore variations, edit generated images, and engage with the process through an intuitive Streamlit web interface. As we integrate state-of-the-art AI with user-friendly design, the project invites users to embark on a journey of collaborative creation, where human imagination converges with the capabilities of DALL-E 2, opening up new horizons for creative expression.

**LITERATURE REVIEW**

Generative models have recently gained attention for their use in producing fake images. The emergence of artificial intelligence (AI)-generated fake images, referred to as “deep fakes” presents several challenges such as developing synthetic images that look realistic, images with multiple objects, and reliable evaluation metrices that align with human judgment [13].

However, new technologies appear promising for image generation. Developing a system that can create images representing a given textual description inspired by how humans perceive is a significant step towards computer intelligence [13]. Prior to the development of generative models, the process of creating an image from text relied on image querying. Which involved selecting the best collection of images from an image database to illustrate text description. Recent advances in artificial intelligence and computer vision have facilitated the creation of images based on text descriptions. The goal of text-to-image synthesis is to generate images from textual descriptions.

Research on text-to-image creation has sparked much interest due to its applications in art, marketing, business, and education, among others. Several frameworks and improvements have been proposed to produce more visually realistic images. The representation of an image as the numeric values of its pixels is termed as image data distribution. If there is an image with the dimensions *m* x *n* pixels.

The image may be interpreted as a vector with *m* x *n* dimensions, which is high dimensional data distribution. Image data distribution of high dimension makes image creation a complex problem. The generated images should be visually realistic and semantically accurate for adequate text-to-image synthesis. Semantic accuracy refers to the agreement between the content of the image and the text description. Generative modelling is when we provide the model with a description of what we want to generate, and the model returns an image. The model automatically learns from the input data and replicates it with variety and accuracy. Suppose we have a description using which we try to generate an image. The generated image will resemble but not be identical to the input sample image because the model uses input images to learn the image’s representation. That is why the representation is unique each time and varies for different models. There are different types of generative models like Autoencoders [16, 21], Generative Adversarial Network (GAN) [16], and Diffusion models [35] for text-to-image generation that several authors have introduced over the years. These models have been compared based on visual realism, diversity, and semantic alignment to understand which model works better at generating an image related to the text used for its generation [9].

After being trained on image data, experiments have shown that producing fake but photorealistic images is feasible. Ramesh et al. [32] showed text-to-image generation based on an autoregressive transformer in terms of zero-shot learning. Zero-shot learning implies creating samples for the text input without being trained on the same input. Scaling can result in more accurate generalization compared to earlier domain specific techniques and the approach achieves the best frechet inception distance [44] and the highest inception score [1] when qualitatively comparing samples from proposed model to those from prior work [32].

This approach is used by the popular text-to-image generation tool DALL.E [9], which became available for public use by OpenAI shortly after the diffusion model was introduced. The diffusion model has been proven to generate high-quality images and give desirable characteristics like distribution coverage, a stationary training aim of creating images from text, and simple scalability [9]. With these improvements, they achieve a new state-of-the-art, outperforming GANs on a variety of metrics and data sets [9].

DALL.E was released in January 2021, followed by DALL-E 2 2, which was released later in November 2022. Meanwhile, diffusion models gained popularity in the vision community. OpenAI chose this method as the basis for DALL-E 2because it uses simple image-denoising nets to reduce a convex regression loss instead of a minmax. DALL.E and other generative AI picture tools are the latest innovation that venture capitalists have been eager to try out. The use of non-fungible tokens (NFTs), a kind of digital asset that can be in the form of digital art, has skyrocketed. NFT artworks are already fetching millions of dollars [4]. With enough imagination, people can create digital art using text-to-image generators, which can be utilized to create NFTs.

Creating realistic graphics from natural language can enable people to produce rich and varied visual content. There has been impressive progress in the first few years since Mansimov et al. [24] started with modern machine learning approaches for text-to-image synthesis [32]. The approach has numerous commercial, educational, and artistic applications. One of the commercial applications is the expensive production of video games and animated films, in which many production artists are employed to perform very mundane tasks. Text-to-image models can create and colourize characters automatically by giving their descriptions. The tool can be used to generated images for books and for teaching which makes visual learning easier and more accessible. The authors of short story books can use the tool for creating images related to the stories.

**COMPONENTS AND MODELS USED**

**DALL -E 2** is a text-to-image AI system, or a CLIP system (connecting text to images). It is an encoder-decoder model, meaning when text is given, this text is encoded into a machine input, processed by the machine, and then fed through a decoder that decodes this into a visible image. DALL-E 2 sits at the intersection of deep natural language processing and computer vision generation and is known as a **Hierarchical Text-Conditional Image Generation model**. The training set is simply pairs of images and their captions, and DALL-E 2’s goal is to train two models: The first is the prior, which is trained on and takes in a text caption, and produces a CLIP image embedding. The second is our decoder, which takes our CLIP image embedding (and optionally a text caption) and produces a learned image.

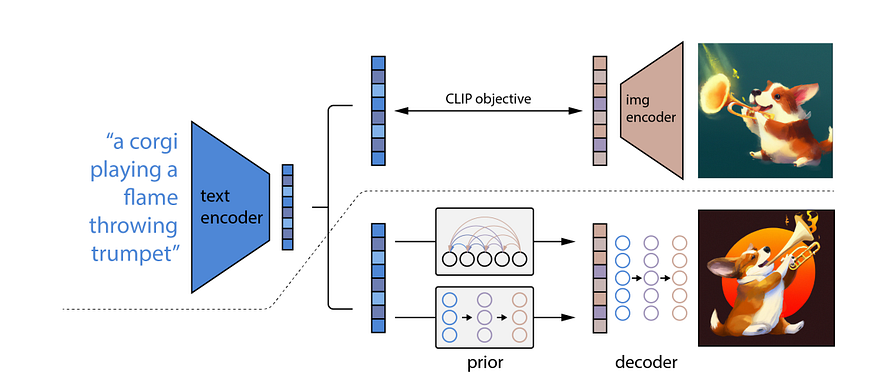


Fig 1. Components of DALL-E 2

The Each components listed plays important roles in this model:

1. CLIP
2. Decoder
3. PRIOR
4. GLIDE(Decoder)

**1. CLIP** (Contrastive Language–Image Pre-training)

CLIP is a neural network model that undergoes training on an extensive dataset comprising 400,000,000 (image, text) pairs. Each pair consists of a picture matched with its corresponding caption, creating a diverse set of 400,000,000 pictures and captions for model training. Remarkably, CLIP possesses the capability to predict the most relevant text snippet when provided with an image input. This implies that you can input an image into the CLIP model, and it will return the most likely caption or summary for that particular image. Intriguingly, CLIP achieves this without directly optimizing for the task, akin to the zero-shot capabilities demonstrated by models like GPT-2 and GPT-3.

Zero-shot learning is a phenomenon observed in models such as CLIP, GPT-2, and GPT-3. These models tend to excel in tasks for which they were not explicitly trained, marking a notable strength in their adaptability. For instance, in "zero-shot learning," a model endeavors to predict a class it has never encountered in the training data. Using a model trained exclusively on cats and dogs to detect raccoons is an example of this capability. CLIP, leveraging the textual information in (image, text) pairs, excels in zero-shot learning scenarios. Even when confronted with images vastly different from its training set, CLIP demonstrates a remarkable ability to provide accurate captions, showcasing its versatility and robust performance in diverse visual contexts. CLIP encodes text and image in same embedding space. This gives us opportunity to work with intersection of image and text modalities

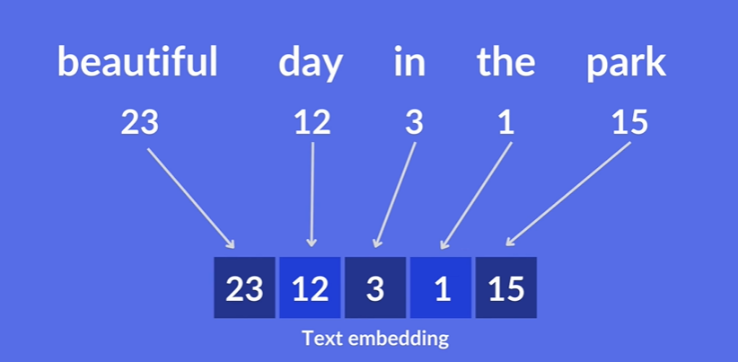
There are Two Embeddings text and image embeddings are both used to create a shared representation space, enabling direct comparisons between images and their associated textual descriptions.

Fig 2. Text embedding

In the image below, you'll see a set of purple text cards going into the text encoder. The output for each card would be a series of numbers. For example, the top card, pepper the aussie pup would enter the text encoder – the thing smashing it into mathematical space – and come out as a series of numbers like (0, 0.2, 0.8).

The exact same thing will happen for the images: each image will go into the image encoder and the output for each image will also be a series of numbers. The picture of, presumably Pepper the Aussie pup, will come out like (0.05, 0.25, 0.7).

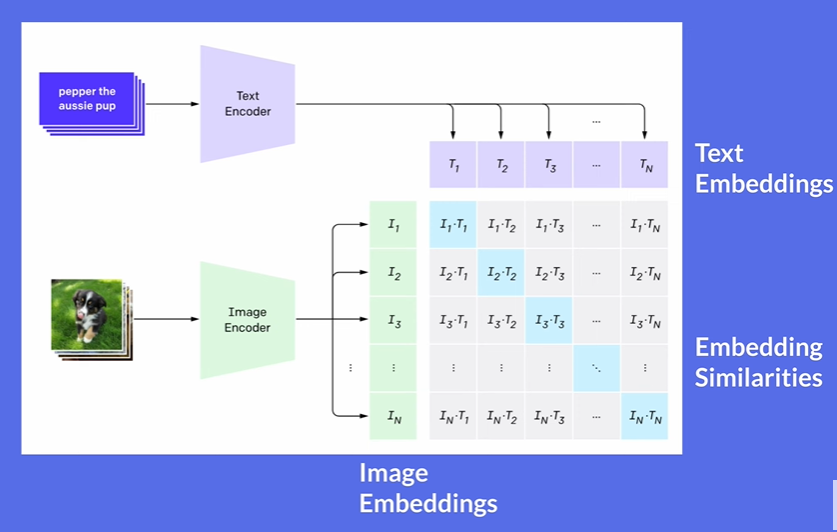


Fig 3. clip Architecture

The embedding of the first image's caption and values in the matrix being the similarity between intersecting embeddings we want the blue highlighted cells to have the highest value and the grey ones to have the lowest value. The text encoder and image encoder get fit at the same time by simultaneously maximizing the cosine similarity of those blue squares and minimizing the cosine similarity of the grey squares, across all of our text + image pairs.

**2.PRIOR**

In machine learning and generative modelling, a "prior" often refers to a probability distribution that captures prior knowledge or assumptions about the data. This distribution is incorporated into the model to guide the generation process. In the case of generative models like DALL-E 2, the "prior" might play a role in shaping the latent space or influencing the diversity and characteristics of generated samples.

The term "prior" typically refers to the prior distribution in the VQ-VAE-2 (Vector Quantized Variational Autoencoder 2) architecture, which is a component of the model.

**Prior Distribution in VQ-VAE-2:**

The prior distribution in VQ-VAE-2 refers to the probability distribution from which the discrete latent codes are sampled. In the context of DALL-E 2, this distribution plays a crucial role in the vector quantization process.

**Vector Quantization:**

The prior distribution is used in the vector quantization process to map continuous values from the encoder to discrete codes. During the encoding phase, the continuous values in the latent space are compared to the discrete codebook, and the prior distribution is used to determine which discrete code is most likely to represent the input.

**Discretization of Latent Space:**

The prior is instrumental in discretizing the latent space into a set of discrete codes. This discretization is a key feature of DALL-E 2, as it allows for more efficient training and improves the model's ability to generate diverse and novel images by representing concepts with distinct codes.

**Sampling Discrete Codes:**

During both training and generation, discrete codes are sampled from the prior distribution. This sampling. Process introduces a level of stochasticity, contributing to the variety of images that dall-e 2 cangenerate even when provided with similar or identical textual prompts.

**Training Objective:**

The prior distribution is incorporated into the overall training objective of DALL-E 2. The model is trained to adjust its parameters in such a way that the sampled codes from the prior, when passed through the decoder, result in images that closely match the input data.

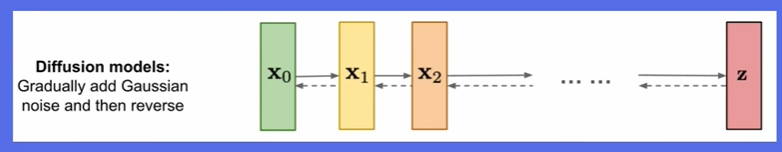
**Balancing Trade-offs:**

The prior distribution helps in balancing the trade-offs between fidelity and diversity in image generation. By influencing the sampling of discrete codes, the prior contributes to the overall strategy of DALL-E 2 in producing varied and high-quality images based on the encoded text.

**3.Diffusion Model:**

Diffusion models are inspired by non-equilibrium thermodynamics. They define a Markov chain of diffusion steps to slowly add random noise to data and then learn to reverse the diffusion process to construct desired data samples from the noise. Unlike VAE or flow models, diffusion models are learned with a fixed procedure and the latent variable has high dimensionality (same as the original data).

The Diffusion model is one of the prior model, they are generative models their working principles are quite simple, where as it take a piece of data for e.g., a picture it randomly add noise to it over time steps until it is not recognized , and from that point it try’s to reconstruct the image to its original form . By doing this they learn to generate data or images over time.



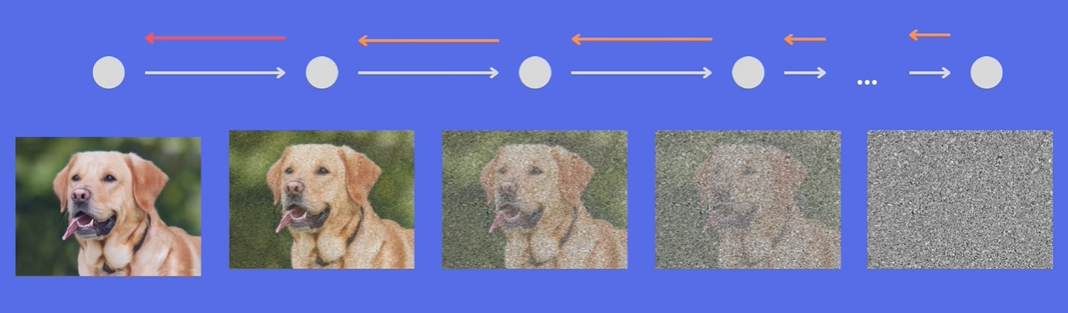


Fig 4. Diffusion model

**4. GLIDE** (Decoder)**:**

The GLIDE is a Diffusion model which is used as a decoder in this model, it is an image generation model, but differently from a pure diffusion model, it also includes the text embedding that was given to the model to support the image creation Process, so at the end you will be creating an image based on the text. In DALL-E 2 after setting up the decoder not only it generates the image based on the glide, additionally it also include the CLIP embeddings to support the Image generation.

However (64x64 px) is generated there will be two up-sampling steps are included to increase the resolution upto (1024x1024 px).

Fig 5. GLIDE Architecture

**ARCHITECTURE AND WORKING**

The DALL-E 2 is OpenAI ‘s High resolution Image generative model that can create high resolution images by taking Text Prompts as Input. It can produce original and realistic Images. On top of that it can also edit images, add new information and produce variations.

The Architecture consists of Two parts:

1. Converting captions into a representation of an Image called the PRIOR.
2. Convert the representation into an actual Image called Decoder

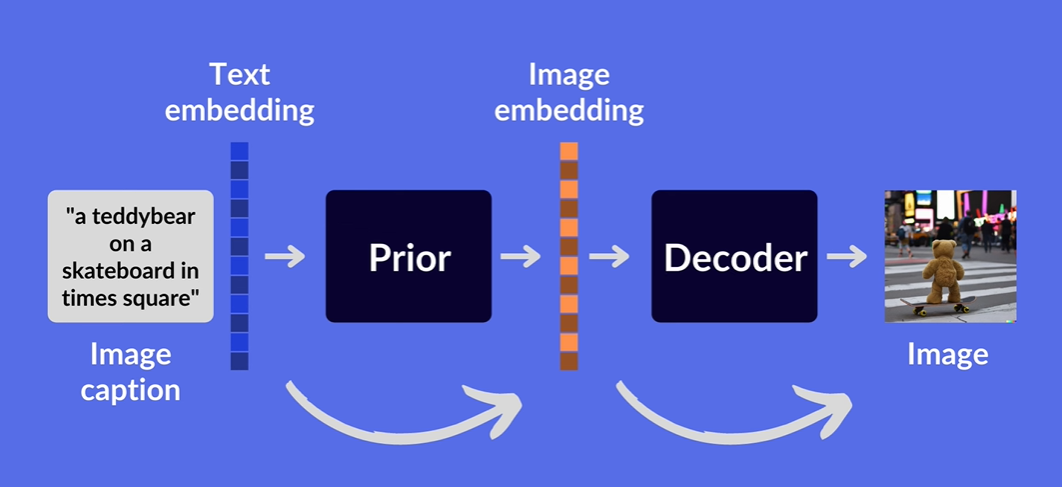


Fig 6. Dall-e 2 Architecture

The text and image representation used in Dall-e 2 are coming from another technology , developed by OpenAI called CLIP.

It is a neural network model that returns the best caption for the given image it basically does the opposite of the Dall-e 2 , It is still helpful it is a contrastive model so it doesn’t try to classify images instead it match images to their corresponding captions , it is trained on image and caption pairs that is collected from the Internet.

To do the matching CLIP trains two encoders:

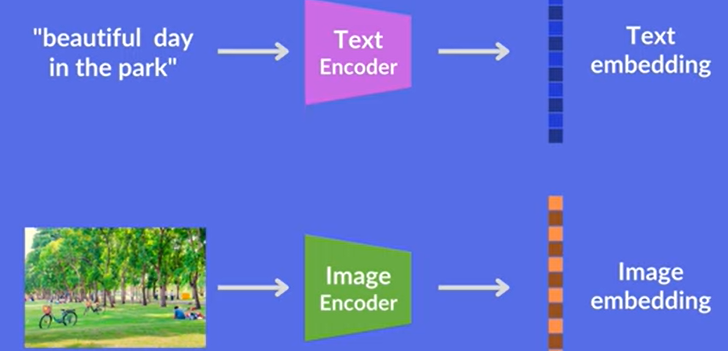
* + Encoder 1 – Turns Images into Image embeddings
  + Encoder 2 – Turns Text or caption into Text embeddings
  + Text Embedding - taking a sentence and mapping the text into vectors

Fig 7. Text and Image encoder

The problem CLIP is optimizing for making sure that the similarity between the embedding of an image and the embedding of its caption is as high as possible.

The Blue vector in the image is the clip text embedding and the orange vector is CLIP image embedding, the prior takes the clip text embedding which is easily generated from the caption through the clip text encoder and creates a clip image embedding out of it in the research paper they try two different options for the prior which are **Autoregressive and Diffusion prior.** The diffusion prior worked better for the Dall-e 2.

Once the prior creates the clipped image embedding the next step is to create the image itself and the decoder is responsible for that, what happens when we just pass the caption or text embedding directly to the decoder. When we pass the caption and the text embedding directly to the decoder they found that having the prior actually yields the best result.

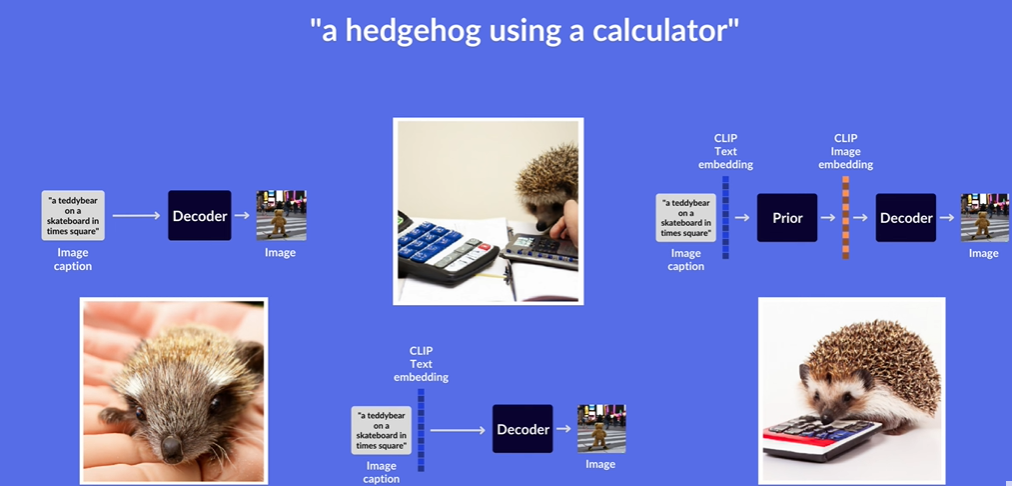


Fig 8. Passing Text and Captions without PRIOR VS with PRIOR

**Example:**

The above fig 8 , for the caption “ a hedgehog using a calculator ” passing the caption and text directly to the decoder gives the first and the second images as a result , while passing them with the prior gives the third image.

The **Decoder** used in the dall-e 2 is a diffusion model called GLIDE, it is also an image generation model but differently from pure diffusion model it includes the embedding of the text that was given to the model, thus the decoder is setup so that is also includes the clip embeddings at the end to support the image creation process you will be creating the image based on the text and the clip embeddings at the end.

**Image Variation**

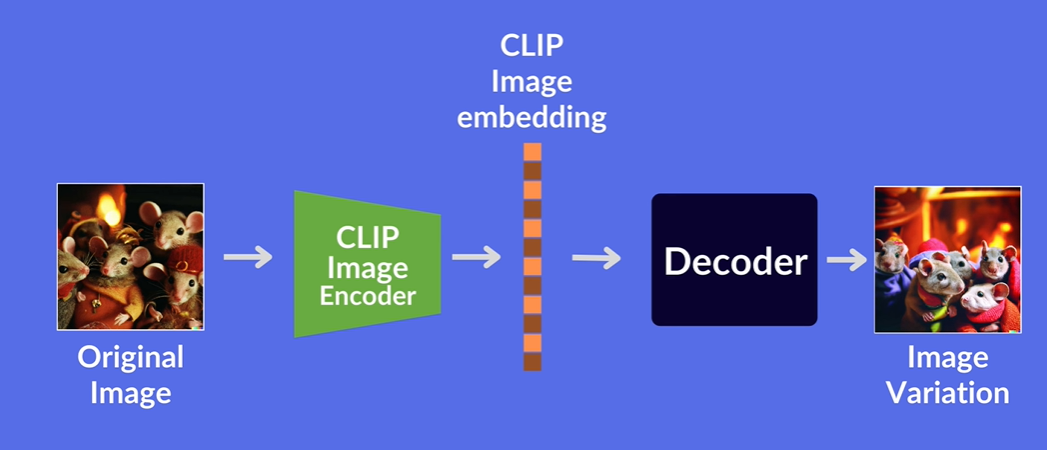
Making a variation of a given image means that keeping the main elements and style of the image as same but change the trivial details. In Dall-E 2 this is done by obtaining the images clip image embedding and running that through the decoder by encoding an image using CLIP and then decoding the image embedding using the diffusion decoder, we can also see what information was caught by clip and what information was lost.

Fig 9. Image variation Architecture

**For example:**

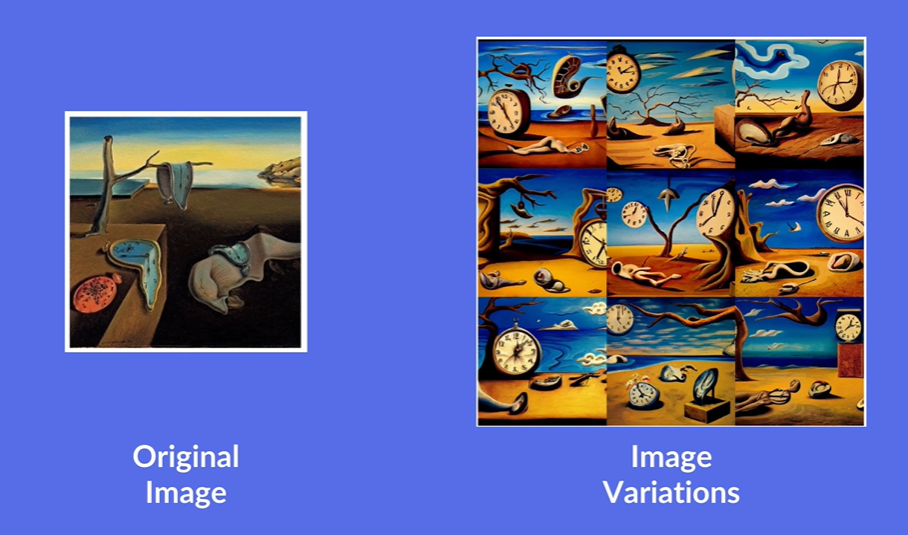
 The below is the Salvador dali’s painting manages to keep the detail of the existence of a clock as well as the stylistic details while varying the trivial details of the image.

Fig 10. Original VS Image variation example

**LIMITATIONS**

**1. Physical Quality Attribution**

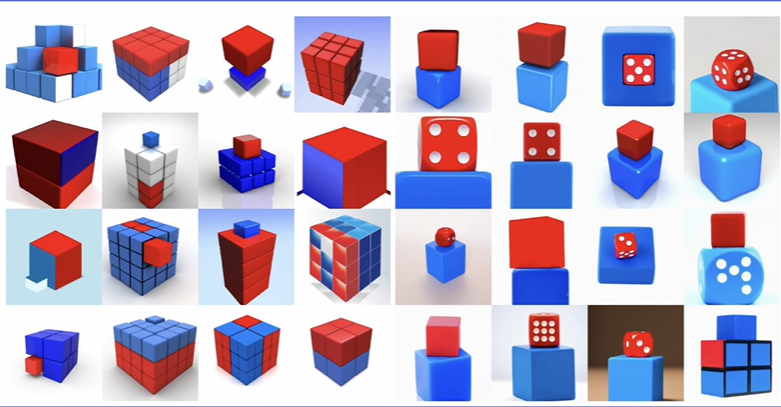
 In the scientific paper that accompanied the unveiling of DALL-E 2, OpenAI points out some limitations of the system. For example, the researchers tested DALL-E's ability to perform compositionality, which is the meaningful merging of multiple object properties, such as color, shape and positioning in the image.The tests show that DALL-E 2 does not understand the logical relationships given in the descriptions and therefore arranges colored cubes incorrectly, for example. The following motifs show DALL-E's attempt to place a red cube on top of a blue cube.

Fig 11. Picture of dall-e 2 try to interpret the red and blue box on top of one another.

**2 .Coherent TEXT IN IMAGES**

It is not good at creating a coherent text in images as shown below , it came out when asked to create an image sign that says “ Deep learning ”



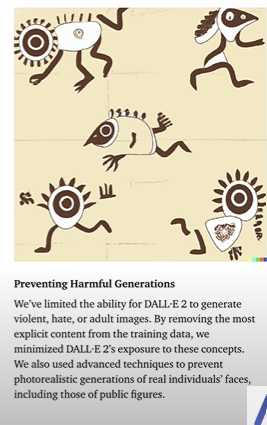
Fig 11. Image generated for prompt of “ Deep learning ” sign

**3.Detailed Scenes**

The model has a hard time producing details in complex scenes for example, when generating an image of the Time Square the screens seem to not have any Readable or understandable details to them.

Fig 12. Generating a image of Time square

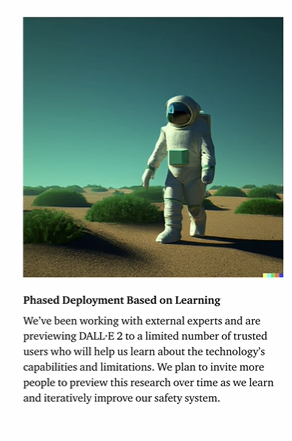
**MITIGATIONS**

**Preventing Harmful Generations**

OpenAi limited the ability for DALL-E 2 to generate violent, hate, or adult images. By removing the most explicit content from the training data, by minimizing DALL-E 2's exposure to these concepts. They also used advanced techniques to prevent photorealistic generations of real individuals' faces, including those of public figures.

**Curbing Misuse**

The content policy does not allow users to generate violent, adult, or political content, among other categories. It won't generate images if the filters identify text prompts and image uploads that may violate the policies. They also have automated and human monitoring systems to guard against misuse.

**Phased Deployment Based on Learning**

We've been working with external experts and are previewing DALL-E 2 to a limited number of trusted users who will help us learn about the technology's capabilities and limitations. We plan to invite more people to preview this research over time as we learn and iteratively improve our safety system.