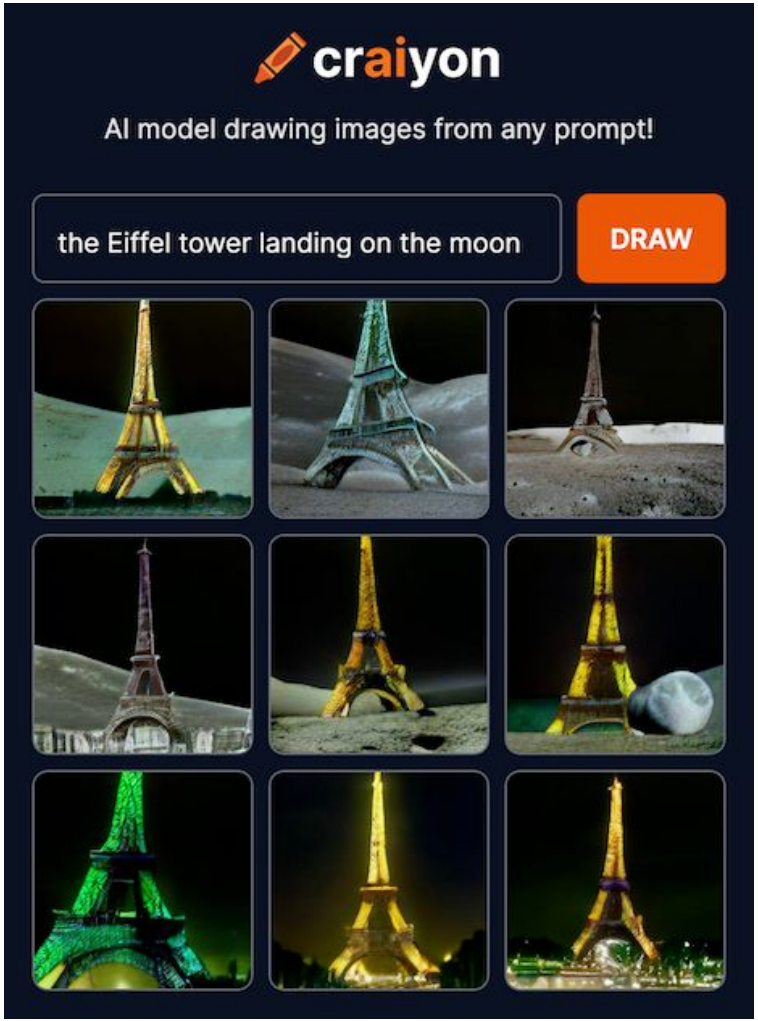
# Hands-on Session on Image Generative Models with Dall.E Mini




# Dall.E Mini — Text to Image

Online Live Version of Dall.E Mini

# Part 1: Building Blocks of Dall.E Mini

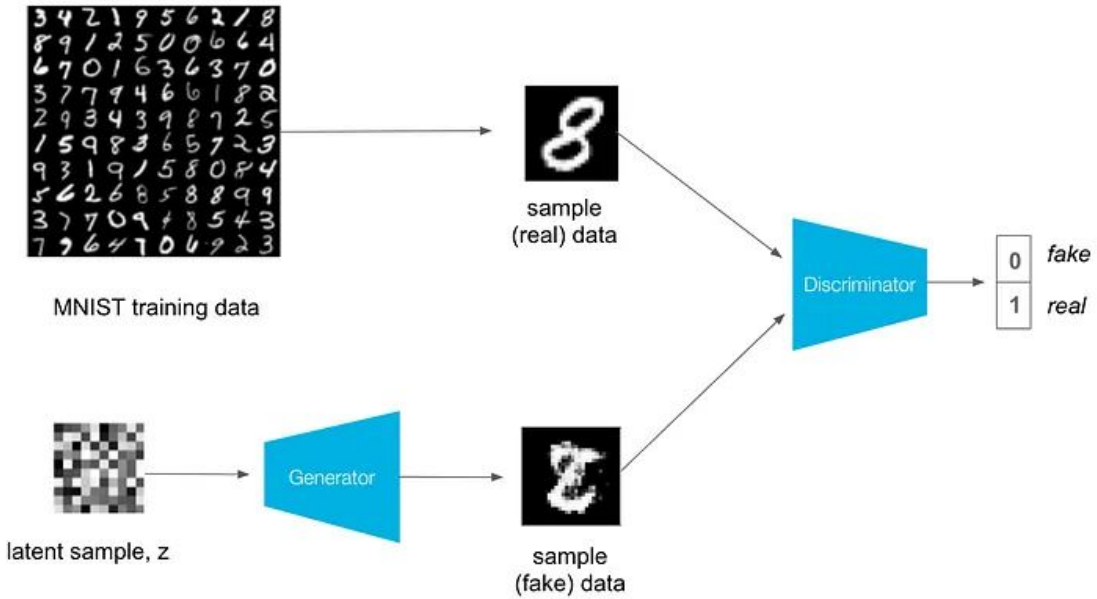
- **BERT-based Encoder-Decoder**: Encodes captions as embedding vectors
- **VQ-GAN**: Decodes caption embeddings into Images
- **CLIP**: Evaluates caption-image relevance

# Part 2: Generative Adversarial Networks (GANs)

- Dall.E Mini uses a variant of GANs called VQ-GANs.
- The evolution of VQ-GANs,
  - Simple GAN
  - Autoencoders (AEs)
  - Variational Autoencoders (VAEs)
  - Vector Quantized Autoencoders (VQ-AEs)
  - Vector Quantized GANs (VQ-GANs)

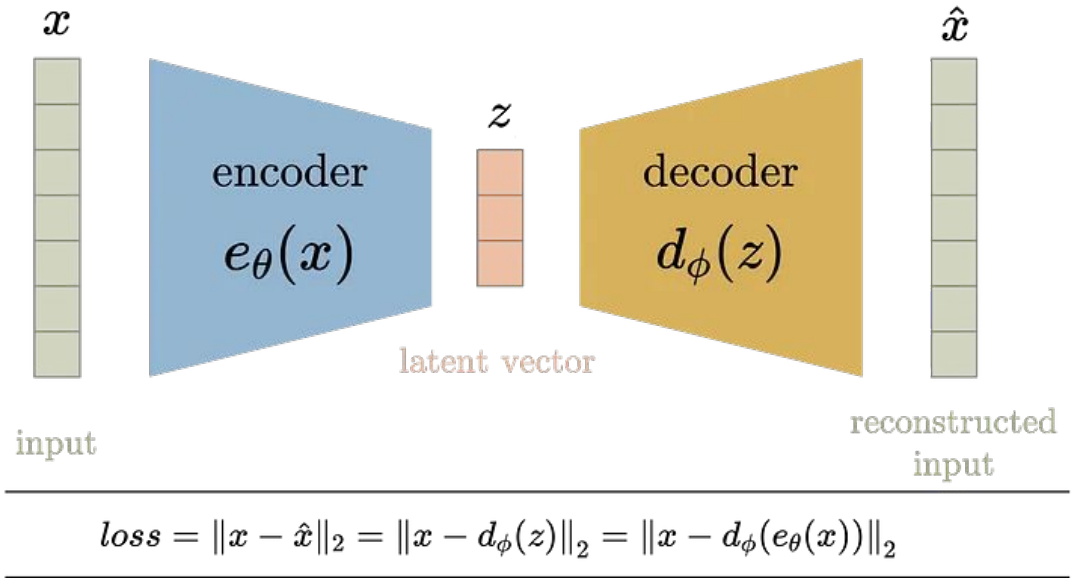
Python Code Demo

# Simple GAN



Python Code Demo

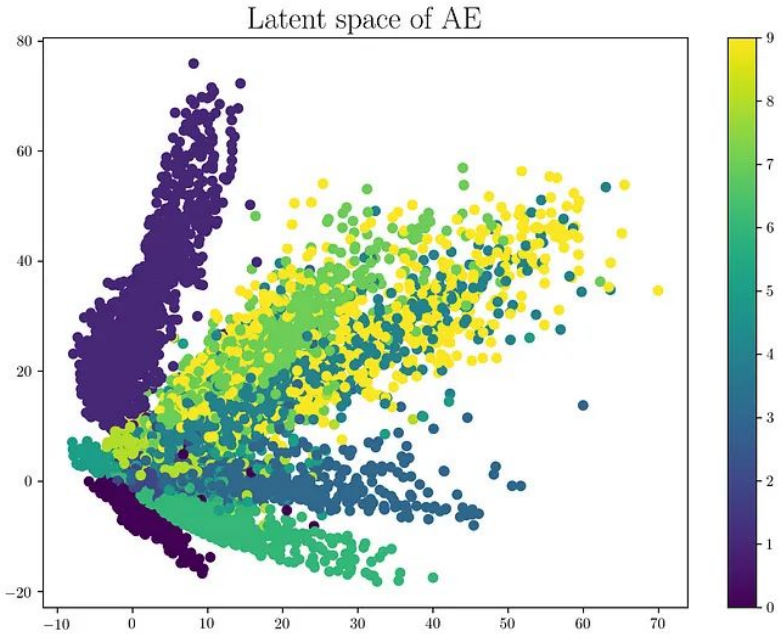
# Autoencoder (AE)




$$loss = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(e_\theta(x))\|_2$$

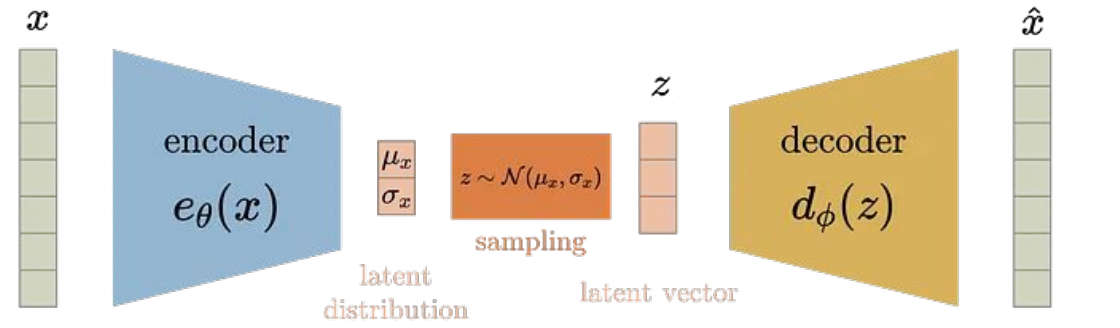
- The latent space is discontinuous and has significant “gaps”.

# Autoencoder (AE)



Python Code Demo

# Variational Autoencoder (VAE)




$$\text{reconstruction loss} = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(\mu_x + \sigma_x \epsilon)\|_2$$

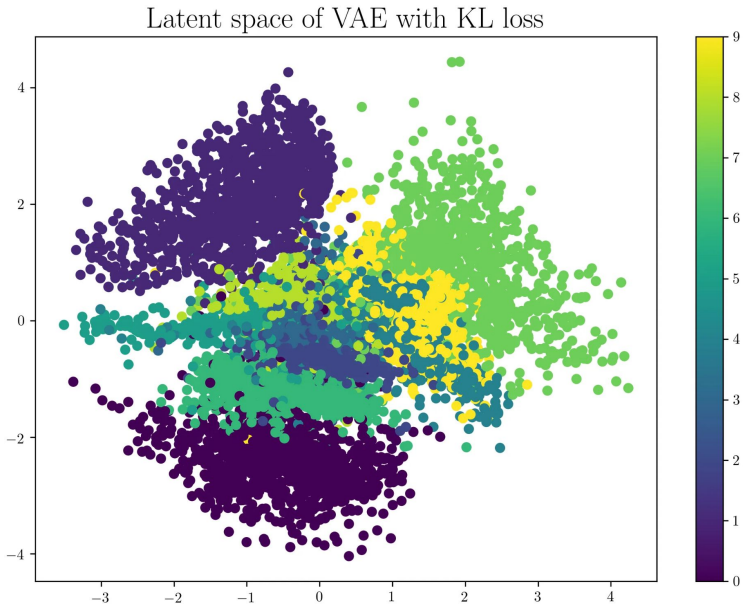
$$\mu_x, \sigma_x = e_\theta(x), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$similarity\ loss = KL\ Divergence = D_{KL}(\mathcal{N}(\mu_x, \sigma_x) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}))$$

$$loss = reconstruction\ loss + similarity\ loss$$

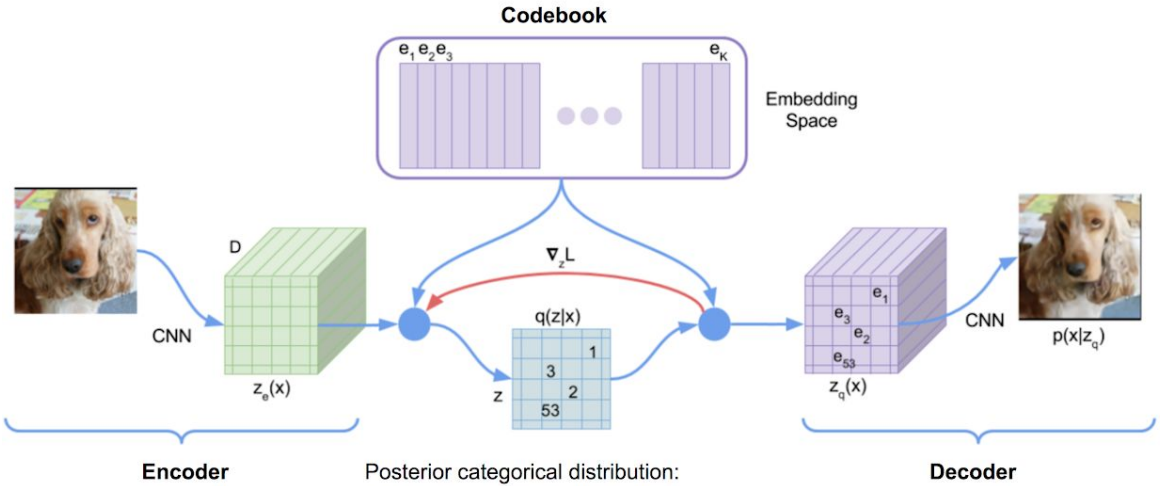
- The latent space is more cohesive — resembles the unit norm.
- Overlapping regions produce “morphed” images.

# Variational Autoencoder (VAE)



Python Code Demo

# Vector-Quantized Variational Autoencoder (VQ-VAE)

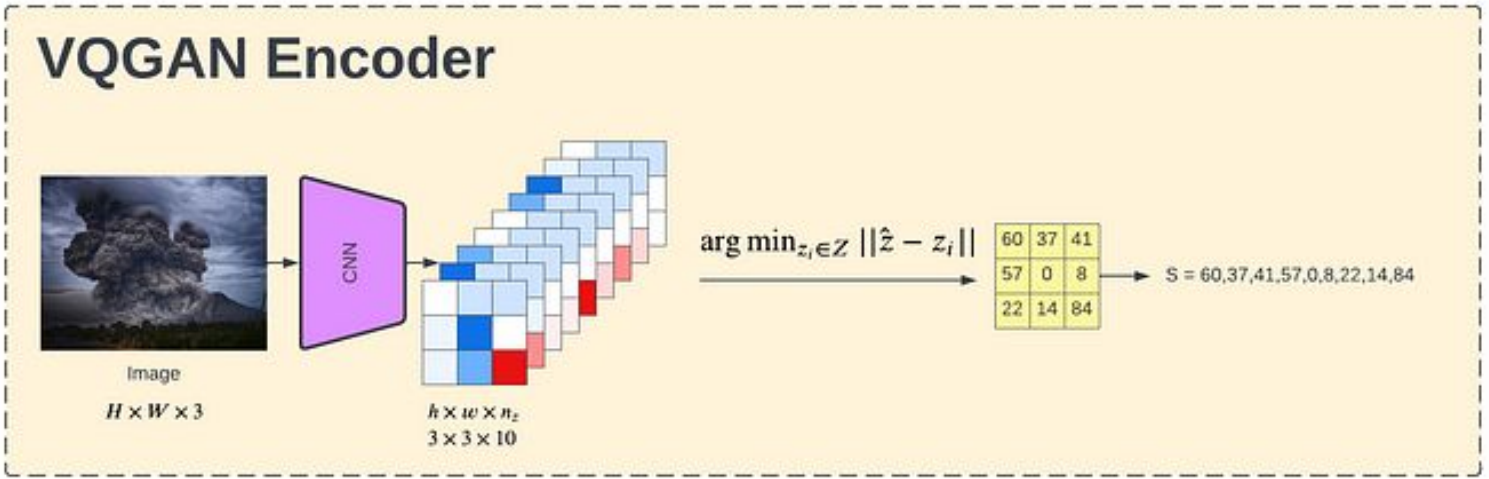



$$q(\mathbf{z} = \mathbf{e}_k|\mathbf{x}) = \begin{cases} 1 & \text{if } k = \arg\min_i \|\mathbf{z}_e(\mathbf{x}) - \mathbf{e}_i\|_2 \\ 0 & \text{otherwise.} \end{cases}$$

- The latent space is discrete.
- No “morphed” outputs.
- Latent space has same dimensions as codebook.

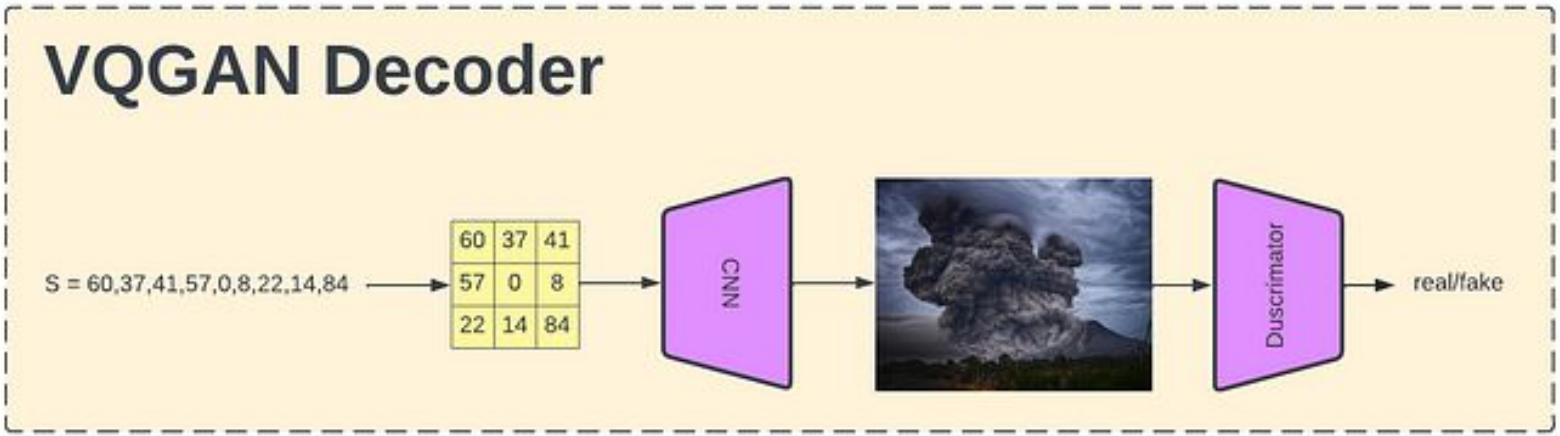
Python Code Demo

# Vector-Quantized Variational GAN (VQ-GAN)



Python Code Demo

# Vector-Quantized Variational GAN (VQ-GAN)

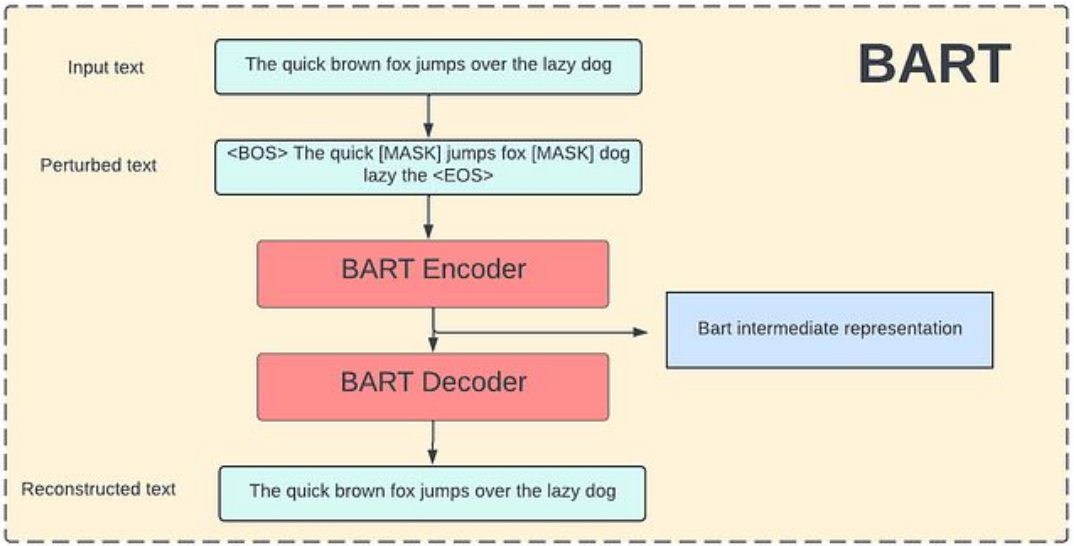


# Part 3: BART Encoder-Decoder

- A BART model is pre-trained to “clean” text captions.
- For Dall.E Mini, the BART model **translates captions into the codebook vocabulary**.
- The codebook of VQ-GAN, in effect, maps text embeddings to image embeddings.

Python Code Demo

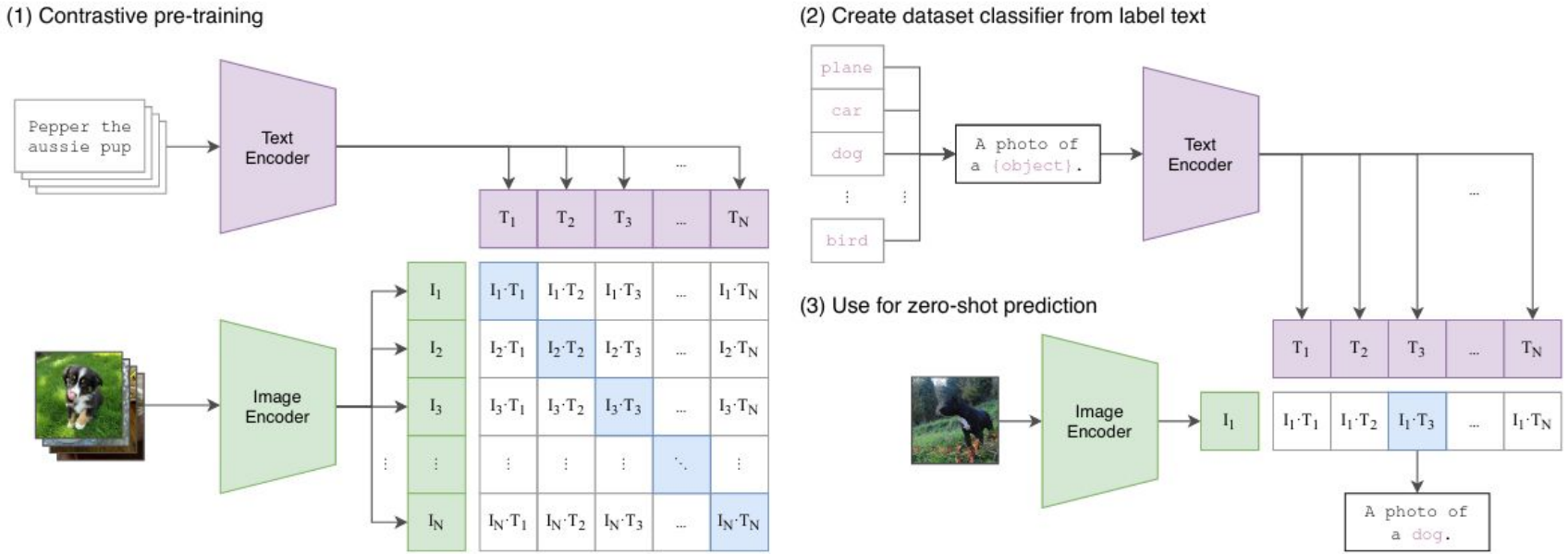
# What BART does for Dall.E.



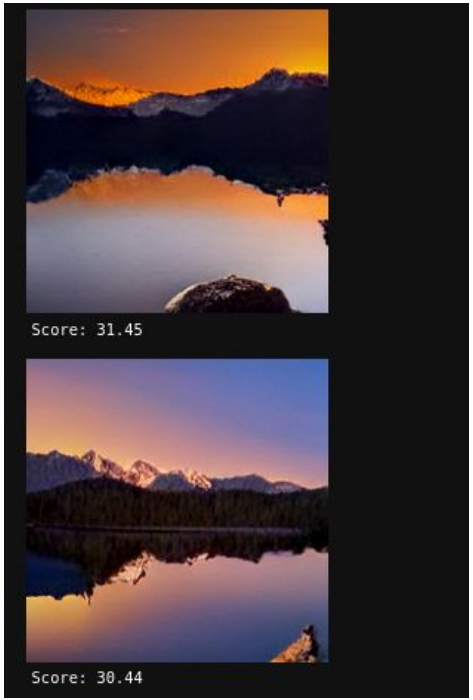

- Translates captions to codebook vocabulary.

Python Code Demo

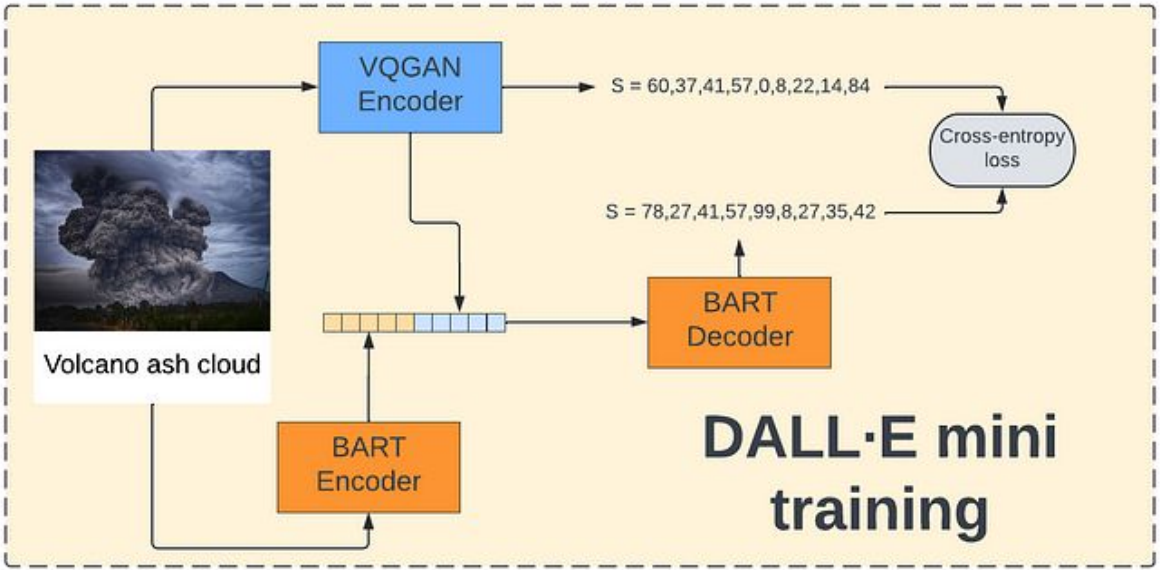
# Part 4: CLIP to Rank Images by Relevance



# Part 4: CLIP to Rank Images by Relevance

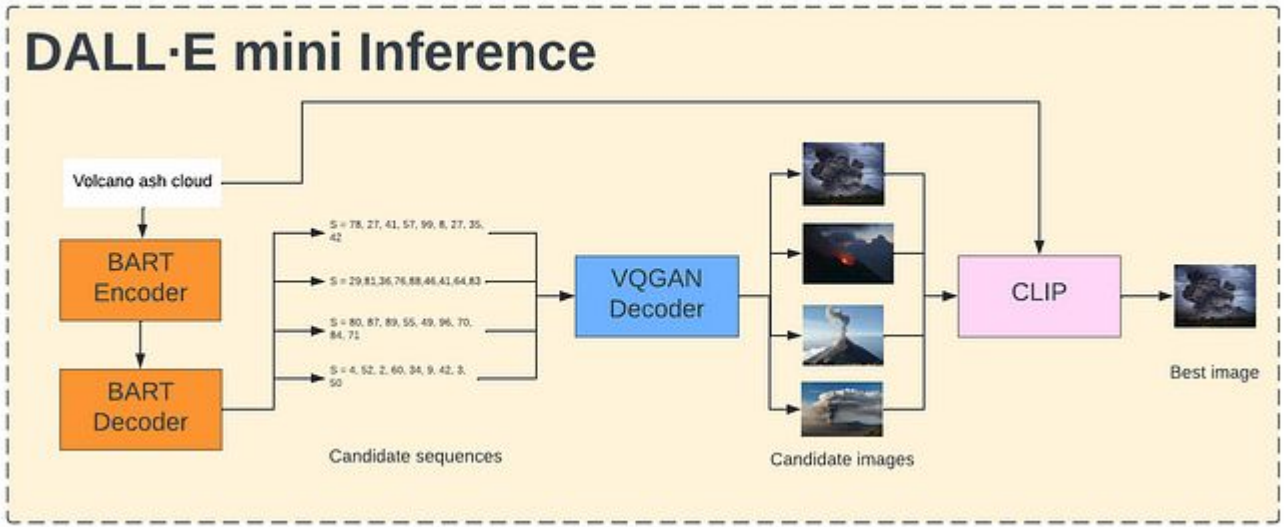



# Part 5: Putting the pieces together.



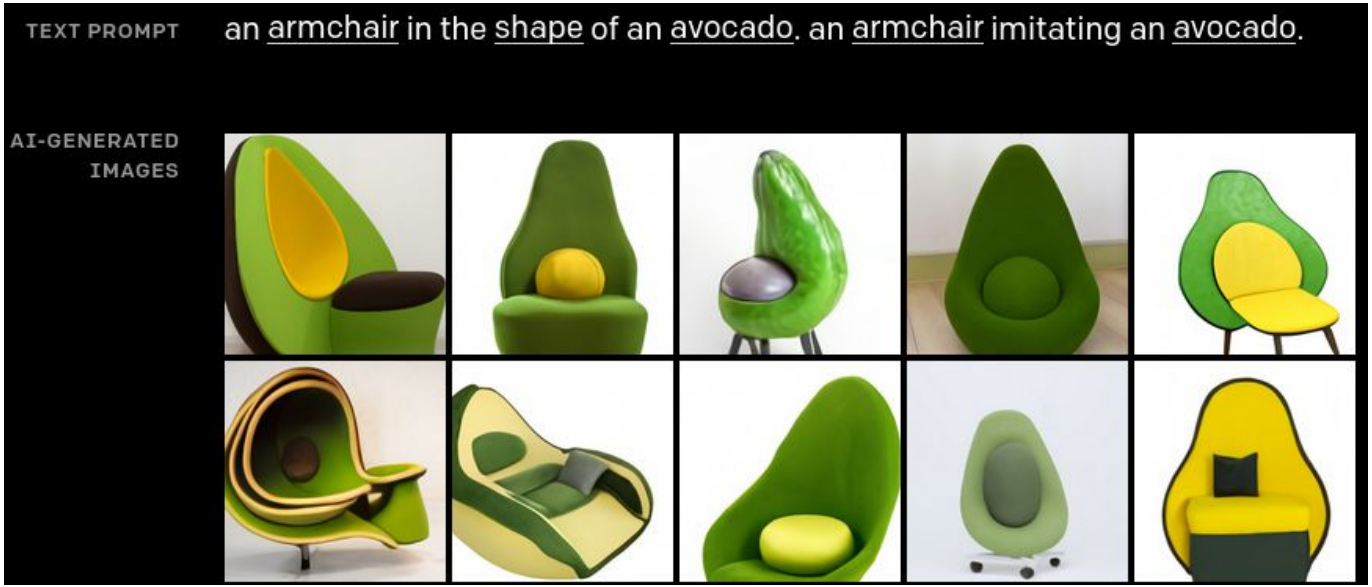
Python Code Demo

# The Dall.E Mini Text-to-Image Pipeline.



# Examples of Generated Images

# Examples of Generated Images



# Examples of Generated Images