# Zero-Shot Text-to-Image Generation

Vladislav Filimonov

CMC MSU

April 13, 2021

# Task overview

The task is to generate images corresponding to text.

(a) an illustration of a baby hedgehog in a christmas sweater walking a dog


(b) a neon sign that reads “backprop”. a neon sign that reads “backprop”. backprop neon sign

# Dataset

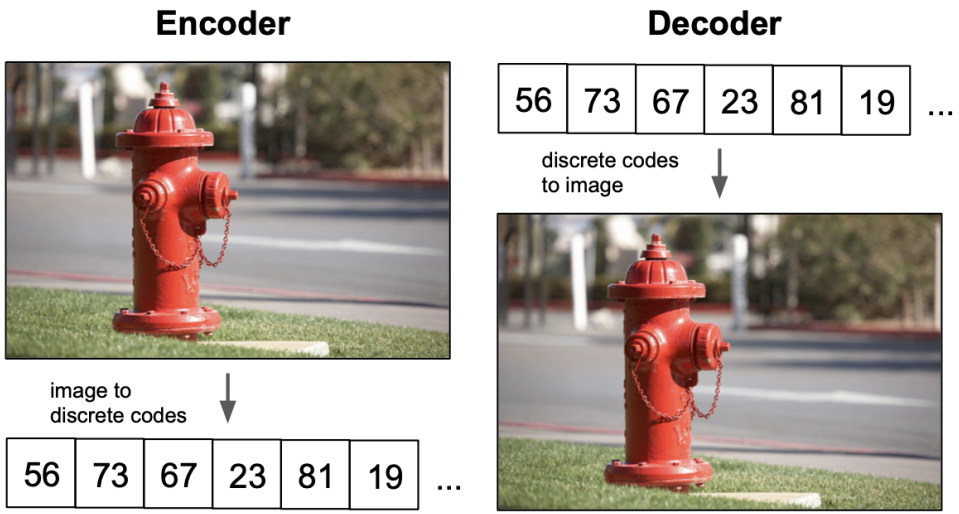
Dataset of 250 million text-image pairs was collected from:

- Conceptual Captions dataset [1] (3.30 million).
- Filtered subset of YFCC100M with technique from [1].
- Crawled data with filtering with technique from [1].

Dataset included part of MS-COCO validation images, but no captures (important for zero-shot experiments)

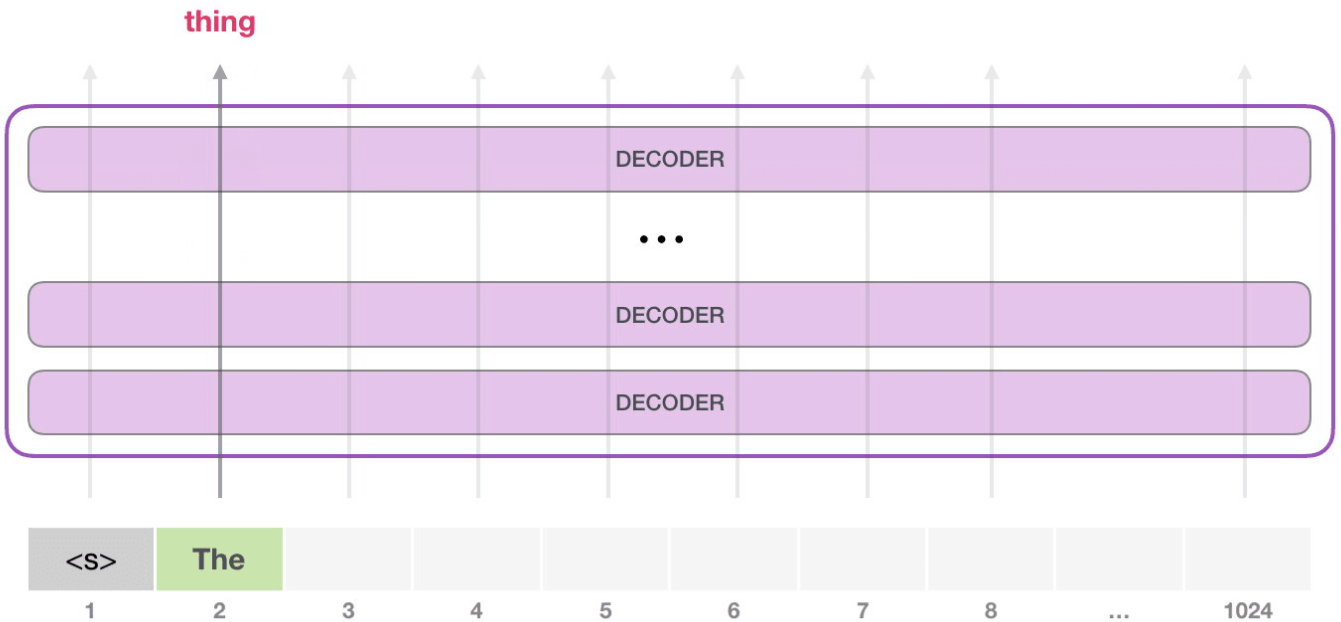
# Model Concept

1. Train Discrete VAE model to map images to a set of discrete latent variables (image tokens).



# Model Concept

1. Train Discrete VAE model to map images to a set of discrete latent variables (image tokens).
2. Train Transformer Decoder model to generate text and image tokens (extracted from DVAE) in language modelling setup.

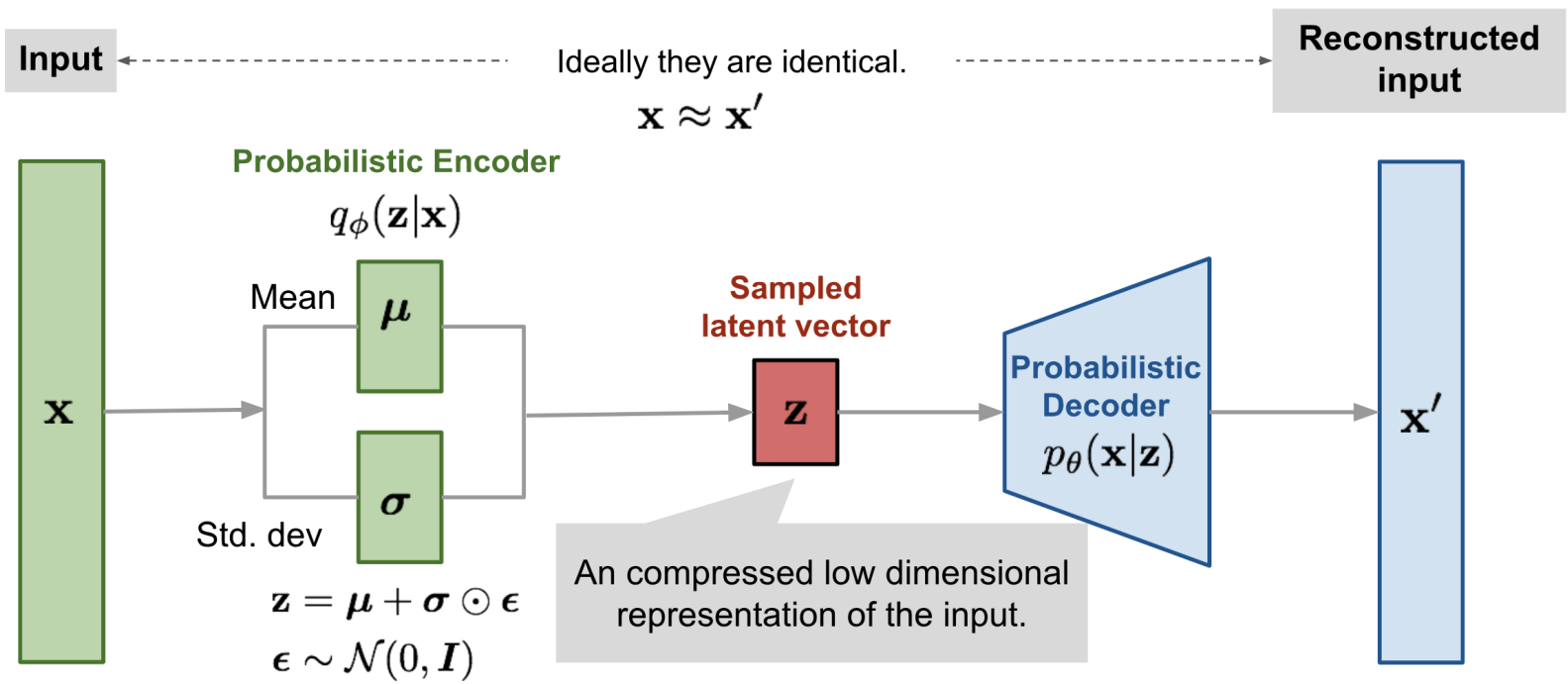



3. Generate image tokens with transformer given text, decode image tokens with decoder of DVAE.

# VAE

VAE model is trained by optimizing ELBO (using Reparametrization Trick or Reinforce):

$$\ln p_\theta(x) \geq \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)} \left( \ln p_\theta(x|z) - D_{\mathrm{KL}}(q_\phi(z|x)||q(z)) \right) \tag{1}$$



# Relaxation of Categorical

## Theorem

*(Gumbel Max Trick[a]) Let* $\pi_1, \ldots, \pi_K$: $\sum_i \pi_i = 1,\ \pi_i \geq 1, \forall i$. *Let*

$$Z = \text{argmax}_k\{\pi_k + G_k\},$$

*where* $G_k$ *is i.i.d* $\sim\ \text{Gumbel}(0, 1)$ *(*$G_k \sim -\log -\log U(0, 1)$*).*
*Then* $P(Z = k) = \pi_k$.

[a] Proof link

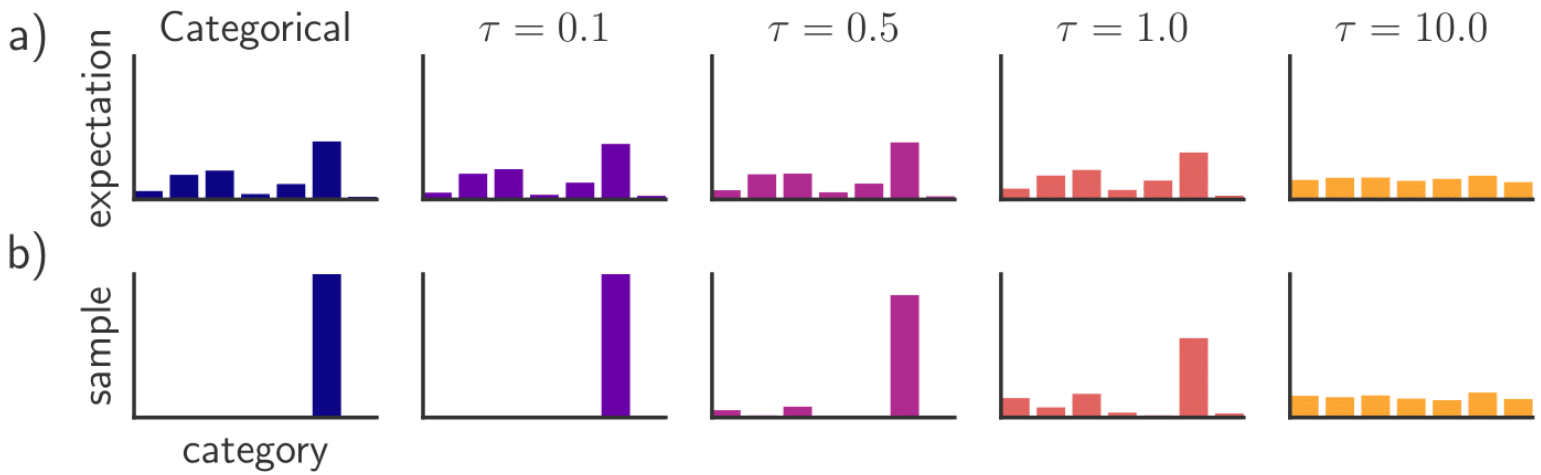
Define continuous relaxation with parameter $\tau$ (temperature):

$$y_i = \frac{\exp((\log(\pi_i) + G_i)/\tau)}{\sum_{j=1}^{K} \exp((\log(\pi_j) + G_j)/\tau)} \quad \text{for } i = 1, ..., K. \tag{2}$$

# Relaxation of Categorical

Properties[1] of relaxation (2):

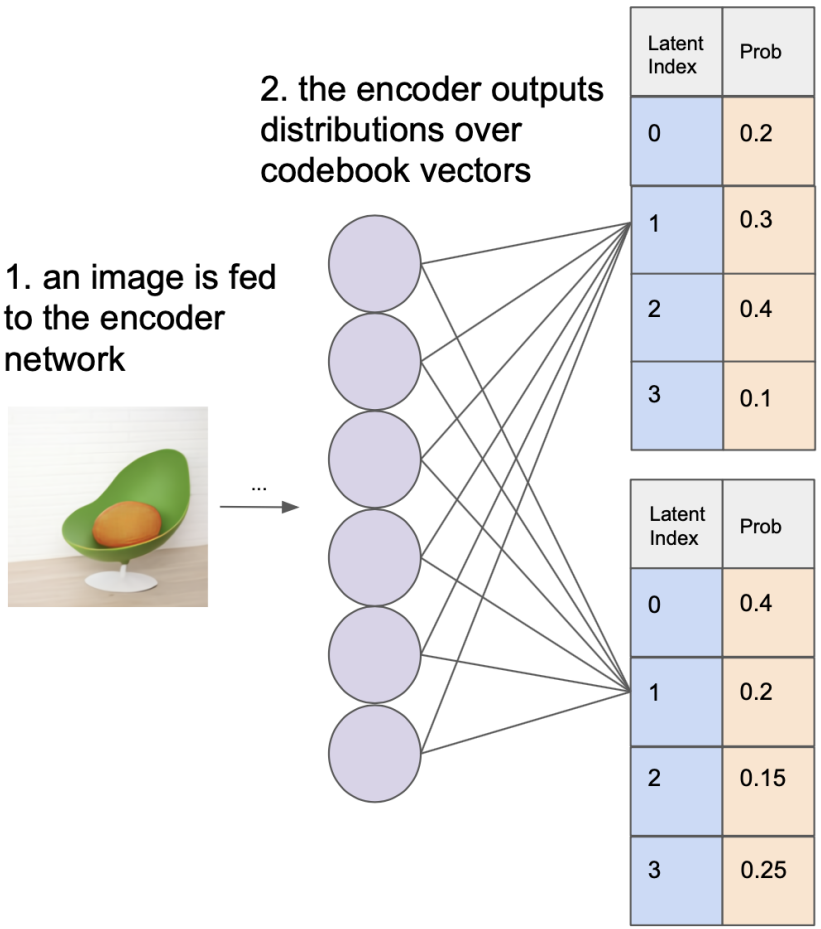
- $P(y_j > y_i) = \frac{\pi_j}{\sum_k \pi_k}, \ \forall i \neq j.$
- $P(\lim_{\tau \to 0} y_j = 1) = \frac{\pi_j}{\sum_k \pi_k}$




[1] Proof link

# DVAE in Dall-E

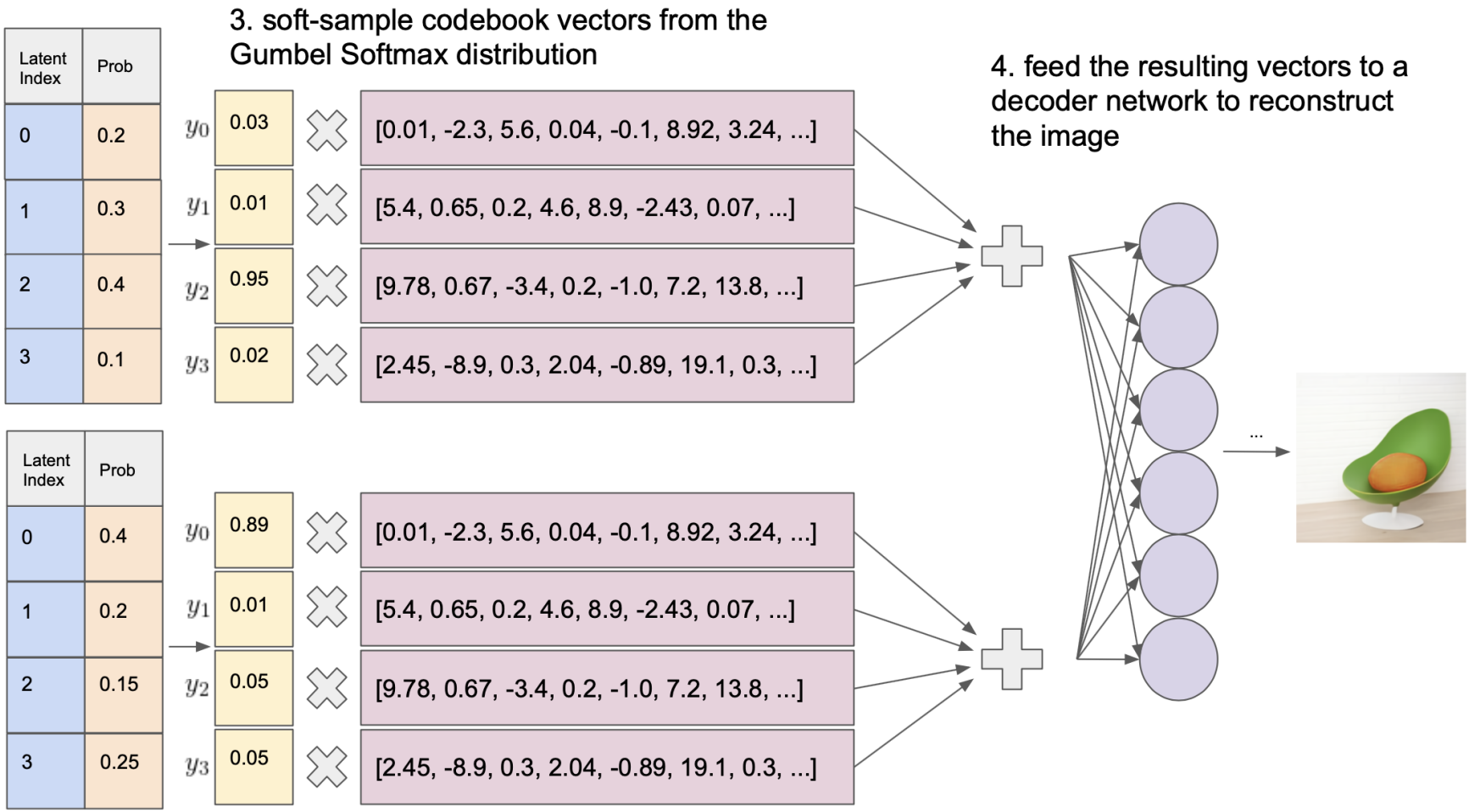
Encoder compress input by 8 times on each spatial dimension and outputs 8192 logits of categorical distribution. Relaxation of latent variables $(y_1, \ldots, y_{8192})$ is computed via Gumbel-Softmax (2) trick.




| Latent Index | Codebook Vector |
|---|---|
| 0 | [0.01, -2.3, 5.6, 0.04, -0.1, 8.92, 3.24, ...] |
| 1 | [5.4, 0.65, 0.2, 4.6, 8.9, -2.43, 0.07, ...] |
| 2 | [9.78, 0.67, -3.4, 0.2, -1.0, 7.2, 13.8, ...] |
| 3 | [2.45, -8.9, 0.3, 2.04, -0.89, 19.1, 0.3, ...] |

# DVAE in Dall-E

Decoder uses set of learnable vectors[2] to map relaxation or one hot vector. into continuous space: $I = y^T W, y \in R^{8192}, W \in R^{8192 \times 128}$.




[2] In practice it is done with conv with kernel size 1, source code link

# DVAE in Dall-E

Commonly used distributions (Normal, Laplace) $p_\theta(x|z)$ is unbounded, but pixels is bounded.
To solve this mismatch:

- Pixel values converted using $\phi : [0, 255] \to (\epsilon, 1 - \epsilon)$, where $\phi(x) = \frac{1-2\epsilon}{255}x + \epsilon$.
- Logit-Laplace distribution is used as likelihood:

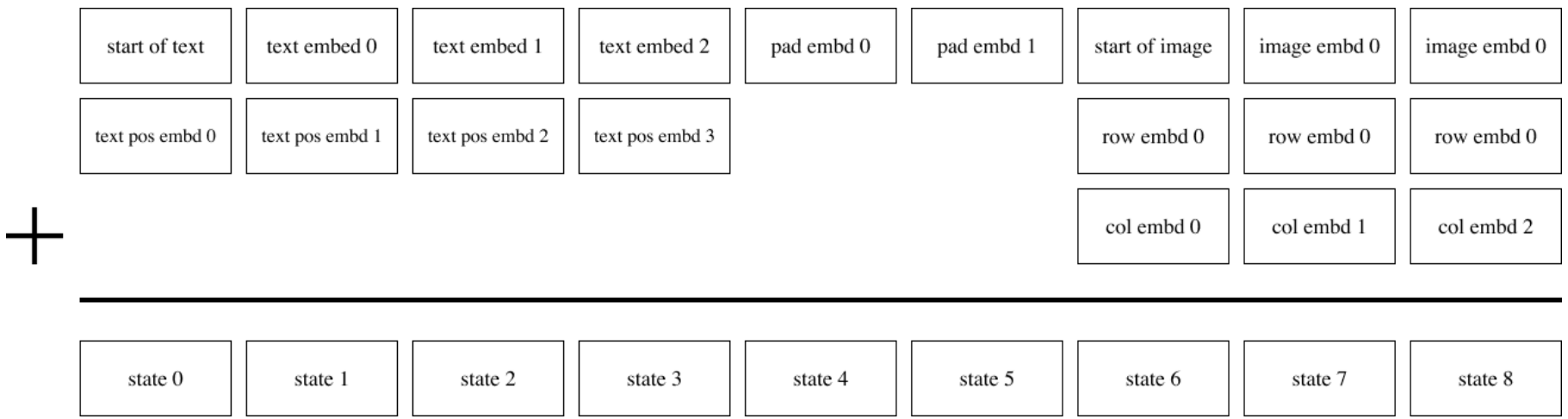
$$p_\theta(x|z) = \frac{1}{2b(z)x(1-x)} \exp\left(-\frac{|\sigma^{-1}(x) - \mu(z)|}{b(z)}\right).$$

- To reconstruct image $\hat{x}$ from decoder outputs $(\mu(z), b(z))$ :

$$\hat{x} = \varphi^{-1}(\sigma(\mu(z))$$

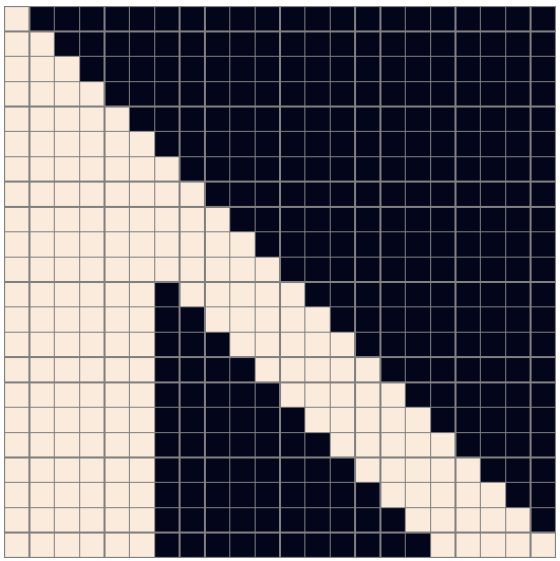
# Model input

Model uses as input 256 BPE [2] text tokens with vocabulary size 16384 concatenated with $32 \times 32 = 1024$ image tokens with vocabulary size 8192.

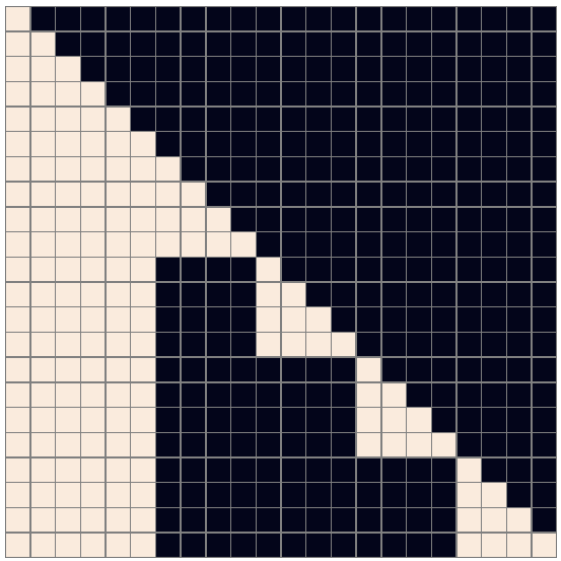


# Sparse Transformer

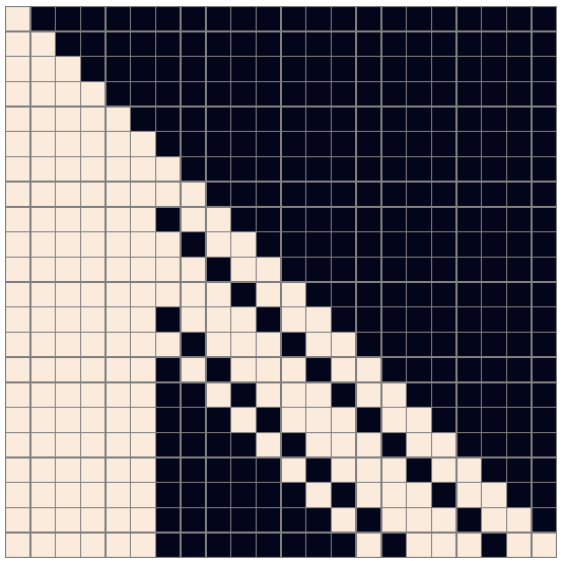
In order to reduce memory consumption and computational time attention is restricted to access only part of elements via masking.
Transformer has 64 attention layers, each with 62 heads with $d_{head} = 64$.



(a) Row attention mask. Used on layers with index $i: \ i - 2 \mod 4 \neq 0$



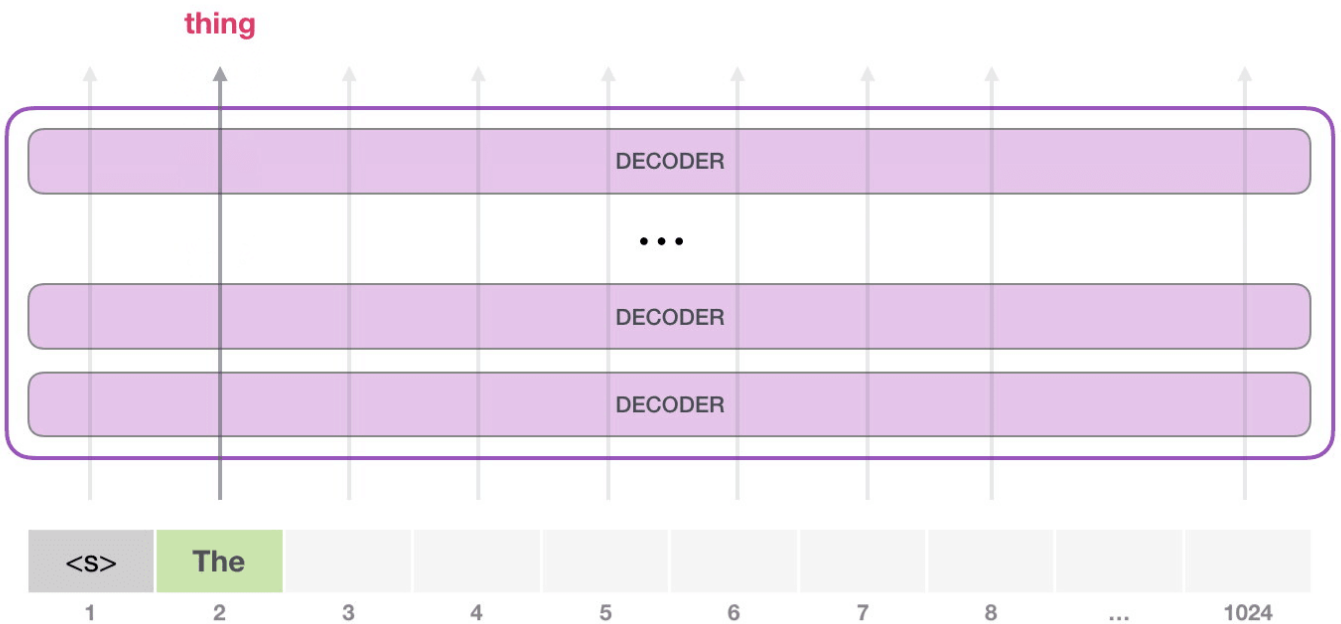
(b) Column attention mask with transposed image states. Used on layers with index $i: \ i - 2 \mod 4 = 0$



(c) Convolutional attention mask. Used only on last attention layer.

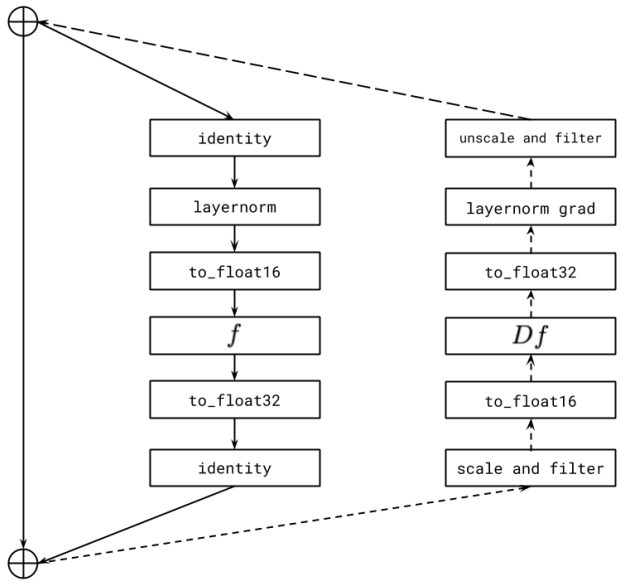
# Transformer in Dall-E

Transformer in Dall-E has 12 billion parameters (24 GB to only store model weights in 16 bit precision).
Model is decoder transformer learned to model text and image in language modelling setup (minimize cross entropy between input and output tokens with causal masking).



# Tricks for leaning huge model

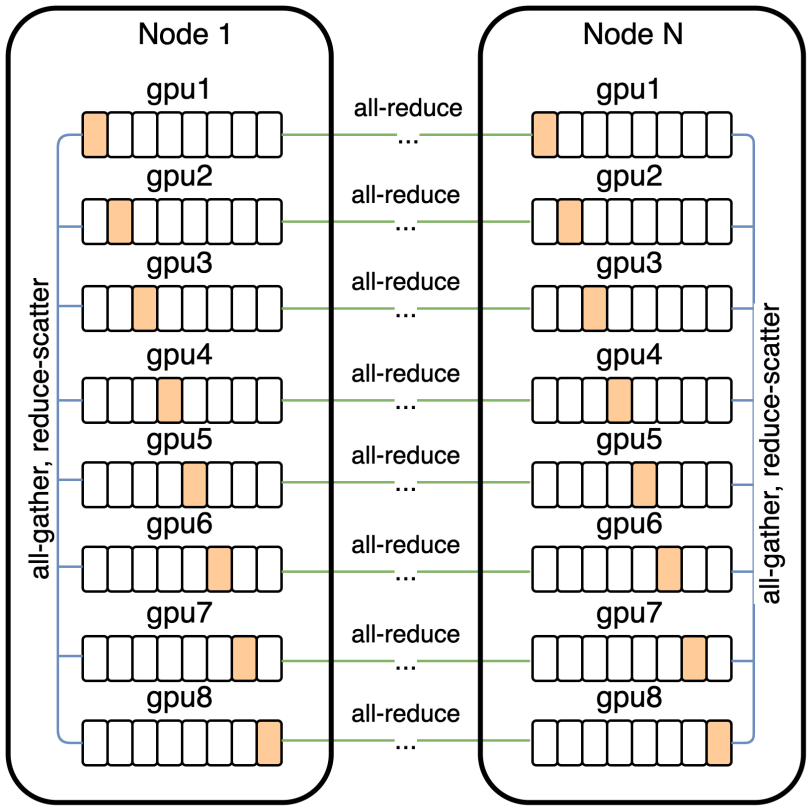
1. To save GPU memory and increase throughput, most parameters, Adam moments, and activations are stored in 16-bit precision.
2. Underflow in the 16-bit gradients is the root of instability.




3. As the model is made deeper and wider, the true exponents of the activation gradients for later resblocks can fall below the minimum exponent of the 16-bit format (Underflow).

# Distributed Optimization

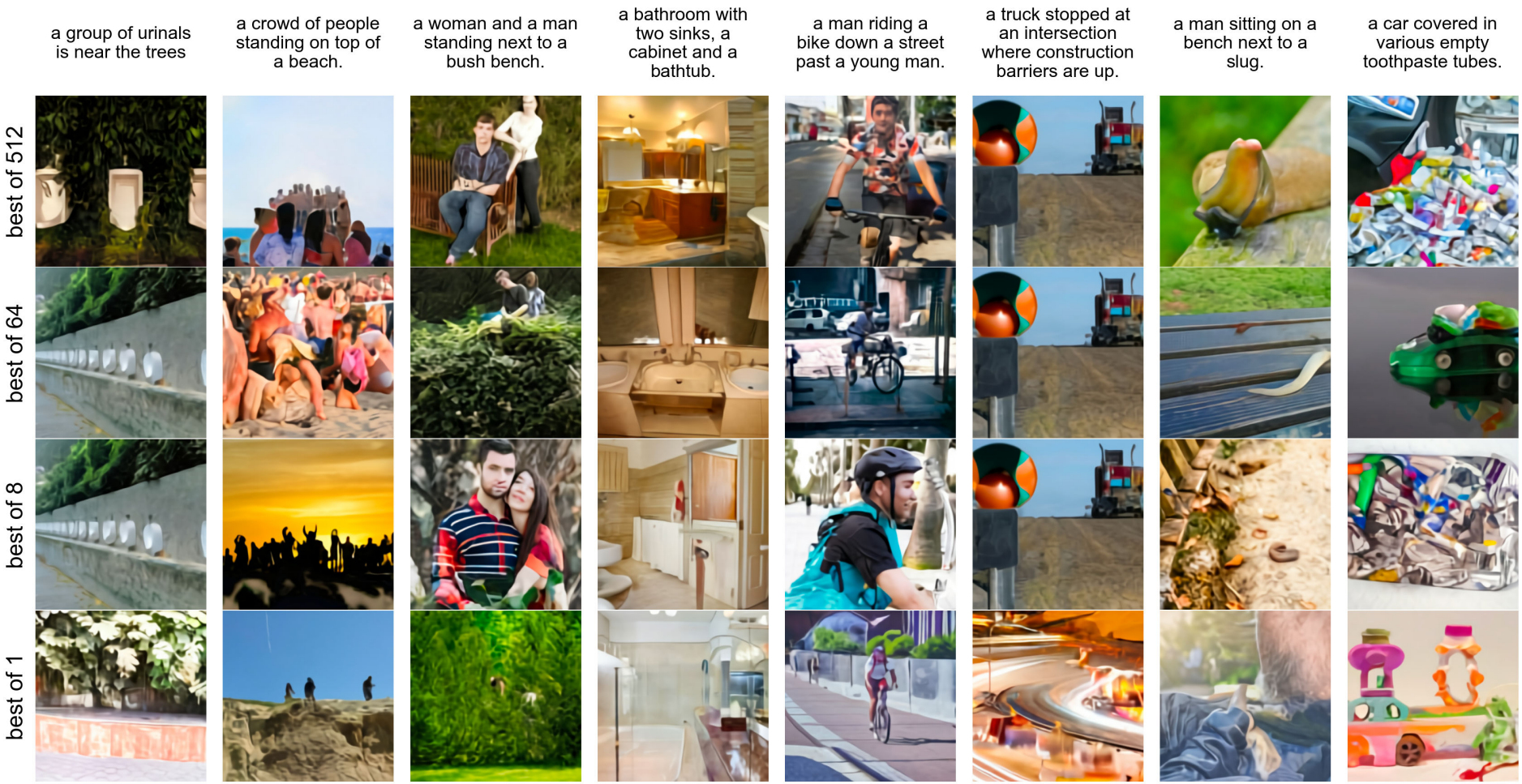
1. Parameter sharding [3] used to store model(24GB) on multiple 16 GB GPUs.
2. Each GPU in a machine computes the low-rank factors using PowerSGD [4] for its parameter shard gradients.
3. Communication between nodes is heavily reduced via sending low-rank factors instead of gradients.



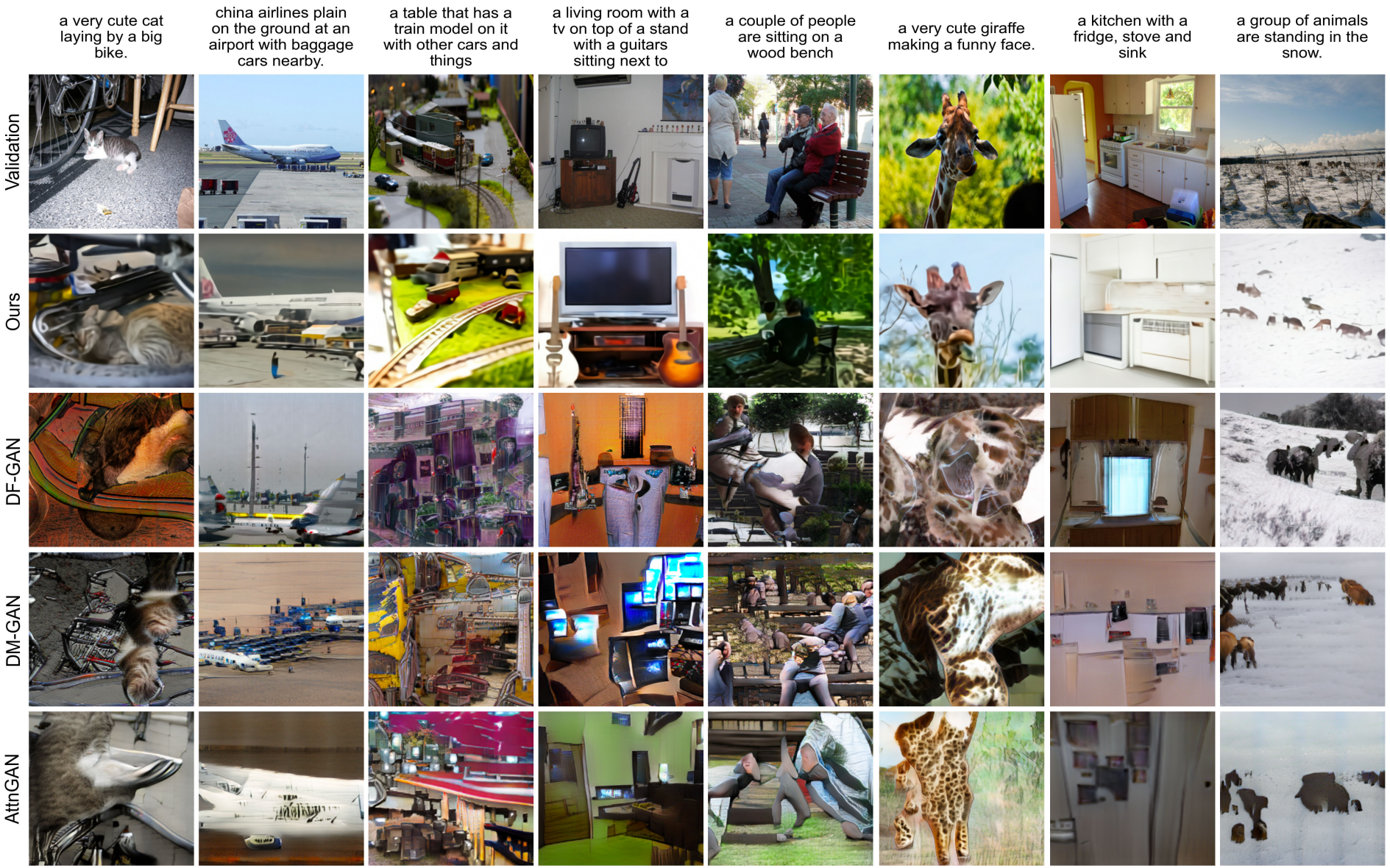

| Effective Parameter Count | Compression Rank | Compression Rate |
|---|---|---|
| $2.8 \cdot 10^9$ ($d_{\text{model}} = 1920$) | 512 | $\approx$83% |
| $5.6 \cdot 10^9$ ($d_{\text{model}} = 2688$) | 640 | $\approx$85% |
| $12.0 \cdot 10^9$ ($d_{\text{model}} = 3968$) | 896 | $\approx$86% |

# Sampling Process

In each experiment 512 text-image pairs are generated and ranked via CLIP model [5] (selecting only top 1 pair).

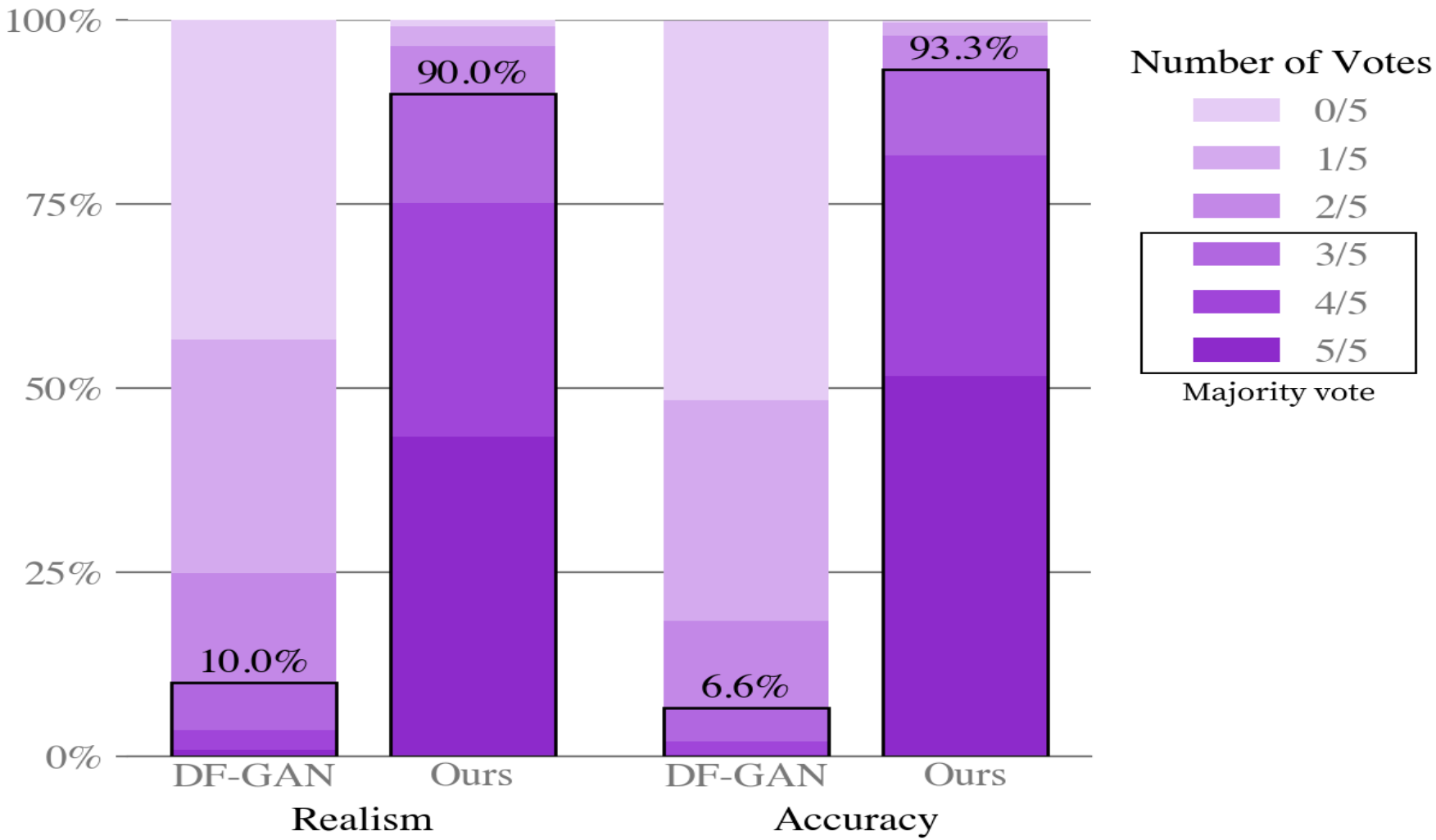


# Compare sample from other approaches

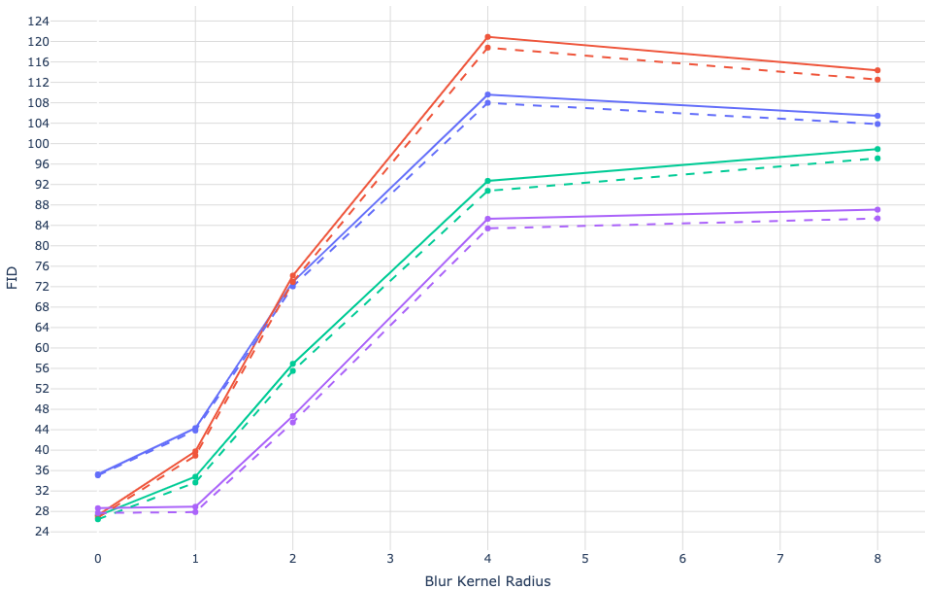


# Human estimation

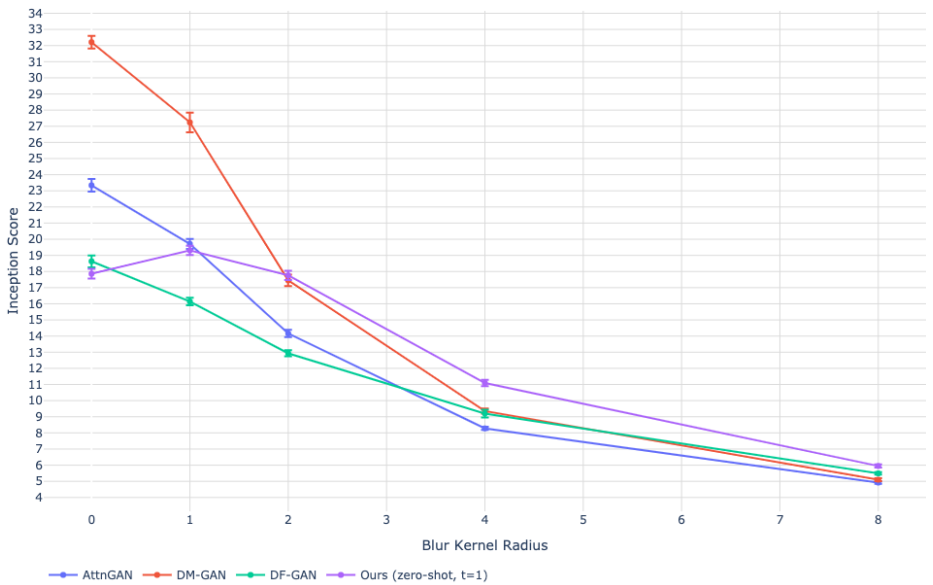
Given text-image pair from Dall-E and other model 5 annotators asked which better matches caption and more realistic.



# MS COCO FID and IS

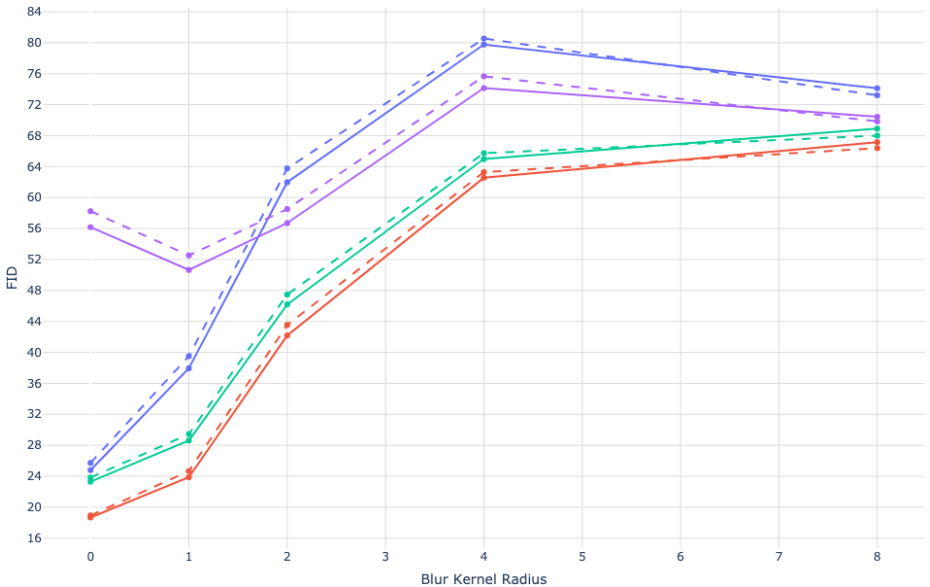



(a) FID on MS-COCO as a function of blur radius.

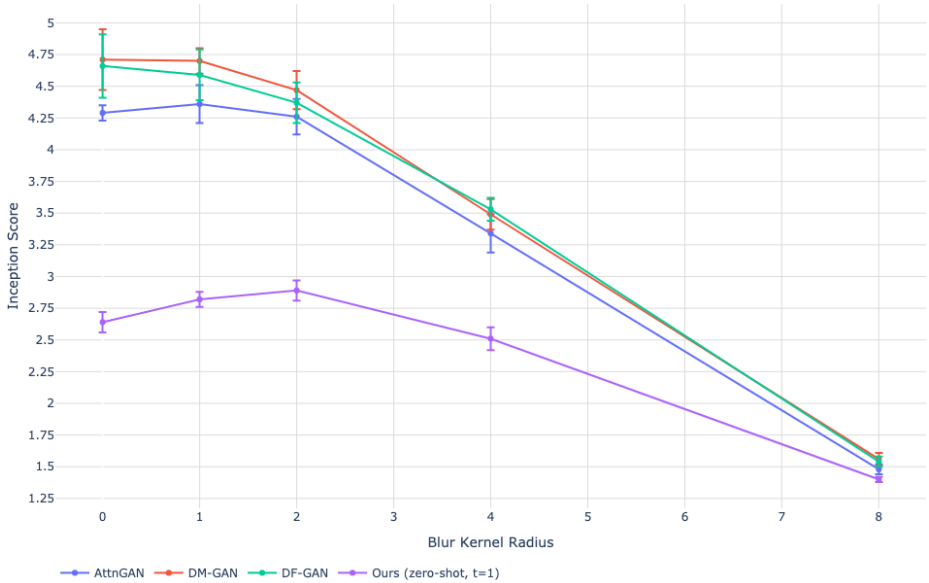



(b) IS on MS-COCO as a function of blur radius.

# CUB FID and IS

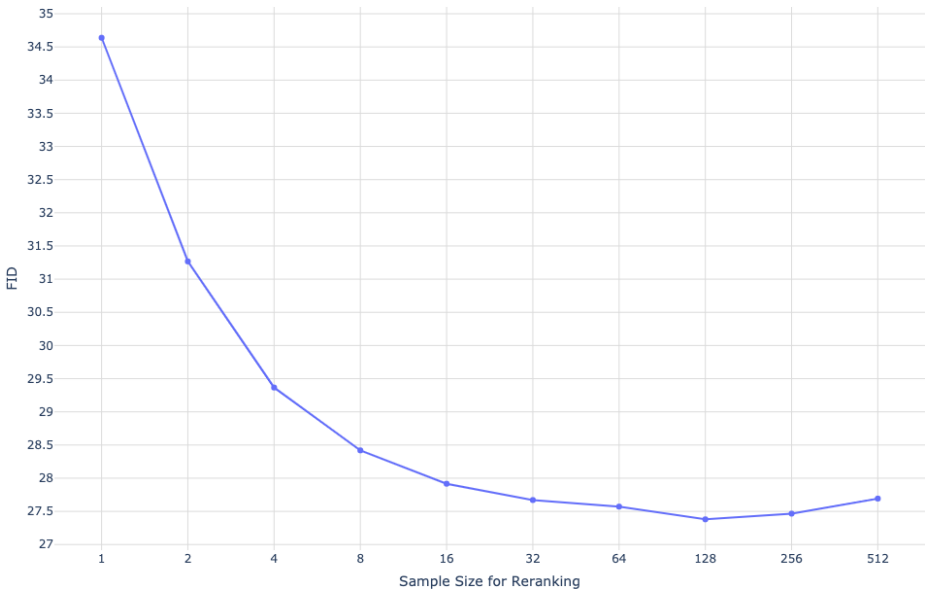



(a) FID on CUB as a function of blur radius.

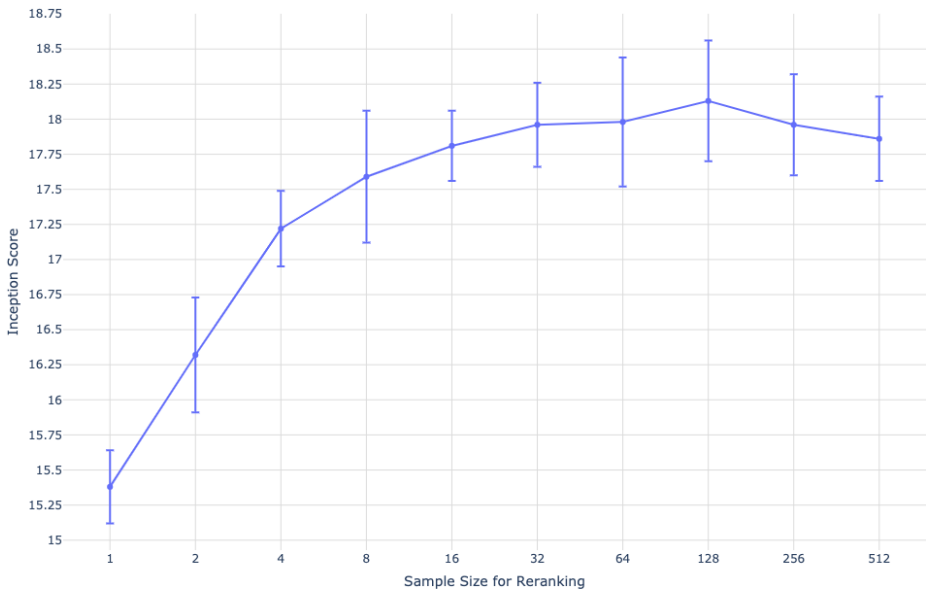



(b) IS on CUB as a function of blur radius.

# Reranking size

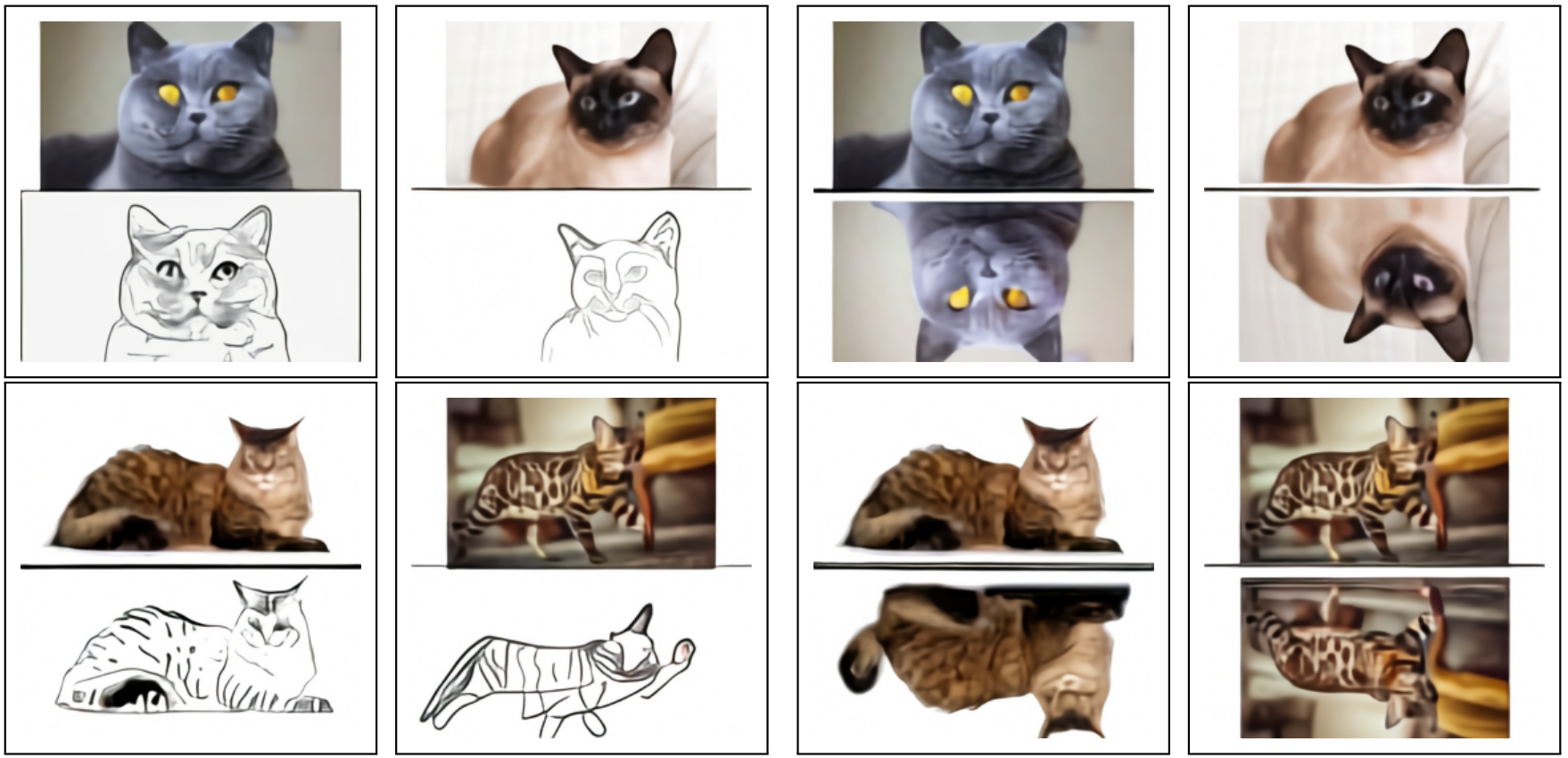



(a) FID on MS-COCO as a function of the sample size used for reranking.




(b) IS on MS-COCO as a function of the sample size used for reranking.

# Image-to-Image translation



(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"

# Image-to-Image translation

(a) "2 panel image of the exact same cat. on the top, a photo of the cat. on the bottom, the cat with sunglasses."

(b) "the exact same cat on the top as a postage stamp on the bottom"

# Summary

- Dall-E approach: learn VAE on images using discrete prior  learn transformer to generate prior conditionally on text.
- 12 billion transformer can be trained, but it is very challenging.
- No ablation study (Do we really nead that huge model?)
- Dataset is not released, collection process is not clear.
- DVAE model released, but it is useless without transformer prior.

[1] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut.
Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[2] Rico Sennrich, Barry Haddow, and Alexandra Birch.
Neural machine translation of rare words with subword units.
*CoRR*, abs/1508.07909, 2015.

[3] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He.
Zero: Memory optimization towards training A trillion parameter models.
*CoRR*, abs/1910.02054, 2019.

[4] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi.
Powersgd: Practical low-rank gradient compression for distributed optimization, 2020.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.
Learning transferable visual models from natural language supervision, 2021.