DALL-E 2

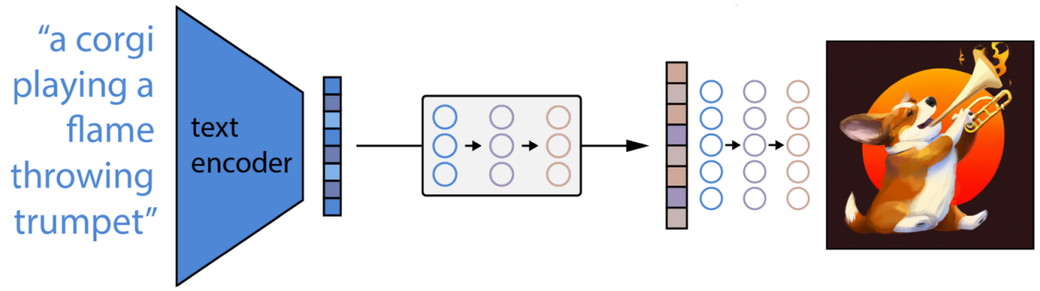
VCHS AI Club

DALL-E Under the Hood:
A Bird's Eye View

AI Club

At the highest level, DALL-E 2's works very simply:

1. A text prompt is input into a text encoder that is trained to map the prompt to a representation space.
2. A model called the prior maps the text encoding to a corresponding image encoding that captures the semantic information of the prompt contained in the text encoding.
3. An image decoder stochastically generates an image which is a visual manifestation of this semantic information.



DALL-E Under the Hood: A Detailed Look

AI Club

Step 1

Linking Textual and Visual Semantics

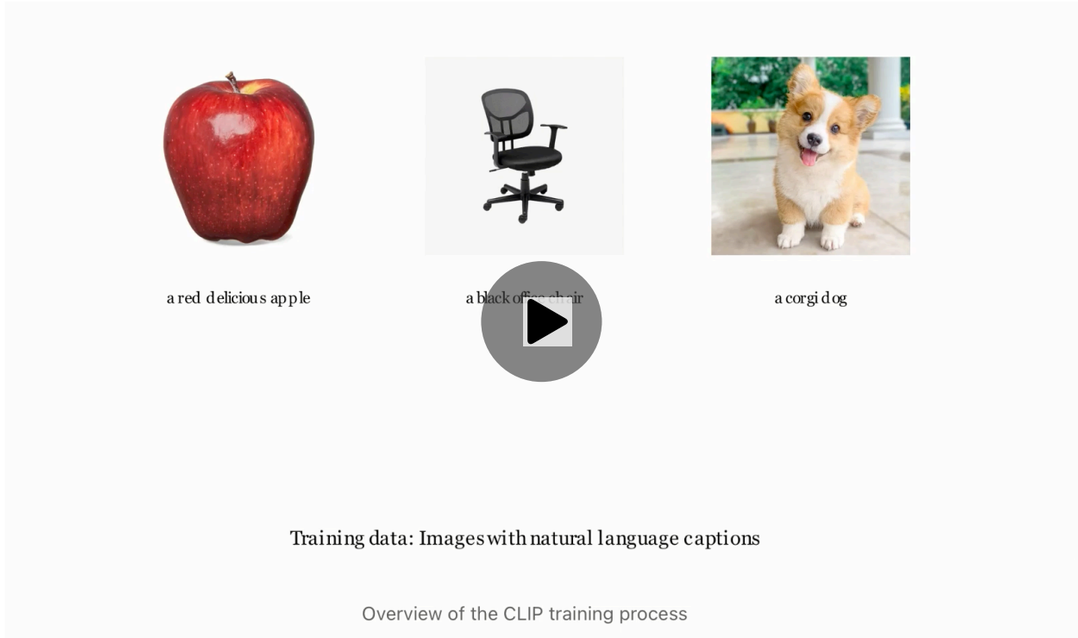
a teddy bear riding a skateboard in Times Square

AI Club

CLIP Training

The fundamental principles of training CLIP are quite simple:

1. All images and their associated captions are passed through their respective encoders, mapping all objects into an m-dimensional space.
2. The cosine similarity of each (image, text) pair is computed.
3. The training objective is to simultaneously maximize the cosine similarity between N correct encoded image/caption pairs and minimize the cosine similarity between N2 - N incorrect encoded image/caption pairs.

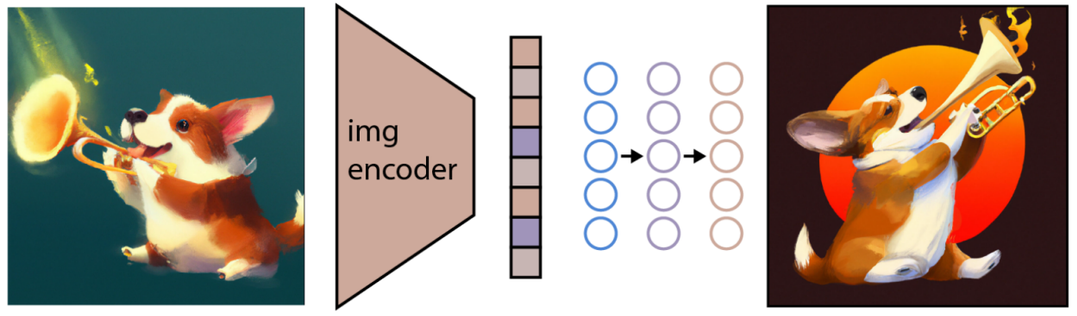


Overview of the CLIP training process

AI Club

Step 2

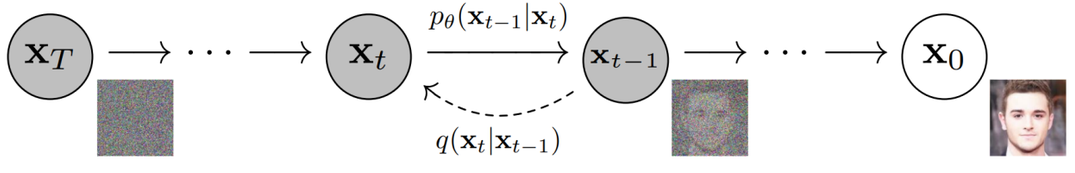
Generating Images from Visual Semantics



An image of a Corgi playing a flame-throwing trumpet passed through CLIP's image encoder.

AI Club

What is a Diffusion Model?

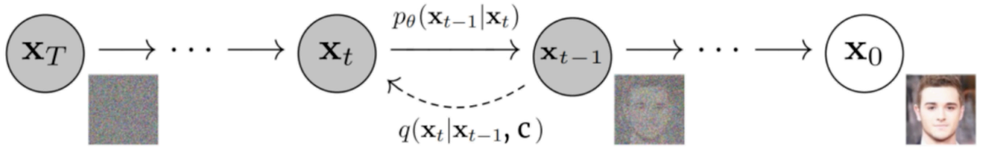


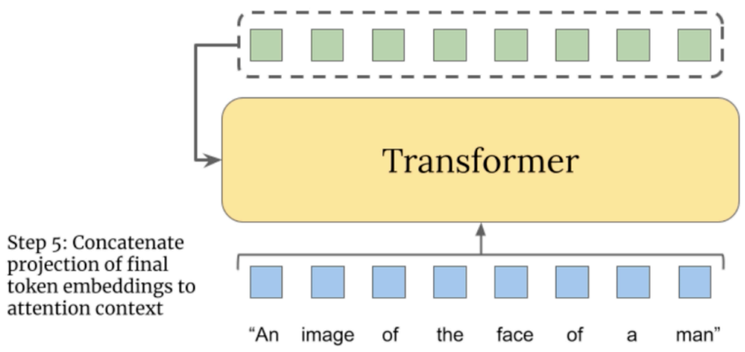
Diffusion Model schematic

AI Club

A Markov chain or Markov process is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.

GLIDE Training

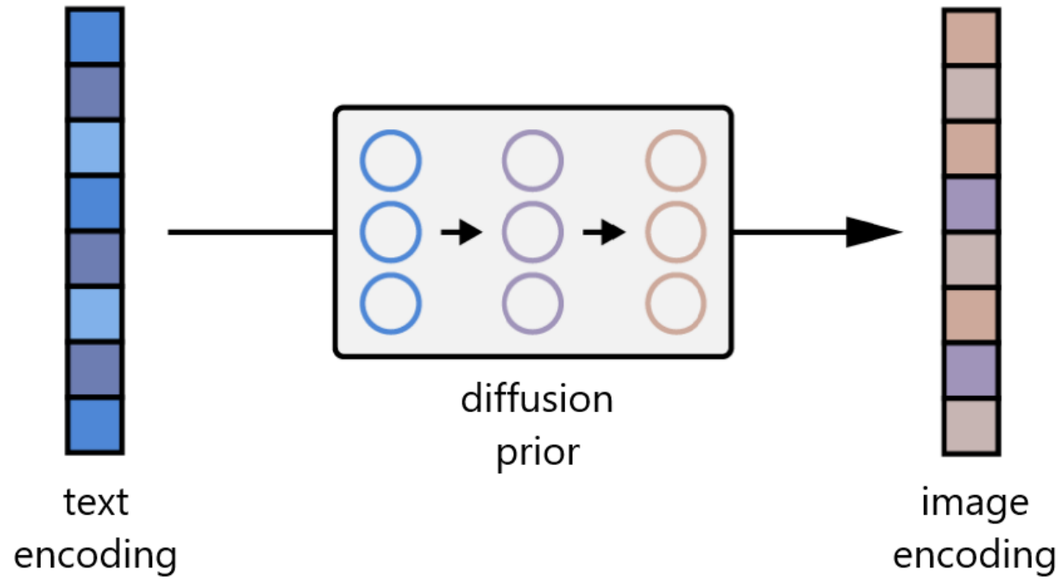




GLIDE training process.

Step 3

Mapping from Textual Semantics to Corresponding Visual Semantics



Prior Training

The Diffusion Prior in DALL-E 2 consists of a decoder-only Transformer. It operates, with a causal attention mask, on an ordered sequence of

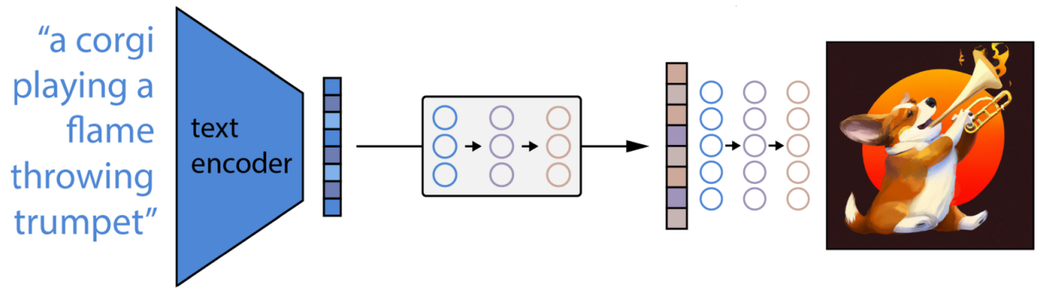
1. The tokenized text/caption.
2. The CLIP text encodings of these tokens.
3. An encoding for the diffusion time-step.
4. The noised image passed through the CLIP image encoder.
5. Final encoding whose output from Transformer is used to predict the unnoised CLIP image encoding.

Step 4

Putting It All Together

At this point, we have all of DALL-E 2's functional components and need only to chain them together for text-conditional image generation:

1. The CLIP text encoder maps the image description into the representation space.
2. The diffusion prior maps from the CLIP text encoding to a corresponding CLIP image encoding.
3. The modified-GLIDE generation model maps from the representation space into the image space via reverse-Diffusion, generating one of many possible images that conveys the semantic information within the input caption.



High-level overview of the DALL-E 2 image-generation process

AI Club