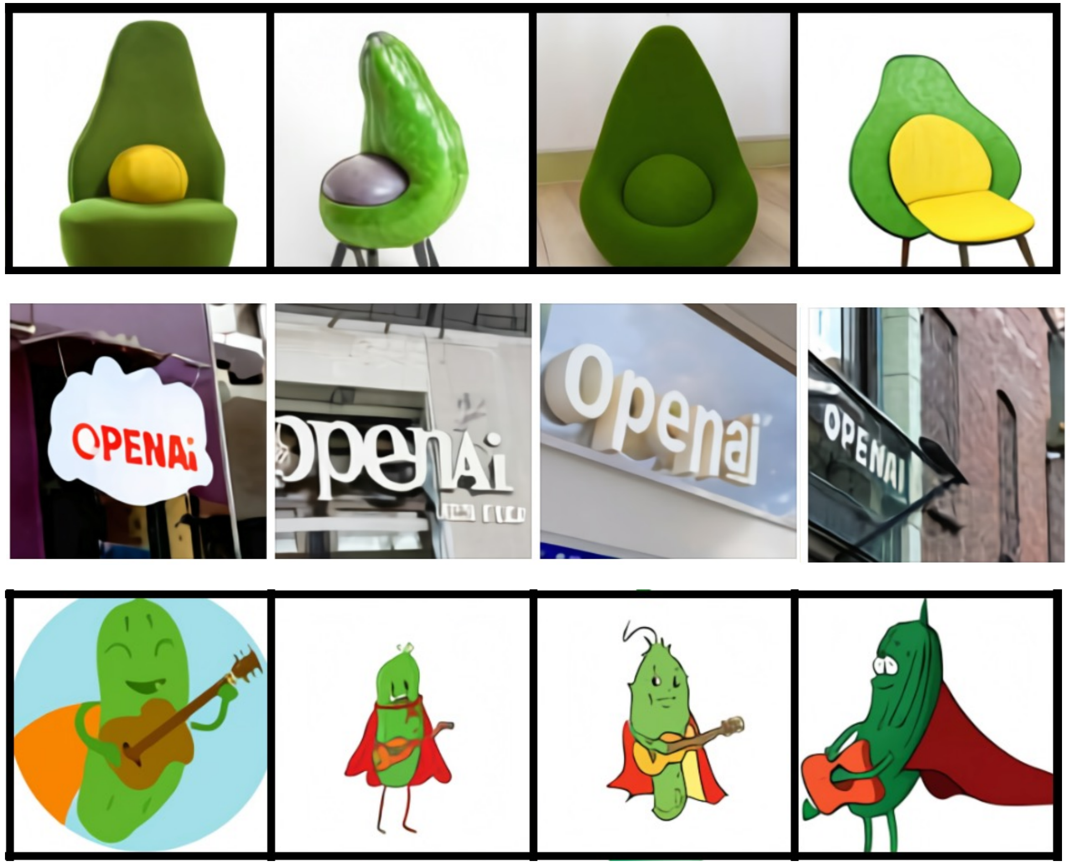
# OpenAI DALL-E

(Zero-shot Text to Image Generation[1])

30 Nov 2022

RIT

# DALL·E's Diverse Capabilities




an armchair in the shape of an avocado. — combining unrelated concepts in plausible ways

a store front that has the word 'openai' written on it — rendering text

an illustration of a baby cucumber in a cape playing a guitar. — creating anthropomorphized versions of animals and objects

**Input Text Prompt**

**Generated Images**

# Text to Image Generation

Input Text Prompt

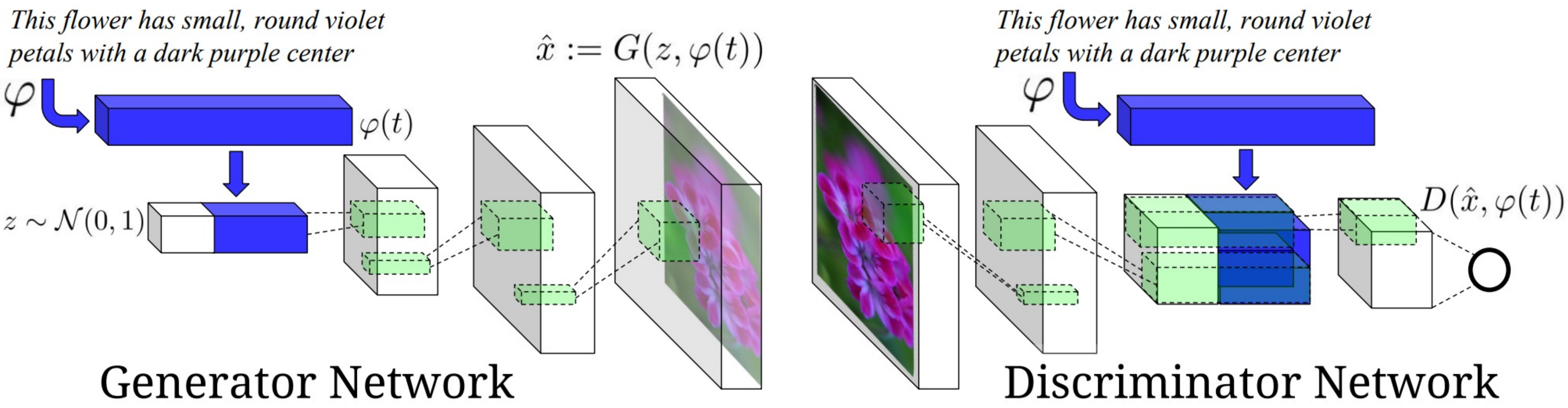
| a group of urinals is near the trees | a crowd of people standing on top of a beach. | a woman and a man standing next to a bush bench. | a bathroom with two sinks, a cabinet and a bathtub. | a man riding a bike down a street past a young man. |
|---|---|---|---|---|

Generated Images

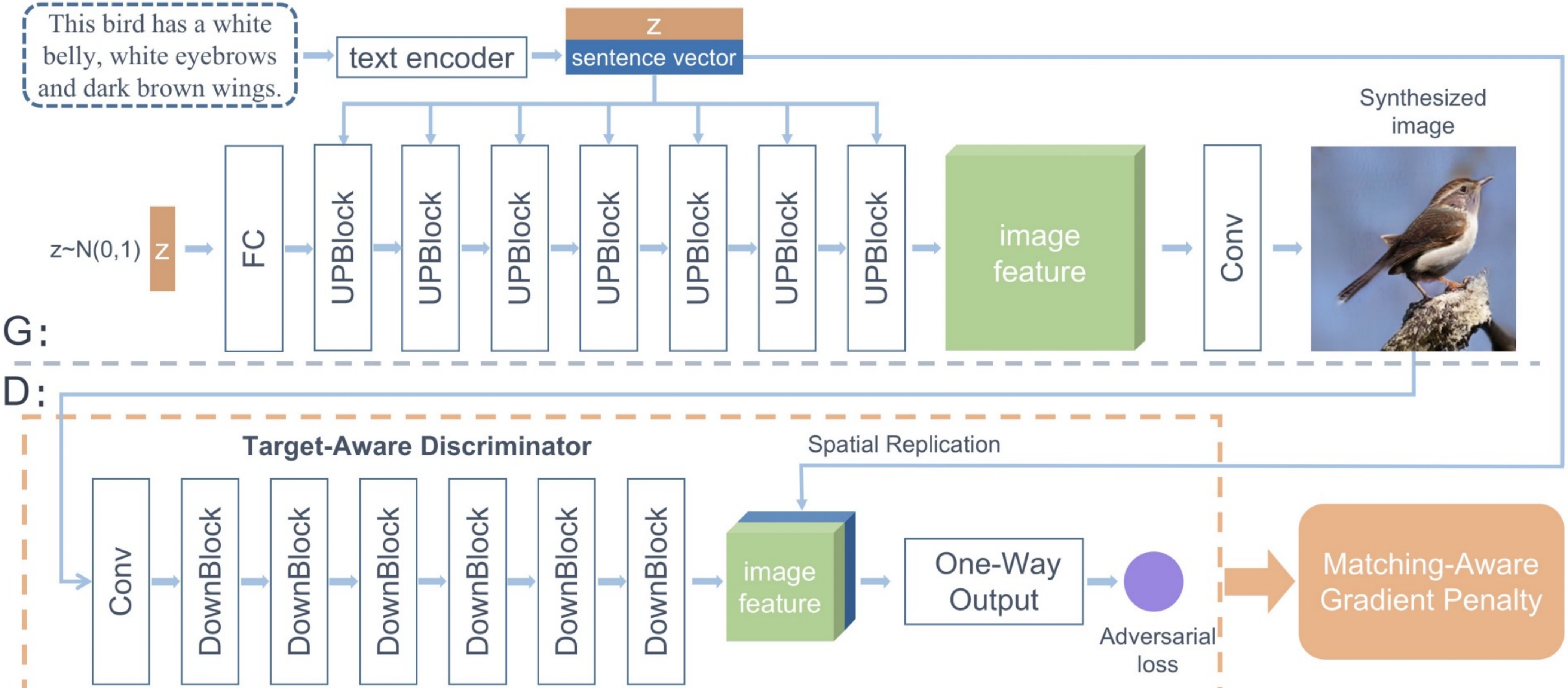
# Related Works: Text to Image Generation

- AlignDRAW (DRAW generative model + condition on image caption)
- GAN based conditional image generation – Reed et. Al.
- StackGAN 2017
- StackGAN++ 2018
- AttentionGAN
- DMGAN
- DFGAN
- TReCS – uses mouse traces

# GAN based Conditional Image Generation[1]



# Deep Fusion GAN (DF-GAN[2])




G: generator network D: discriminator network FC: fully connected layer UPBlock: upsample + residual block + DFBlock DownBlock: downsample + residual block

# DALL·E

Could scaled dataset size and model enhance performance?

with Transformer??

# Technical Details

# DALL·E

- 12B parameter transformer decoder + discrete VAE
- 250M image-text pairs

# Data Collection

- Statistics: 250M image-text pairs

- Sources
  - Google's Conceptual Captions
  - text – image pairs from Wikipedia
  - filtered subset of YFCC100M (obtained from Flicker)

- Data Overlap Test
  - Train a contrastive model
  - Sort and manual inspection for threshold to remove images

# DALL·E Modeling

Joint distribution of text $x$, image $y$ and image tokens $z$,

$$p_{\theta,\psi}(x, y, z) = p_\theta(y|x, z)\ p_\psi(x, z)$$

where,

$p_\theta(y|x, z)$ - likelihood of image $y$ given caption $x$ and image tokens $z$

$p_\psi(x, z)$ - joint distribution over caption $x$ and image tokens $z$.

# DALL·E Modeling

$$
\begin{aligned}
\log p_{\theta,\psi}(y|x) &= \log \int p_{\theta,\psi}(y,z|x)dz. \\
&= \log \int \frac{p_{\theta,\psi}(y,z|x)}{q_\phi(z|y)} q_\phi(z|y)dz \\
&= \log \mathbb{E}_{q_\phi(z|y)}\left[\frac{p_{\theta,\psi}(y,z|x)}{q_\phi(z|y)}\right] \\
&\geq \mathbb{E}_{q_\phi(z|y)}\left[\log \frac{p_\theta(y,z|x)}{q_\phi(z|y)}\right] \\
&= \mathbb{E}_{q_\phi(z|y)}\left[\log \frac{p_\theta(y|z,x)\, p_\psi(z|x)}{q_\phi(z|y)}\right] \\
&= \mathbb{E}_{q_\phi(z|y)}\left[\log p_\theta(y|z,x)\right] + \mathbb{E}_{q_\phi(z|y)}\left[\log \frac{p_\psi(z|x)}{q_\phi(z|y)}\right] \\
&= \mathbb{E}_{q_\phi(z|y)}\left[\log p_\theta(y|z,x)\right] - \mathbb{KL}\Big[q_\phi(z|y) \mid p_\psi(z|x)\Big] \quad \text{Lower Bound}
\end{aligned}
$$

# DALL·E Modeling

$\beta$-VAE

Lower Bound

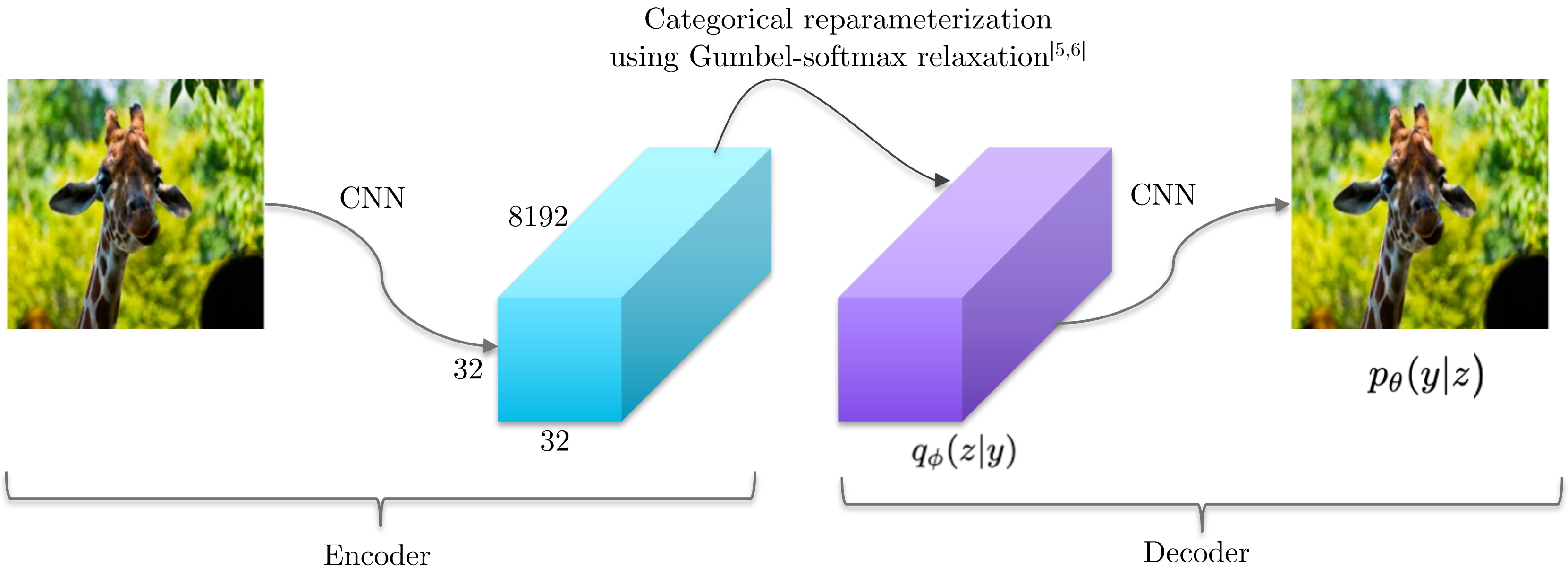
$$\mathcal{L}(\theta, \phi, \psi) = \mathbb{E}_{q_\phi(z|y)} \Big[ \log p_\theta(y|z, x) \Big] - \beta \ \mathbb{KL}\Big[ q_\phi(z|y) \mid p_\psi(z|x) \Big]$$

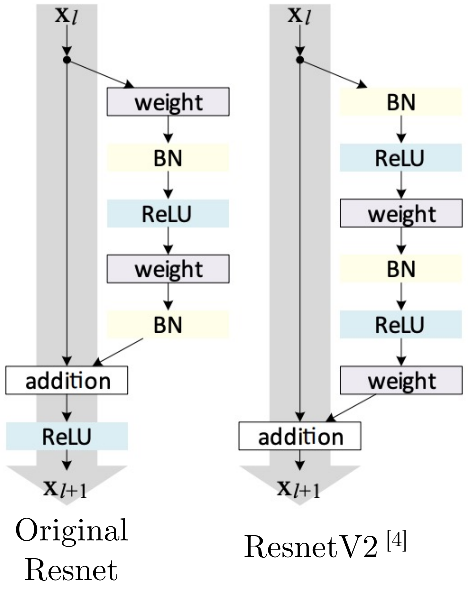
- Stage1:
  - Optimize lower bound w.r.t. $\theta$ and $\phi$
  - trains discrete VAE (dVAE) to learn image encoding

- Stage 2:
  - Optimize lower bound w.r.t. $\psi$
  - trains transformer to model conditional distribution of image tokens given text

# DALL·E Stage 1 - dVAE

- Encoder Output
  - Final embedding - 32 x 32 x 8192
  - ResnetV2 structure (improved over original Resnet)
- Decoder
  - Decoder block: ResnetV2 block + Nearest Neighbor upsampling



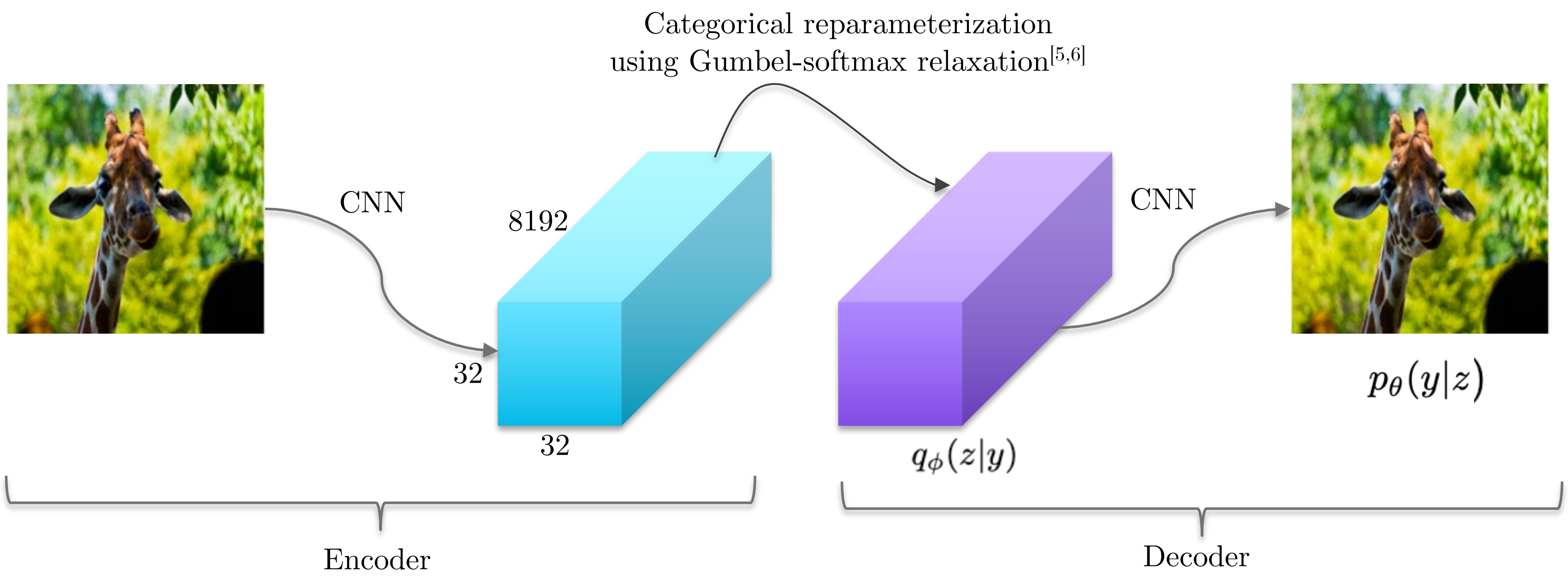



# DALL·E Stage 1 - dVAE

$$\hat{\theta}, \hat{\phi} = \underset{\theta,\phi}{\operatorname{argmax}} \; \mathcal{L}(\theta, \phi, \psi) = \mathbb{E}_{q_\phi(z|y)} \left[ \log p_\theta(y|z, x) \right] - \beta \, \mathbb{KL}\left[ q_\phi(z|y) \mid p_\psi(z|x) \right]$$

- $p_\psi(z|x)$ – uniform distribution over the K $=$ 8192 codebook vectors
- $q_\phi(z|y)$ – categorical distribution for each spatial position in 32 x 32 grid output by encoder
- $p_\theta(y|z, x)$ – evaluated using *logit-Laplace distribution*

$$f(x \mid \mu, b) = \frac{1}{2bx(1-x)} \exp\left( -\frac{|\operatorname{logit}(x) - \mu|}{b} \right)$$



# DALL·E Stage 2

$$\hat{\psi} = \underset{\psi}{\operatorname{argmax}}\ \mathcal{L}(\hat{\theta}, \hat{\phi}, \psi) = \mathbb{E}_{q_{\hat{\phi}}(z|y)}\left[\log p_{\hat{\theta}}(y|x,z)\right] - \beta\, \mathbb{KL}\left[q_{\hat{\phi}}(z|y) \mid p_{\psi}(z|x)\right]$$

- Learn prior distribution over text and image tokens
  - Text – 256 tokens (vocabulary size = 16384)
  - Image – 32 x 32 = 1024 tokens (codebook size = 8192)
- Model Architecture
  - autoregressive sparse transformer decoder (64 self-attention layers)

$$p_{\psi}(x, z) = \prod_m p_{\psi}(x_m | x_{<m}) \prod_i p_{\psi}(z_i | z_{<i}, x_i)$$

  - Masking
    - text-2-image attention – no mask
    - text-2-text attention – casual
    - image-2-image attention – row, column or convolutional attention masks
- Loss
  - Categorical cross entropy (weighted averaging for text (1/8) and image (7/8) – emphasize image modeling)

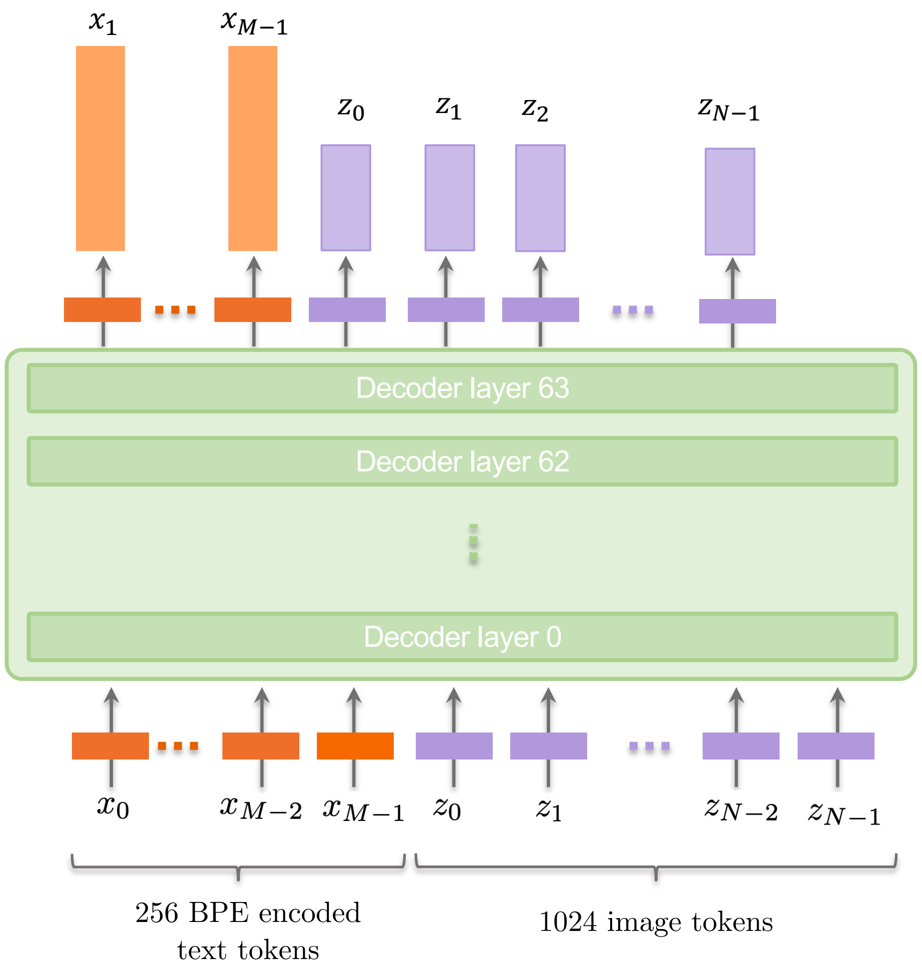
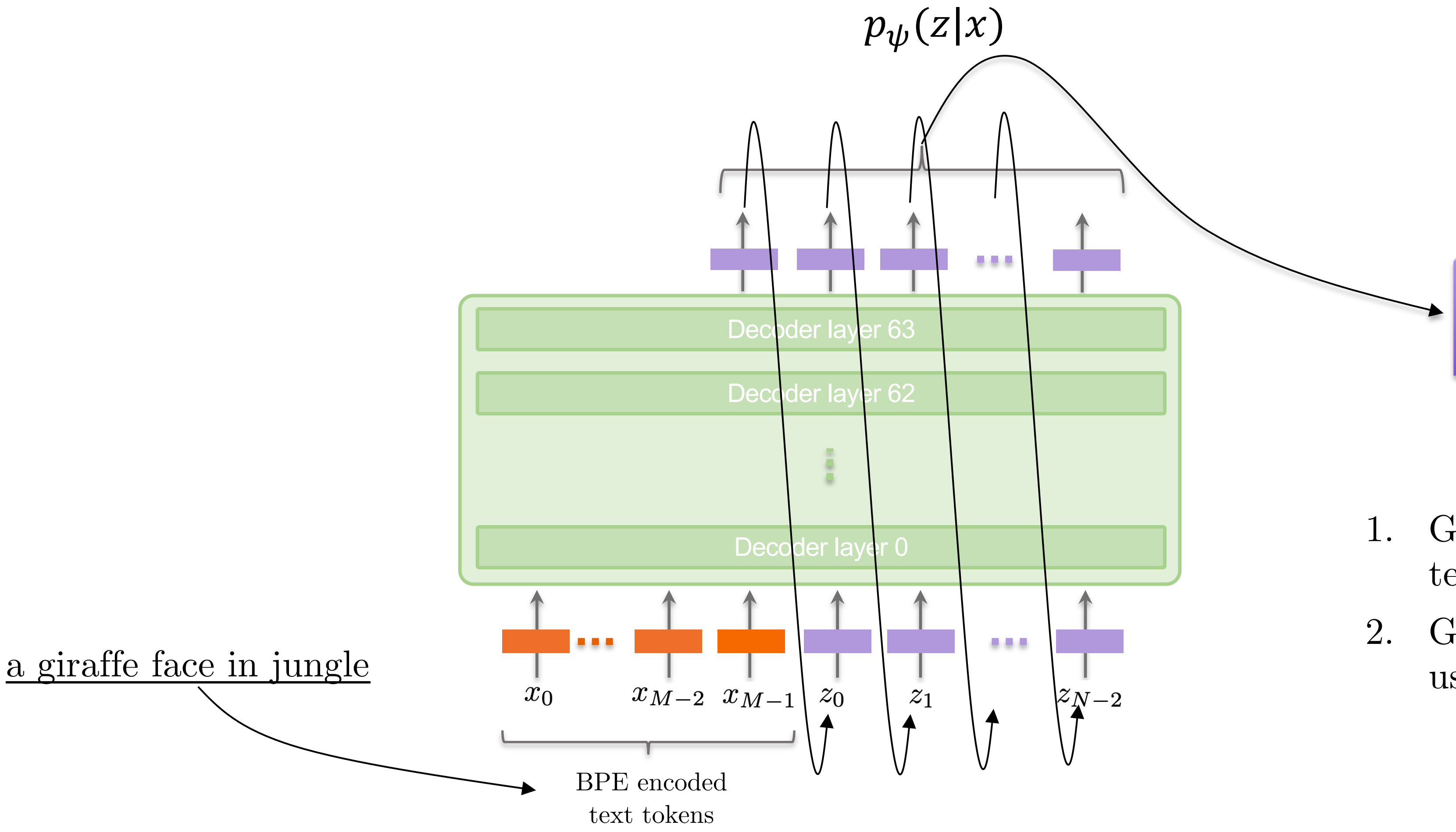
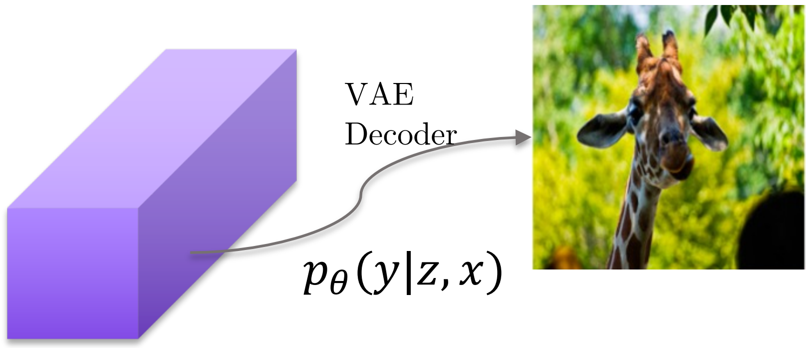



Fig: DALL-E Transformer

# Sample Generation Process




1. Generate image tokens by conditioning on text tokens using transformer
2. Generate image pixels from image tokens using VAE decoder

# Sample Evaluation

- Re-rank generated samples using a pretrained CLIP model

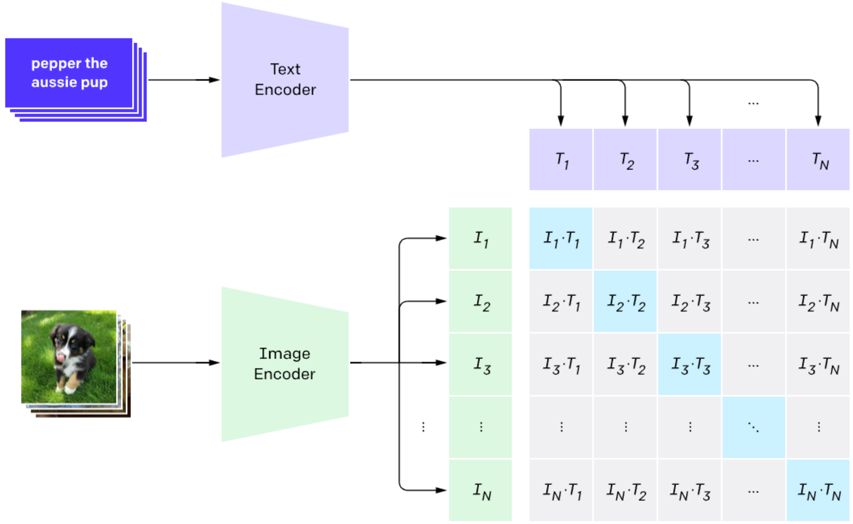



Fig: CLIP model[9]

# Evaluation

# Comparison

- DALL-E
- DF-GAN
- DM-GAN
- AttnGAN

# DALL-E v/s DFGAN Human Evaluation

- 5 independent human annotators (Amazon Mechanical Turk)

**Task:** Evaluate the two images and answer the questions below.

Image 1


Image 2

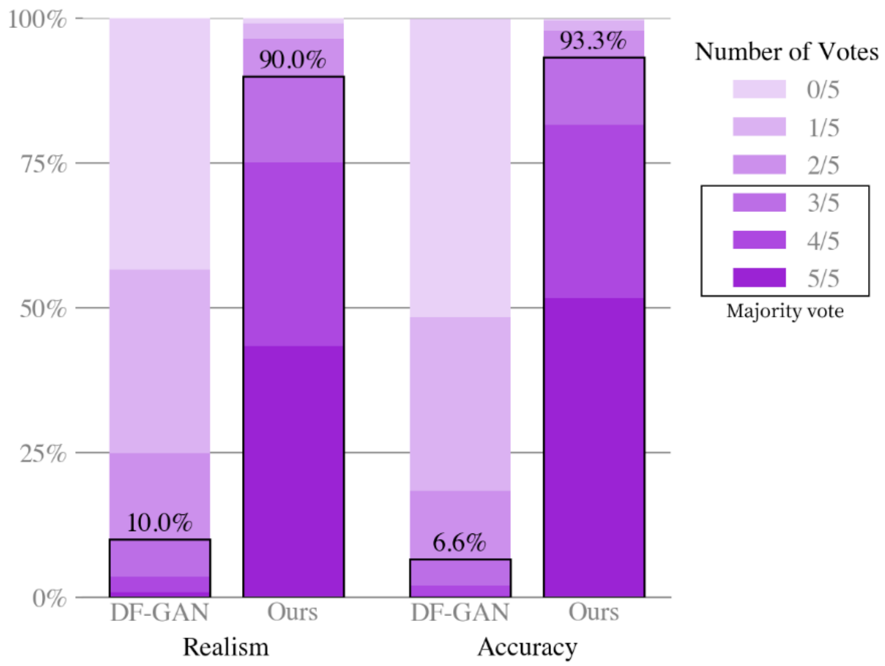
Which image is more realistic?

○ Image 1 is more realistic ○ Image 2 is more realistic

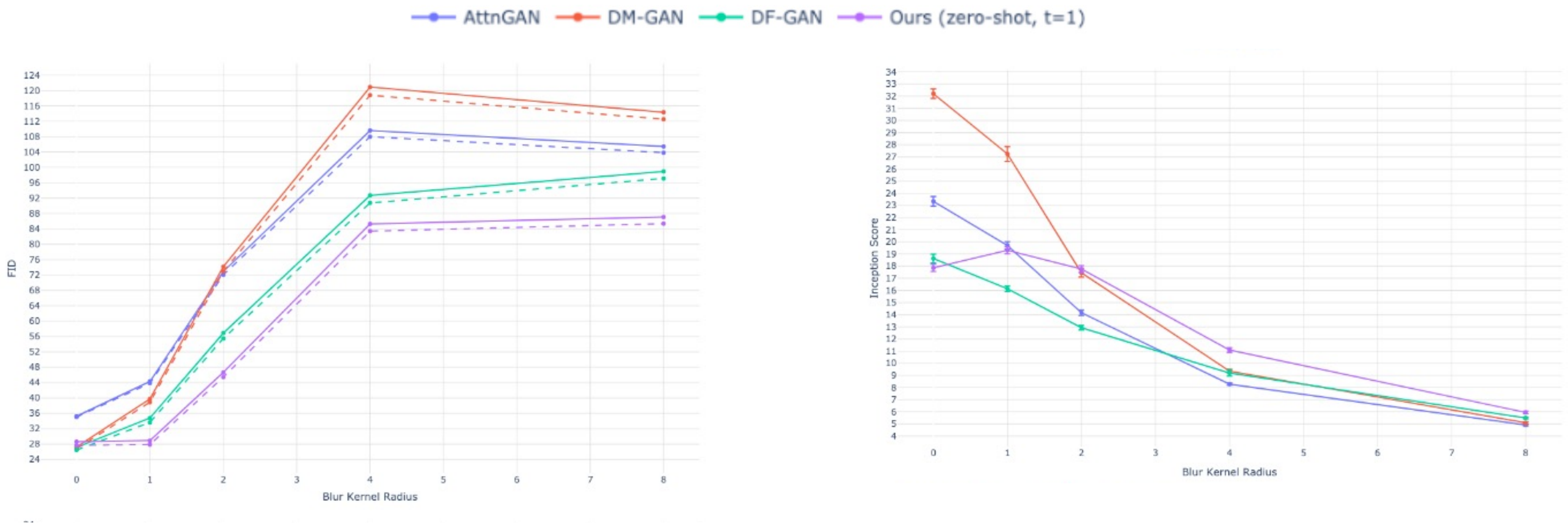
Which image matches with this caption better? **Caption:** "a man walks across a street with a stop sign in the foreground."

○ Image 1 matches better ○ Image 2 matches better ○ Neither 1 nor 2 match

**Submit**



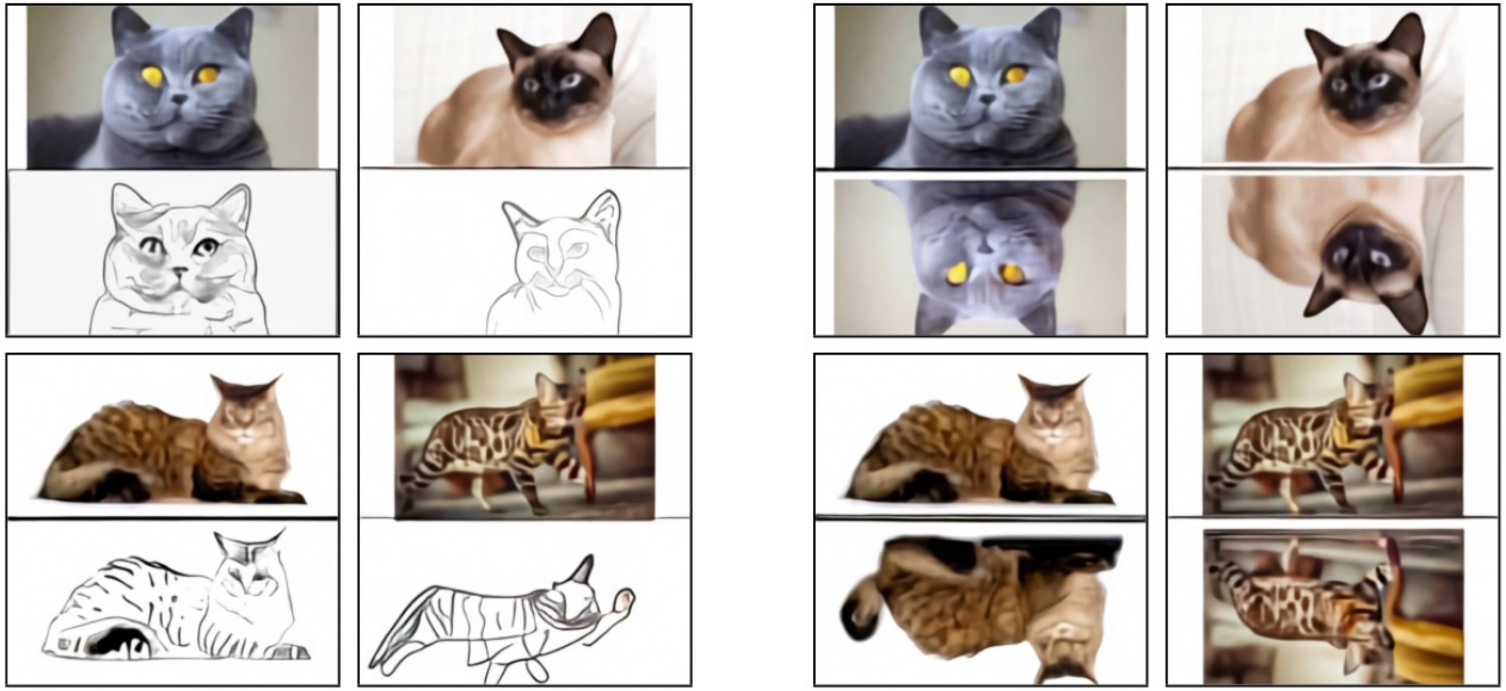
# DALL-E v/s DFGAN Comparison




(a) FID and IS on MS-COCO as a function of blur radius.

# Zero Shot Image to Image Translation

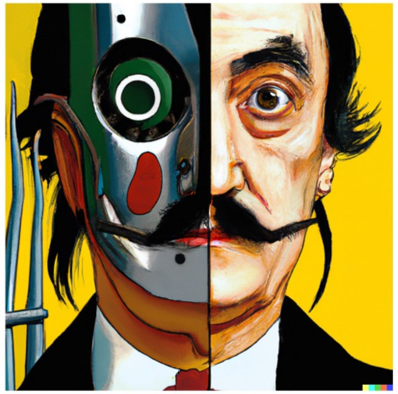
- Unanticipated and no explicit modification in training
- Emerged during test



(a) "the exact same cat on the top as a sketch on the bottom"

(b) "the exact same photo on the top reflected upside-down on the bottom"

# Recent Works



vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

DALL-E 2[8] (OpenAI)

Make-A-Video[7] (meta)

# Thank you!

# References

1. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H. (2016). Generative adversarial text to image synthesis. In ICML 2016.
2. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K., & Xu, C.. (2020). DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis.
3. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I.. (2021). Zero-Shot Text-to-Image Generation.
4. He, K., Zhang, X., Ren, S., & Sun, J.. (2016). Identity Mappings in Deep Residual Networks.
5. Jang, E., Gu, S., Poole, B. (2016). Categorical reparametrization with Gumbel-softmax.
6. Maddison, C., Mnih, A., Teh, Y. W. (2016). The Concrete distribution: a continuous relaxation of discrete random variables
7. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., & Taigman, Y.. (2022). Make-A-Video: Text-to-Video Generation without Text-Video Data.
8. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M.. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents.
9. Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I.. (2021). Learning Transferable Visual Models From Natural Language Supervision.