**Stable Diffusion**

It was developed by the company Stability AI. We are currently using the version 3. The first version came in August 2022.

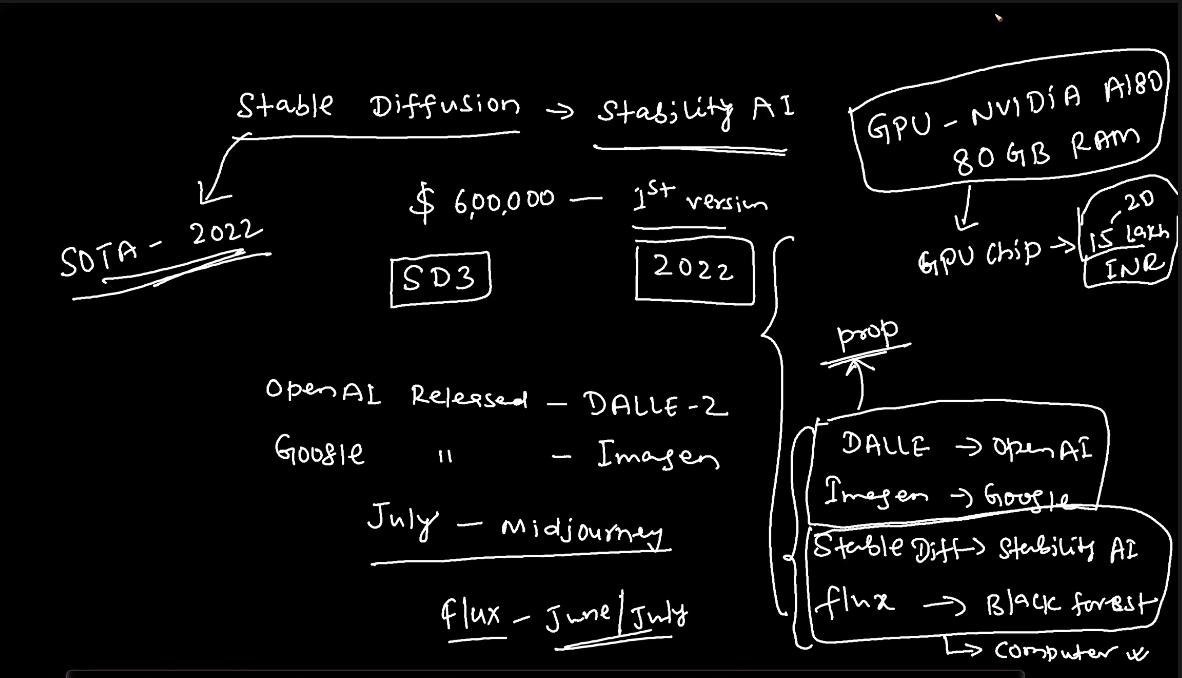
It has competitors like OpenAI – DALL-E-2 (April 2022)

Google – Imagen (May 2022)

MidJournery (2022). But none of them is able to beat Stable Diffusion. Another model called Flux came in June-July 2024 which beated Stable diffusion also.

Flux was developed by the company called Black Forest. Stable diffusion and flux both work on a physics principle with the same name.

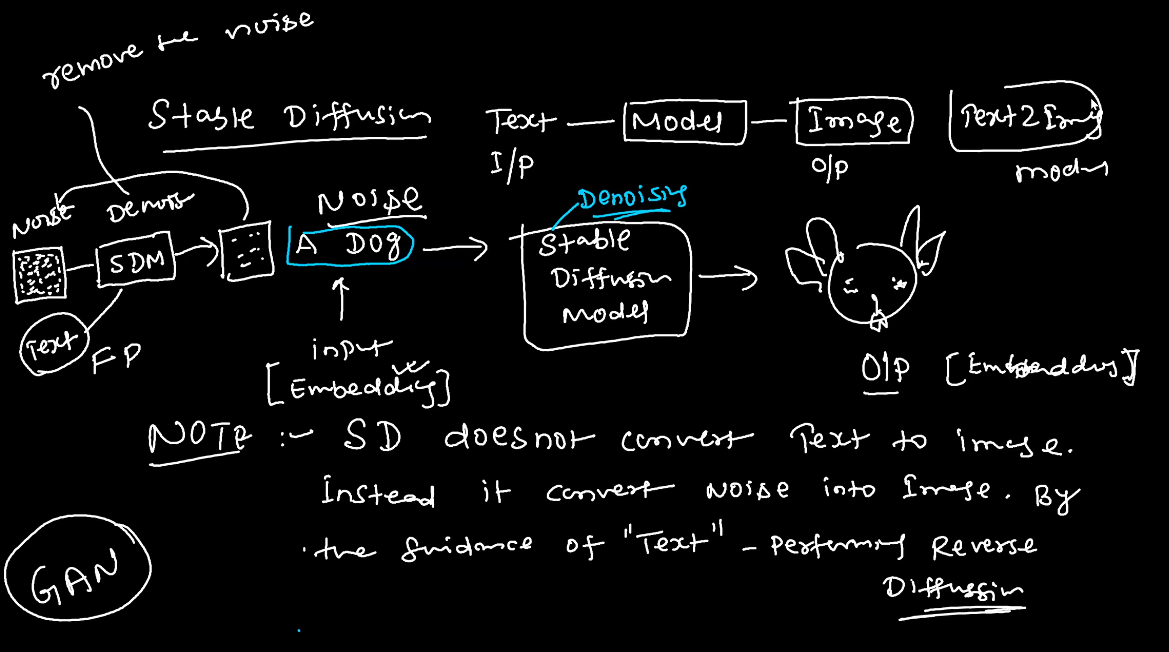
Open AI (Dalle) and Imagen (Google) are both proprietary (they have not disclosed their code) while Stable diffusion and Flux are both open source but costly and require a huge computational capacity.



Note:- Stable diffusion does not converts the text into image.

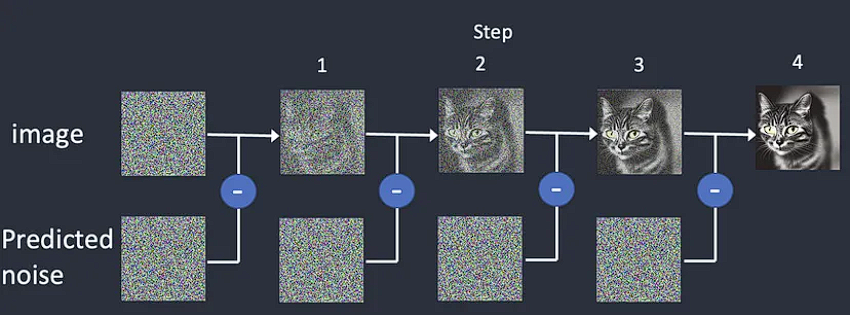
We pass input as a text data which is treated as an embedding (or noise) and stable diffusion de noises the data and generates image which is basically pixel values. We can also say that it converts noise into image. By the guidance of text, performing reverse diffusion.

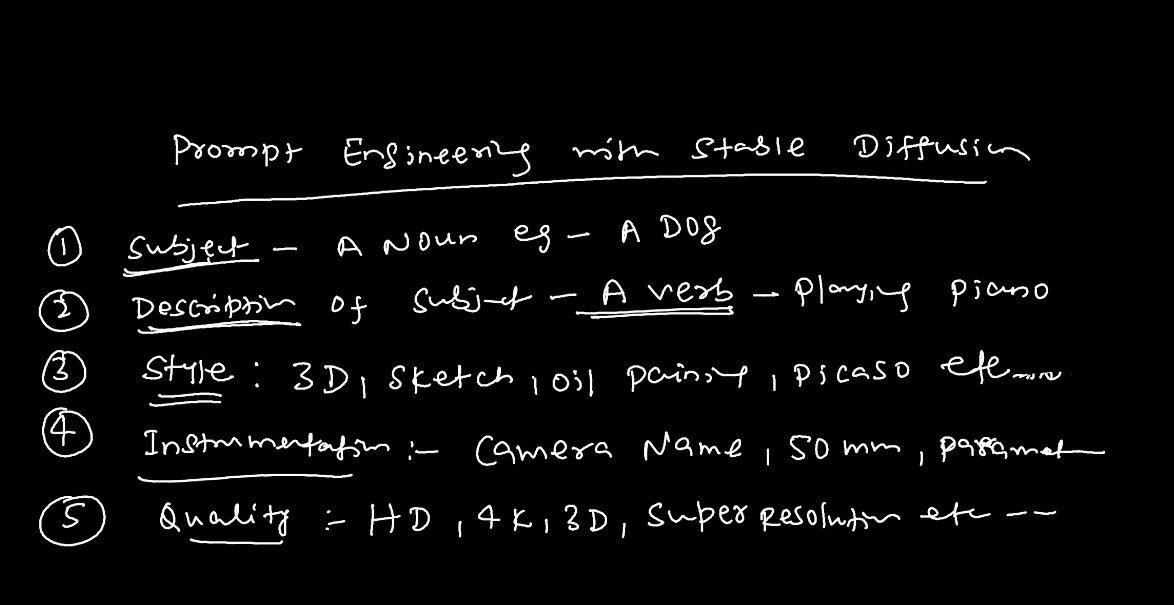
Note:- GANs generate images by image noise by SD generated them by using text noise.



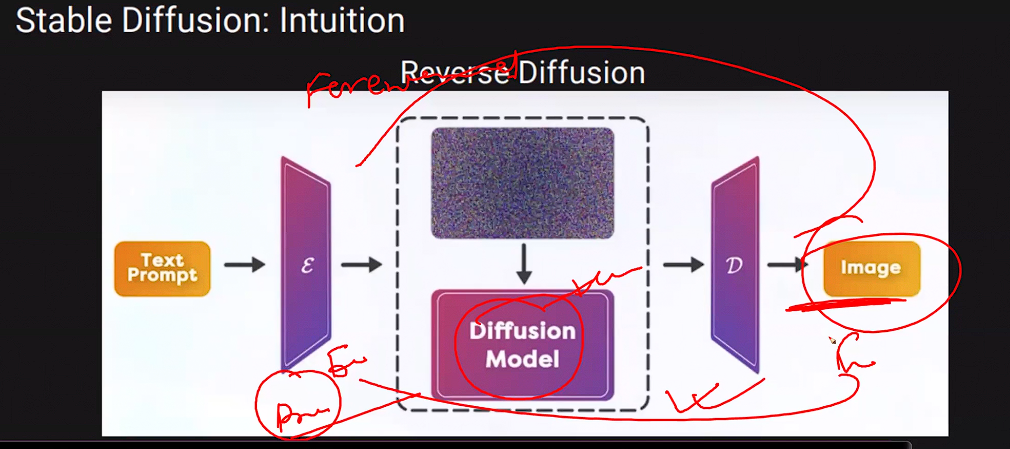
The iteration (FP + BP) keeps happening until the image pixel values becomes exactly equal to the text embedding (noise)



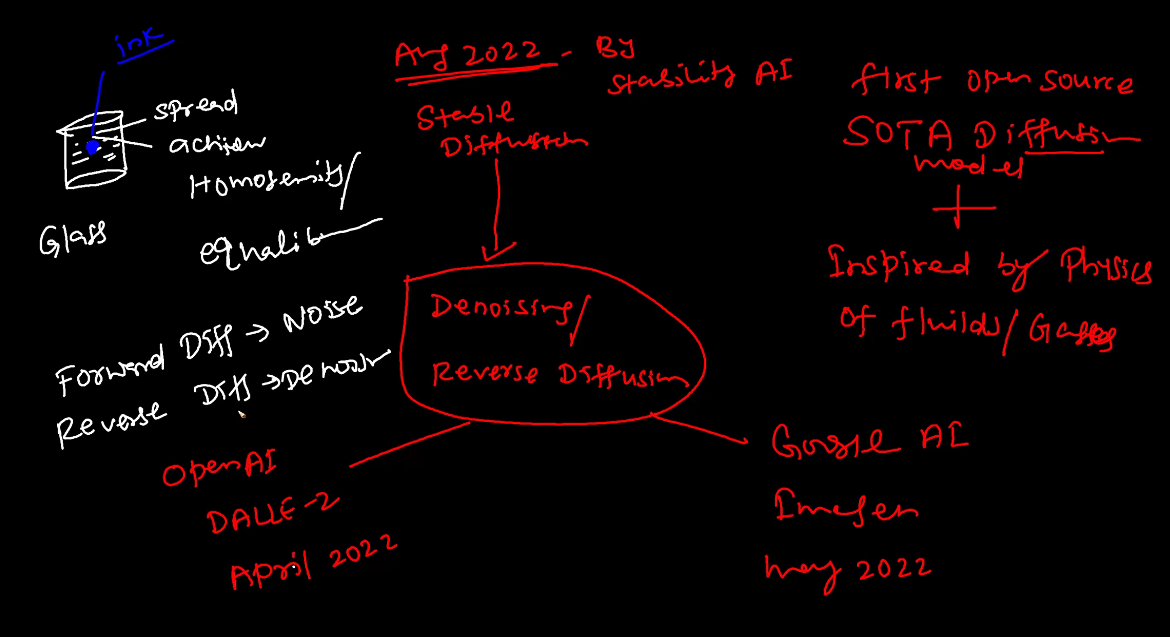




This prompt structure is very crucial to fetch the image.



Stable diffusion structure (deep dive):-



Stable diffusion which came in August 2022 is actually Denoising or reverse diffusion. It was the first open source SOTA model. It was inspired by the physics of fluids and gases. It is like spreading ink in a water which ultimately reaches equilibrium and the smell of perfume which gets spread evenly in the air.

It is actually a Neural Network in which there is already a noise data present after which we give an input prompt and in the forward propagation, the loss is calculated basis the noise embedding (pixel values) and text embedding. Then in the reverse propagation, the model de noises the data based on the given prompt until the text embedding matches with the image embedding and the image is generated.

Two things happen in the reverse propagation,

1. Calculating how much noise is left.

2. Remove the noise.

3. Repeat the process.

Stable diffusion models the process of "reverse diffusion" — going from chaos (random noise) to order (a coherent image).

**Key Detail**: The "reverse propagation" here isn’t the same as backpropagation in training. It refers to the iterative denoising process in image generation.

Summary:- Stable Diffusion is a denoising diffusion model inspired by the physics of diffusion (like ink spreading in water). It starts with random noise and removes the noise step by step, guided by a text prompt, to create an image. The model predicts how much noise is present in the noisy image and removes it iteratively until the image aligns with the text embedding, producing a meaningful result. This process is what we call reverse diffusion or denoising.

The **CLIP method** in Stable Diffusion plays a crucial role in aligning the generated image with the text prompt. Let’s break it down step by step in a simple, intuitive way.

**What is CLIP?**

CLIP stands for **Contrastive Language-Image Pretraining**. It’s a neural network developed by OpenAI that can understand and connect **text descriptions** with **images**. In simpler terms, it helps the model determine how well an image matches a given text prompt.

**Why is CLIP Important in Stable Diffusion?**

In Stable Diffusion, the text prompt ("A cat sitting on a beach") guides the generation process. But for this to work, the model needs a way to compare the meaning of the text with the content of the image. That’s where **CLIP** comes in.

CLIP does two key things:

1. **Encodes Text and Images into Embeddings**:
   * Texts and images are converted into numerical representations (embeddings) that capture their meaning.
   * For example:
     + "A cat sitting on a beach" becomes a **text embedding**.
     + The noisy image (at any step) becomes an **image embedding**.
2. **Aligns the Embeddings**:
   * CLIP compares the text embedding and the image embedding to see how well they match.
   * If the match is poor, the Stable Diffusion model adjusts the image (by removing noise) to make it closer to the text.

**How Does the CLIP Method Work in Stable Diffusion?**

1. **Input the Text Prompt**:
   * You provide a text prompt, like "A futuristic city under the stars."
2. **Generate Noise and Start Refining**:
   * The model starts with a noisy image and gradually removes noise in steps.
   * At each step, the image is evaluated by CLIP.
3. **CLIP Scores the Match**:
   * CLIP checks how similar the text embedding ("futuristic city under the stars") is to the image embedding (current noisy image).
   * If the score is low (poor match), the model adjusts the image further to make it more aligned with the text.
4. **Iterative Improvement**:
   * This process repeats over many steps. With each step, the image becomes clearer and aligns better with the prompt.

**Intuitive Example:**

Imagine you’re describing a picture to an artist:

* You say, "Draw a cat sitting under a tree."
* The artist starts with random scribbles (like random noise).
* After each attempt, you tell the artist, "Hmm, the cat doesn’t look right, and there’s no tree."
* The artist adjusts their drawing based on your feedback until the drawing matches your description.

In this analogy:

* **You** are CLIP, giving feedback on how well the image matches the text.
* **The artist** is Stable Diffusion, refining the image step by step.

**Why is CLIP Called "Contrastive"?**

The term "contrastive" comes from how CLIP was trained:

* It was shown **pairs of images and text** (e.g., an image of a dog and the text "a photo of a dog").
* It learned to associate matching pairs while contrasting them with non-matching ones (e.g., "a photo of a cat" paired with a dog image).

This allows CLIP to evaluate how closely an image matches a text description.

**Summary of CLIP in Stable Diffusion:**

* CLIP acts as a guide or "critic" that evaluates how well the current image matches the text prompt.
* It converts both text and images into embeddings and aligns them.
* It ensures that the generated image reflects the prompt accurately.

CLIP was trained with 400 million data. It is a like a Imagenet dataset.