DALL·E: Zero-Shot Text-to-Image Generation

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever

OpenAI

Arxiv

Presenter: Minho Park

Motivation

- Image-text datasets (e.g., MS-COCO, CUB-200) are relatively small in the Text-to-Image task.
- They train a 12-billion parameter autoregressive transformer on 250 million image-text pairs.
  12,000,000,000 * 32bit (w/o mixed precision) = 48GB…
  - The image generation is evaluated on the MS-COCO dataset zero-shot.

(a) a tapir made of accordion. a tapir with the texture of an accordion.

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

(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign

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

Figure 2. With varying degrees of reliability, our model appears to be able to combine distinct concepts in plausible ways, create anthropomorphized versions of animals, render text, and perform some types of image-to-image translation.

Method

- Train a transformer to autoregressively model the text and image tokens as a single stream of data.
- However, they can not use pixels as image tokens directly.

- Using a two-stage training procedure.
- Stage 1: Train a discrete variational autoencoder (dVAE)
  - To compress 256×256 images into 32×32 grid of image tokens.
- Stage 2: Train an autoregressive transformer.
  - Concatenate up to 256 BPE-encoded text tokens with the $32 \times 32 = 1024$ image tokens.
  - To model the joint distribution over the text and image tokens.

Formulation

- The variational bound (Evidence Lower BOund, ELBO)
- $p_\theta(x) = p_\theta(x, z) = p_\theta(x|z)p(z)$

$$\log p_\theta(x) \geq \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}\left(q_\phi(z|x) \parallel p_\theta(z)\right)$$

- Image x, captions y, and the tokens z.
- $p_{\theta,\psi}(x, y) = p_{\theta,\psi}(x, y, z) = p_\theta(x|y, z)p_\psi(y, z)$

$$\log p_{\theta,\psi}(x, y) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|y, z)] - \beta D_{KL}\left(q_\phi(y, z|x) \parallel p_\psi(y, z)\right)$$

In practice, $\beta > 1$ is helpful

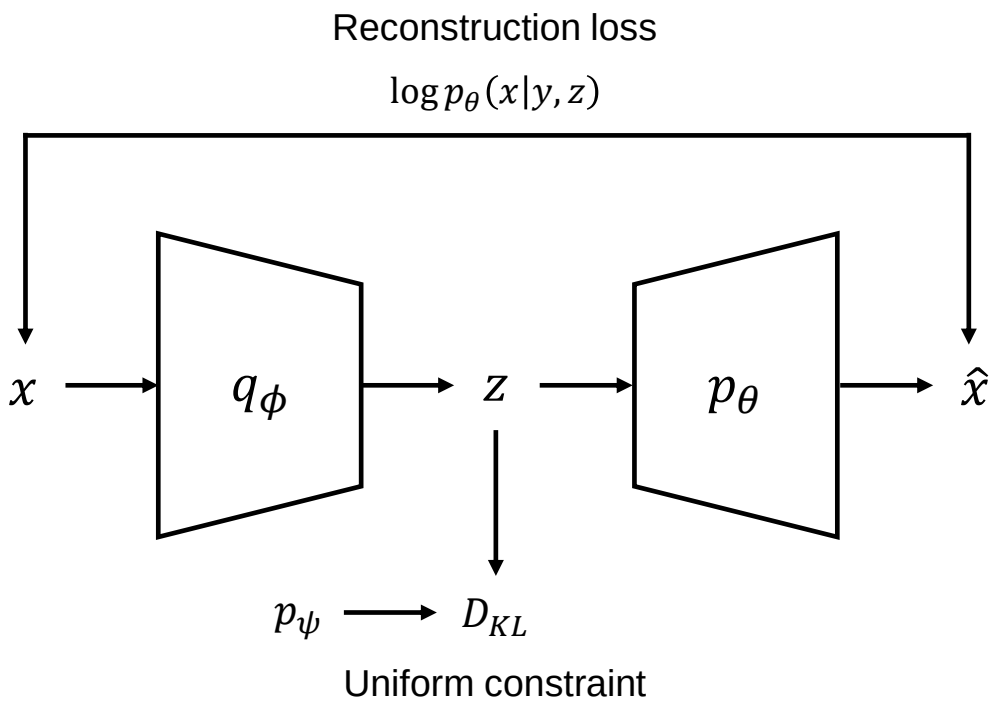
Formulation

$$\log p_{\theta,\psi}(x,y) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|y,z)] - \beta D_{KL}\left(q_\phi(y,z|x) \,\|\, p_\psi(y,z)\right)$$

- q_ϕ: The distribution over the image tokens generated by the dVAE encoder given the image x.
- p_θ: The distribution over the images generated by the dVAE decoder given the image tokens.
- p_ψ : The distribution over the text and image tokens modeled by the transformer.
- Maximize ELBO with respect to ϕ, θ, and ψ.

Stage One: Learning the Visual Codebook

- Maximize ELBO with respect to θ, and ϕ.
- Set the initial prior p_ψ to the uniform categorical distribution.



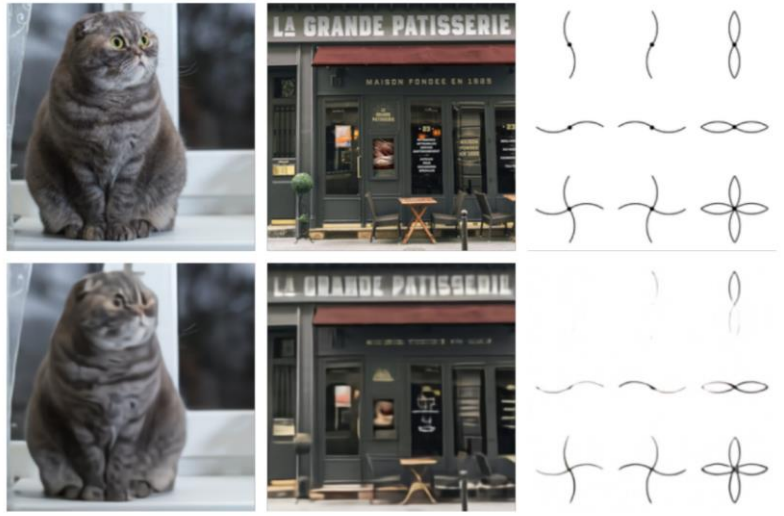
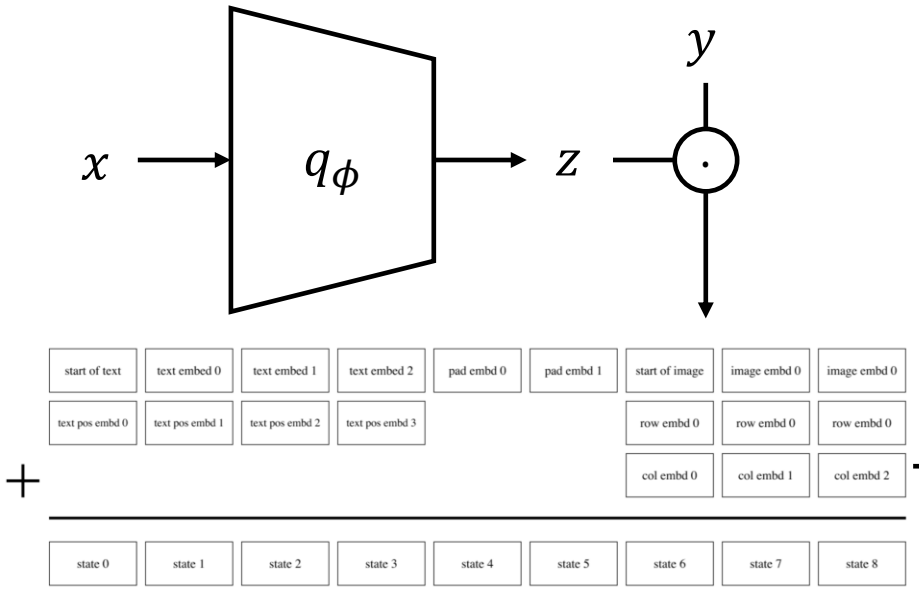


Figure 1. Comparison of original images (top) and reconstructions from the discrete VAE (bottom). The encoder downsamples the spatial resolution by a factor of 8. While details (e.g., the texture of the cat's fur, the writing on the storefront, and the thin lines in the illustration) are sometimes lost or distorted, the main features of the image are still typically recognizable. We use a large vocabulary size of 8192 to mitigate the loss of information.

Stage Two: Learning the Prior

- Maximize ELBO with respect to ψ.
- Standard causal mask in text
- Since we are primarily interested in image modeling, lambda of text and image for XE loss are 1/8, 7/8 respectively.



Decoder-only Transformer

64 Self-attention layers

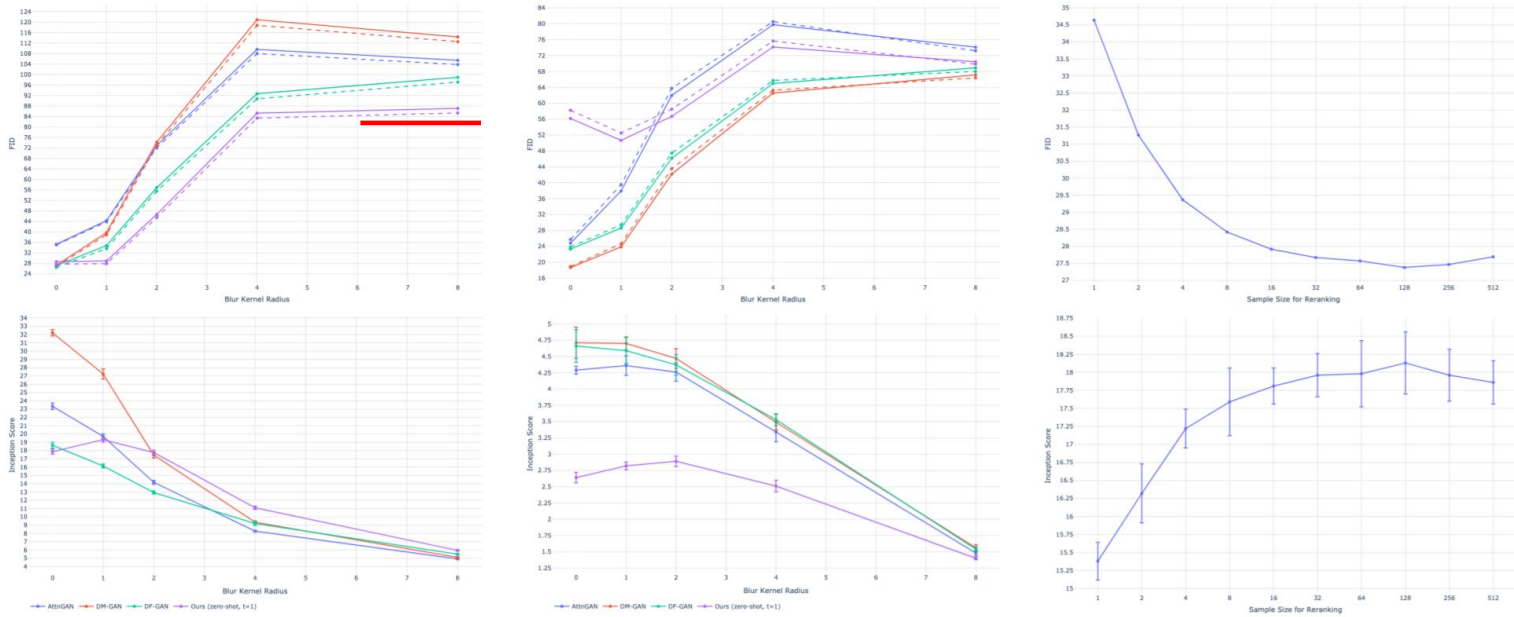
12 Billion Parameter

Implementation Details

- Data Collection: Create a dataset by collecting 250 million text-image pairs from the internet.
  - Text-image pairs from Wikipedia, and a filtered subset of YFCC100M (not include MS-COCO captions).
- Mixed-Precision Training: Most parameters, Adam moments, and activations are stored in 16-bit precision.
- Distributed Optimization: Using parameter sharding (24GB).

Quantitative Results

- FID and IS on MS-COCO and CUB.



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

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

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

Figure 9. Quantitative results on MS-COCO and CUB. Solid lines represent FID computed against the original validation sets, and dashed lines represent FID computed against validation sets with overlapping images removed (see Section 3.2). For MS-COCO, we evaluate all models on a subset of 30,000 captions sampled from the validation set. For CUB, we evaluate all models on all of the unique captions in the test set. We compute the FID and IS using the DM-GAN code, which is available at `https://github.com/MinfengZhu/DM-GAN`.

Qualitative Findings

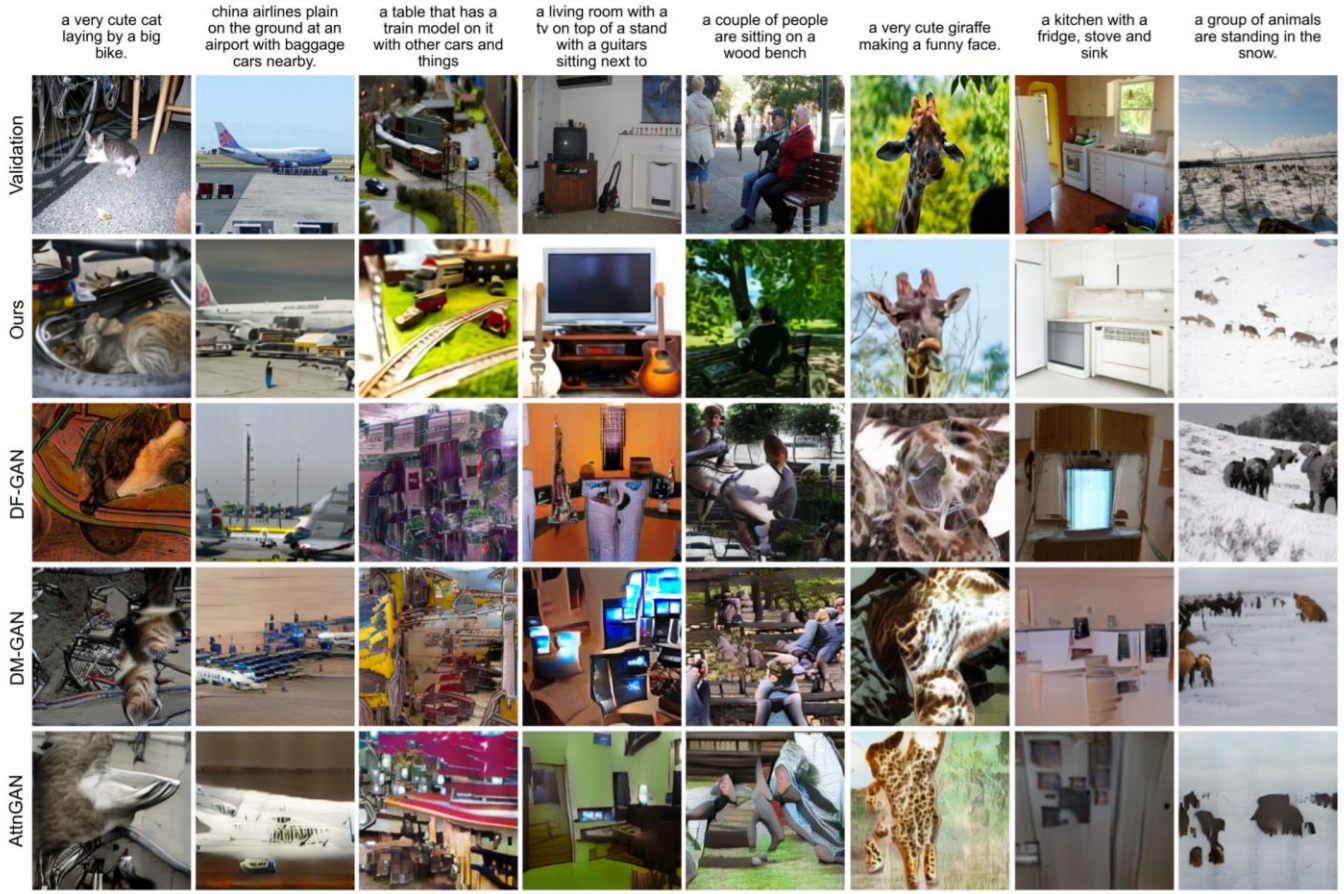


Figure 3. Comparison of samples from our model to those from prior approaches on captions from MS-COCO. Each of our model samples is the best of 512 as ranked by the contrastive model. We do not use any manual cherrypicking with the selection of either the captions or the samples from any of the models.

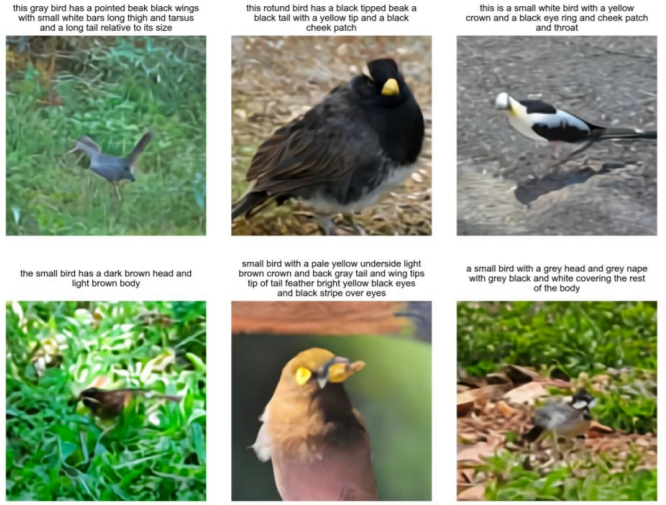


Figure 8. Zero-shot samples from our model on the CUB dataset.

Sample Generation

- Re-rank the samples drawn from the transformer using a pretrained contrastive model (CLIP).

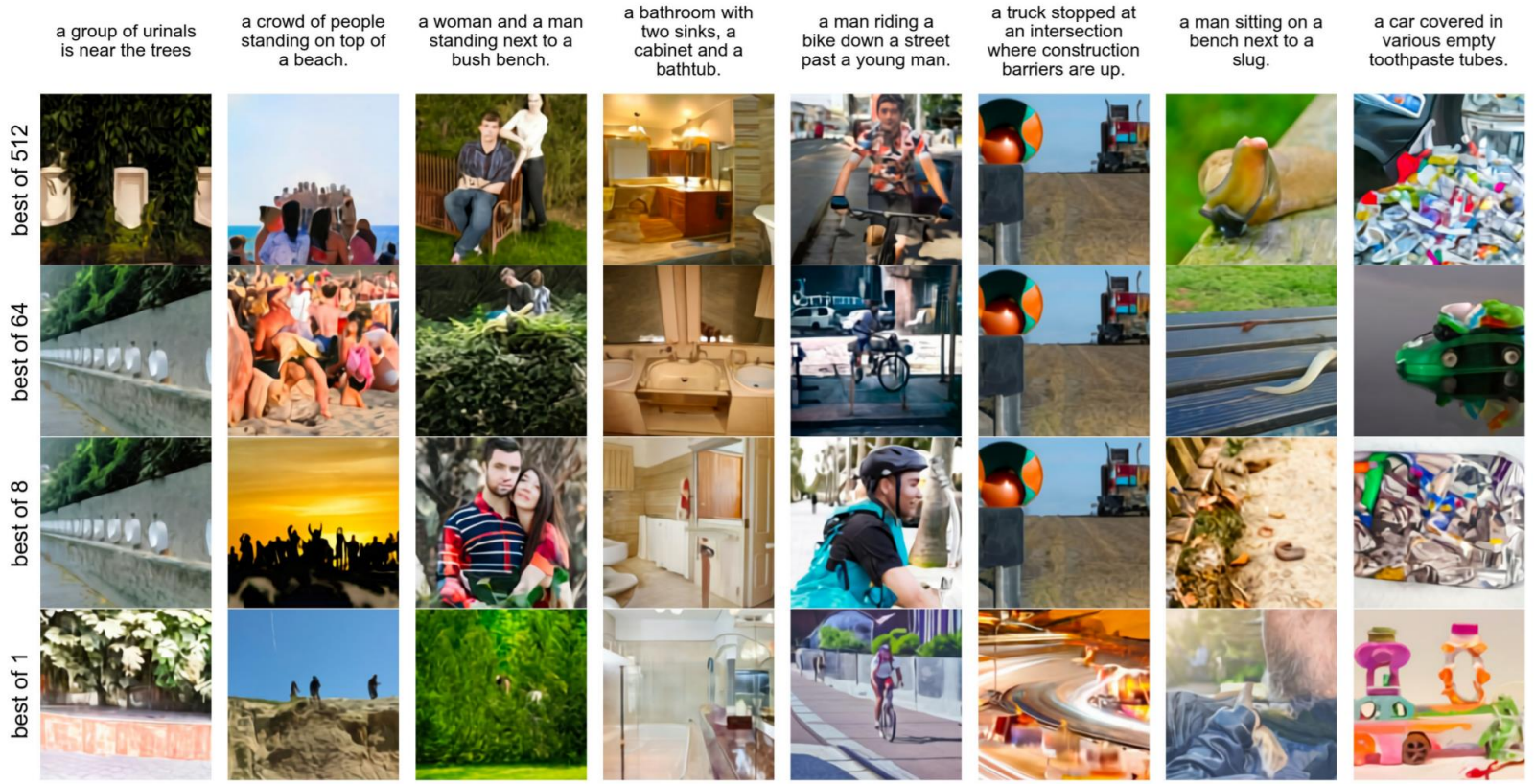


Figure 6. Effect of increasing the number of images for the contrastive reranking procedure on MS-COCO captions.

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *arXiv preprint arXiv:2103.00020* (2021).

References

- Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *arXiv preprint arXiv:2102.12092* (2021).
- Official code, https://github.com/openai/dall-e
- Lucidrains, DALL-E pytorch code, https://github.com/lucidrains/DALLE-pytorch
- Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *arXiv preprint arXiv:2103.00020* (2021).