Diffusion model is used for generating impressive images which are highly realistic . It starts from basic high density Gaussian noise which is nothing but unstructured chaotic data. This data gradually improves it till it is distinguished from random noise to create data. There are two steps in which this is achieved that is by forward diffusion and reverse diffusion.

Forward diffusion:-

In this process, noise is iteratively added to the data, transforming a clear image into something more chaotic and less recognizable. The data becomes increasingly disordered as noise is added step by step. This process is typically defined as a Markov chain, where each step adds a bit more noise, making the data less recognizable over time. The purpose of this step is to learn how a clear image can gradually become noise, which is then used in the reverse process to generate new images.

Reverse Diffusion :-

In this process, the unstructured and chaotic data, which starts as random noise, is gradually refined by removing the noise step by step. As the noise is reduced, the data becomes more structured and recognizable. This is achieved by training a neural network to predict the noise added during the forward diffusion process. By learning how to effectively denoise the data at each step, the model can generate new, highly realistic samples. It begins with random noise and iteratively refines it, transforming it into a clear, detailed image that matches the target distribution.

So basically in text to image/video synthesis at first the text is encoded step by step into frames by forward diffusion process making it more complex data then by reverse diffusion process the frames are step by step converted to the video or image to more recognizable form till we get our desired output .As in the case of our prompt "a cat riding a bicycle on Mars" is first converted into a format that guides the diffusion model. This encoding is crucial as it helps the model understand the specific elements and details required for the final image or video. Next, in the forward diffusion process, this encoded text is used to generate initial frames. Noise is added step by step, making the data increasingly complex and reflective of the detailed features described in the prompt. Underneath the hood in this step there is neural network training where it is trained to identify noise which can be removed in further steps .Once the data is sufficiently noisy and complex, the reverse diffusion process takes over. Here, the model gradually removes the noise, refining the frames incrementally. Each iteration improves the clarity and detail, transforming the chaotic noise into recognizable and structured images or video frames. Finally, after the reverse process is complete, the frames are assembled into a cohesive image or video. The result is a realistic depiction of "a cat riding a bicycle on Mars," with all the intricate details accurately rendered.

These models are used to generate impressive output because :-

Gradual Refinement:- In these models noise is removed or added in form of steps so its slow step by step moment but precise moment towards model generation.

Stable training:- In this model there is stable training which prevents it from collapsing. Generally other models collapse when such diverse outputs are expected to be generated.As this model learns gradually it can produce diverse and accurate outputs.

Flexibility:- These models have been trained of diverse inputs of text so they are more flexible as hey are conditioned to generate images.This flexibility allows it to generate specific and complex images.