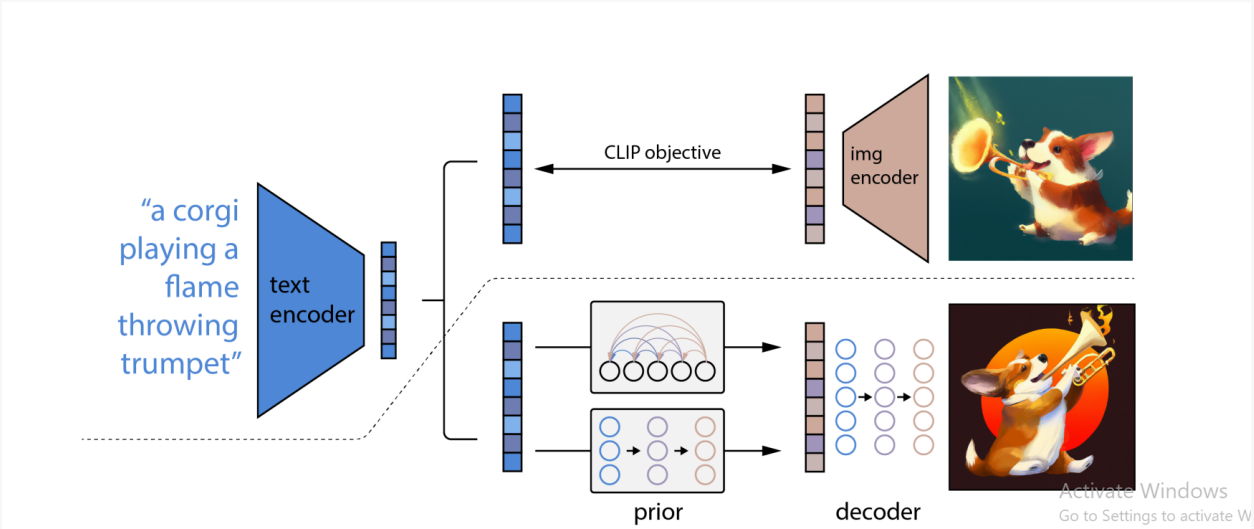
**Dall-E2**

The two-stage model you mentioned consists of two key components: a prior and a decoder

CLIP is a model that can understand both text and images and has been trained to associate text descriptions with images in a shared embedding.

The decoder uses the information contained in the image embedding to guide the generation of an image that aligns with the semantics and style specified in the input text caption.creating pixel-level details of the image.




Architecture of Dall-E2

In this work, we combine these two approaches for the problem of text-conditional image generation. We first train a diffusion decoder to invert the CLIP image encoder. Our inverter is non-deterministic, and can produce multiple images corresponding to a given image embedding..

Training Stage (Above the Dotted Line):

- CLIP Training Process: In this phase, the model undergoes training to create a joint representation space for both text and images. This means that CLIP learns to associate text descriptions with corresponding images, essentially understanding the relationships between them.

Text-to-Image Generation Stage (Below the Dotted Line):

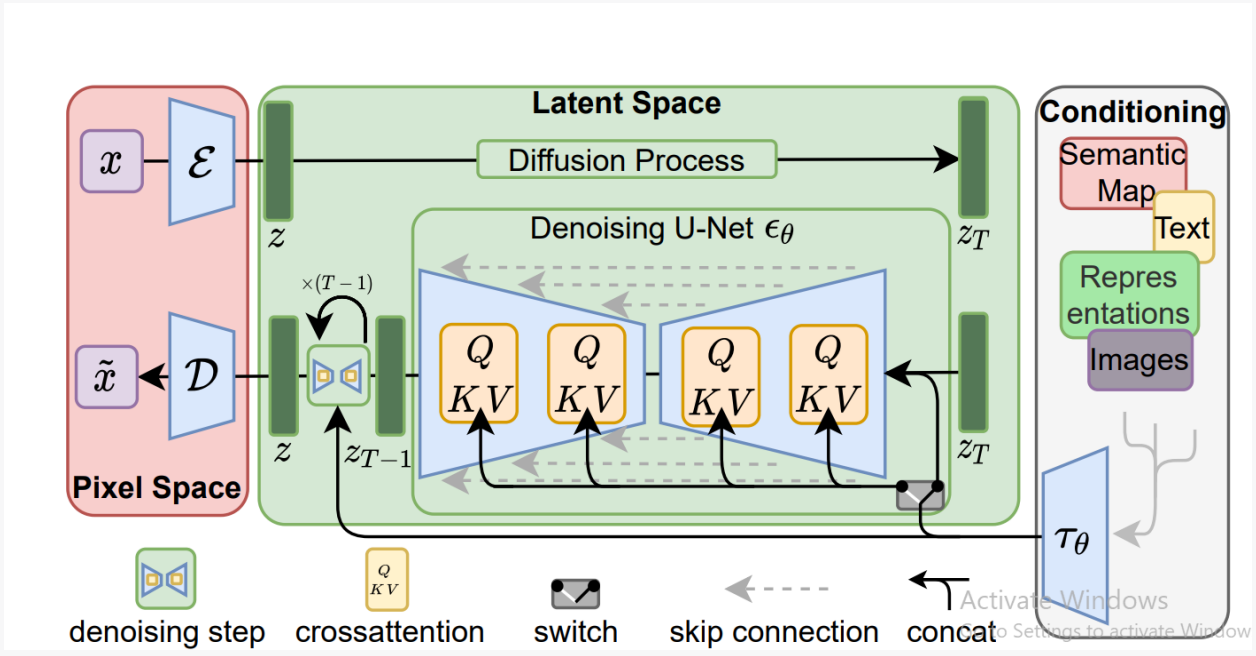
- Text Input: The process of generating images from text captions begins by providing a text caption as input.

- CLIP Text Embedding: The input text caption is converted into a CLIP text embedding. This embedding captures the meaning and style of the text description. Importantly, the CLIP model, which is pretrained and has learned this shared representation space, is used to perform this conversion.
- Prior Model (Autoregressive or Diffusion): The CLIP text embedding is then fed into a prior model. This prior model can either be autoregressive or diffusion-based. The prior's role is to generate an image embedding based on the CLIP text embedding. This image embedding is a numerical representation that encodes the information needed to create an image that aligns with the text's content and style.
- Image Embedding: The prior successfully produces the image embedding, which represents the desired image content and style guided by the input text.
- Diffusion Decoder: The image embedding is used to condition a diffusion decoder. The diffusion decoder is responsible for taking this embedding and transforming it into the final image. The decoder generates pixel-level details and overall image structure based on the information contained in the image embedding.

Variations of an input image by encoding with CLIP and then decoding with a diffusion model.The decoder allows us to invert images given their CLIP image embeddings,

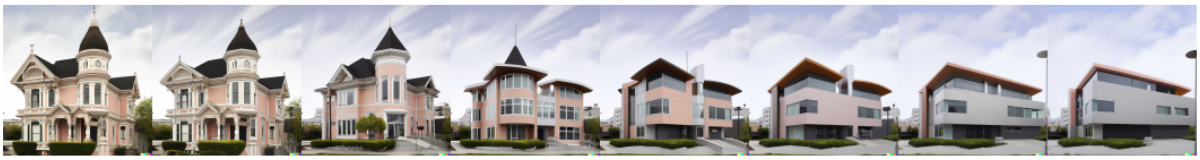
while the prior allows us to learn a generative model of the image embeddings themselves
P(x|y) = P(x, zi |y) = P(x|zi , y)P(zi |y).

Stable Diffusion




Given an image x, we can produce related images that share the same essential content but vary in other apects, such as shape and orientation (Figure 3). To do this, we apply the decoder to the bipartite representation.

a photo of a cat → an anime drawing of a super saiyan cat, artstation



a photo of a victorian house → a photo of a modern house

a photo of an adult lion → a photo of lion cub

a photo of a landscape in winter → a photo of a landscape in fall

A key advantage of using CLIP compared to other models for image representations is that it embeds images and text to the same latent space, thus allowing us to apply language-guided image manipulations (i.e., text diffs),