



Text-to-Image Synthesis with Conditional Diffusion Models

Aman Shrivastava

CS PhD Candidate, University of Virginia

Conditional diffusion models

In conditional diffusion models, an additional input, y (eg. a class label or a text sequence) is available and we try to model the conditional distribution $p(x | y)$ instead.

This allows us to generate data given the conditioning signal.

Some examples generated from Google's Imagen [1], and OpenAI's Dalle-2 [2] on the right.



A medieval painting of the wifi not working



A still of Homer Simpson in Psycho (1960)



An Alpaca is smiling and underwater in the pool



A tulip pushing a baby carriage

[1] Saharia, Chitwan, et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." *arXiv preprint arXiv:2205.11487* (2022).

[2] Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." *arXiv preprint arXiv:2204.06125* (2022).

Conditional diffusion models

In practice, the denoising model $p(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{y})$ is also conditioned on ' \mathbf{y} ' in addition to the image from the previous timestep, \mathbf{x}_t

Reverse process: $p_\theta(\mathbf{x}_{0:T} | \mathbf{c}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}), \quad p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t, \mathbf{c}), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t, \mathbf{c}))$

Variational upper bound: $L_\theta(\mathbf{x}_0 | \mathbf{c}) = \mathbb{E}_q \left[L_T(\mathbf{x}_0) + \sum_{t>1} D_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})) - \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1, \mathbf{c}) \right].$

Practical considerations

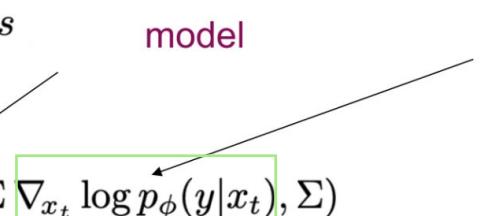
- **Scalar conditioning:** encode scalar as a vector embedding, simple spatial addition or adaptive group normalization layers.
- **Image conditioning:** channel-wise concatenation of the conditional image.
- **Text conditioning:** single vector embedding – spatial addition or adaptive group norm / a seq of vector embeddings – cross-attention.

Score-model based guidance

Using the gradient of an independently pre-trained score model as guidance
Given a conditional model $p(x_t | y)$, we use gradients from an extra score model $p(y | x_t)$ during sampling.

Algorithm 1 Classifier guided diffusion sampling, given a diffusion model $(\mu_\theta(x_t), \Sigma_\theta(x_t))$, classifier $p_\phi(y|x_t)$, and gradient scale s .

```
Input: class label  $y$ , gradient scale  $s$ 
 $x_T \leftarrow$  sample from  $\mathcal{N}(0, \mathbf{I})$ 
for all  $t$  from  $T$  to 1 do
     $\mu, \Sigma \leftarrow \mu_\theta(x_t), \Sigma_\theta(x_t)$ 
     $x_{t-1} \leftarrow$  sample from  $\mathcal{N}(\mu + s\Sigma \nabla_{x_t} \log p_\phi(y|x_t), \Sigma)$ 
end for
return  $x_0$ 
```



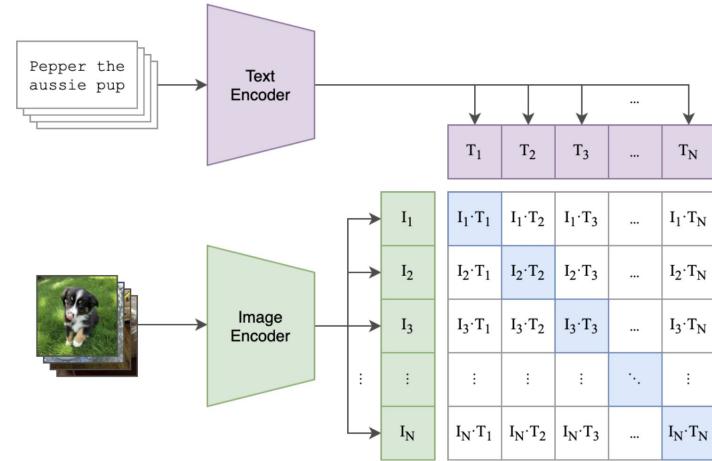
The diagram illustrates the flow of information in the algorithm. It shows two main components: 'model' and 'Classifier gradient'. Arrows point from 'model' to both 'mu, Sigma' and 'Classifier gradient'. Arrows point from 'Classifier gradient' to the term 'nabla_x_t log p_phi(y|x_t)' in the assignment statement.

CLIP guidance

Given an image x and a prompt y , a CLIP model computes the alignment $\text{cos_sim}(x, y)$ which indicates how similar the image and the prompt are.

To use this signal for guidance, we assume that the CLIP similarity score is a good estimation of the function $p(y|x)$

The gradient of this score wrt the noised image, x_t at timestep t is used as the guidance gradient



*Note that this requires the CLIP model to compute score for **noised-images** at intermediate timesteps, hence a noised-CLIP model is trained for guidance*

Classifier-free guidance

Given both a conditional and an unconditional DM, we can design an “implicit” classifier:

$$p(\mathbf{c}|\mathbf{x}_t) \propto p(\mathbf{x}_t|\mathbf{c})/p(\mathbf{x}_t)$$

Conditional diffusion model Unconditional diffusion model

In practice, $p(\mathbf{x}|\mathbf{c})$ and $p(\mathbf{x})$ are trained together by randomly dropping the conditioning signal with a certain probability during training.

Using above, the score-gradient becomes:

$$\begin{aligned}\nabla_{\mathbf{x}_t} [\log p(\mathbf{x}_t|\mathbf{c}) + \omega \log p(\mathbf{c}|\mathbf{x}_t)] &= \nabla_{\mathbf{x}_t} [\log p(\mathbf{x}_t|\mathbf{c}) + \omega (\log p(\mathbf{x}_t|\mathbf{c}) - \log p(\mathbf{x}_t))] \\ &= \nabla_{\mathbf{x}_t} [(1 + \omega) \log p(\mathbf{x}_t|\mathbf{c}) - \omega \log p(\mathbf{x}_t)]\end{aligned}$$

GLIDE | OpenAI

- A 64x64 base diffusion model
- A 64 -> 256 conditional super-resolution model
- Evaluates both classifier-free and **CLIP guidance**

CLIP guidance: Use the CLIP alignment score $p(x, y)$ as a estimation of $p(y | x)$



“a boat in the canals of venice”



“a painting of a fox in the style of starry night”



“a crayon drawing of a space elevator”



“a futuristic city in synthwave style”

DALL.E 2 | OpenAI

- 1kx1k text-conditioned image generation
- Uses a **prior** to produce CLIP embeddings conditioned on the text-caption
- Uses a **decoder** to produce images conditioned on the CLIP embeddings



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



panda mad scientist mixing sparkling chemicals, artstation

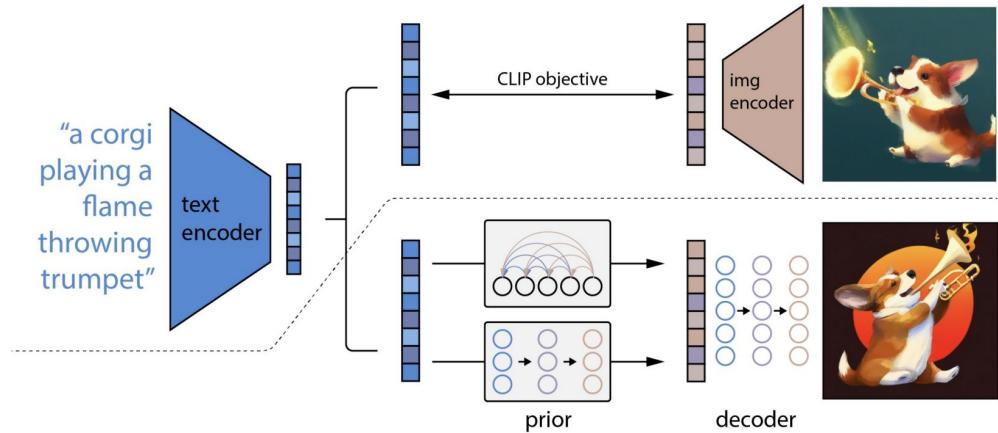


a corgi's head depicted as an explosion of a nebula

DALL.E 2 | Open AI

Conditioning on CLIP-embeddings

- Helps capture multimodal representations
- The bi-partite latent enables several text-controlled image manipulation tasks



DALL.E 2 | Open AI

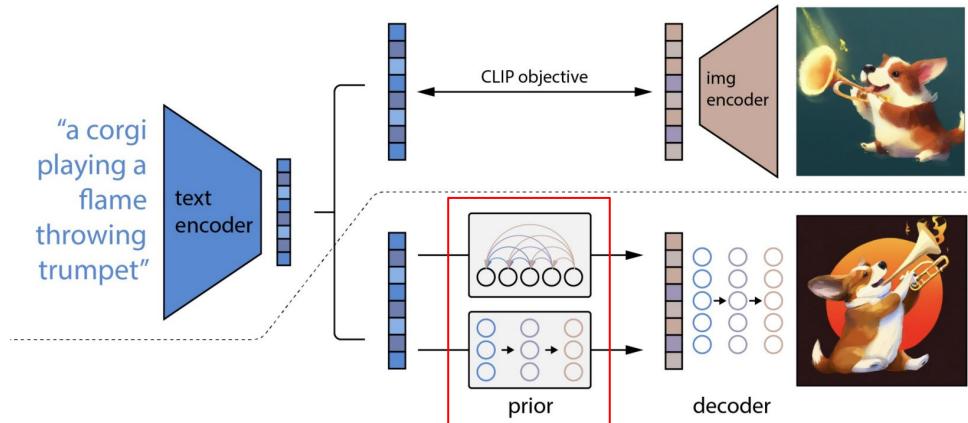
Proposes 2 types of priors:

1. Autoregressive prior

Quantize image embeds into a sequence of discrete codes and predict them autoregressively

2. Diffusion prior

Model continuous image embeddings by DMs conditioned on caption

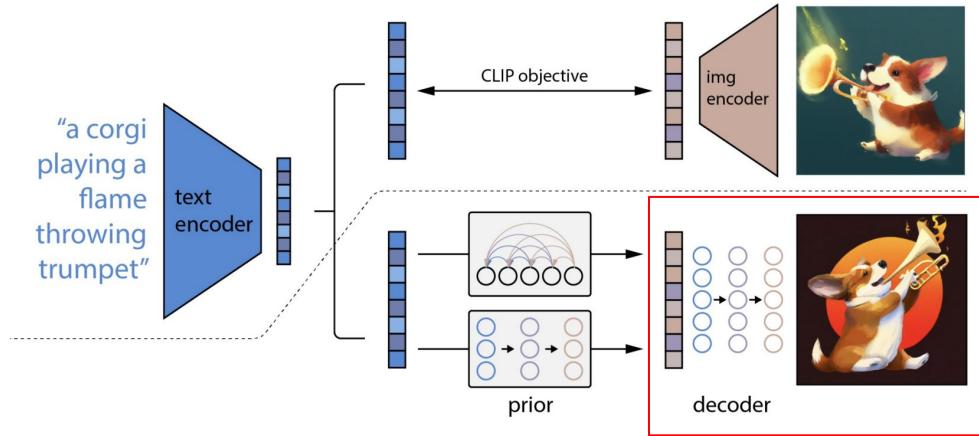


DALL.E 2 | Open AI

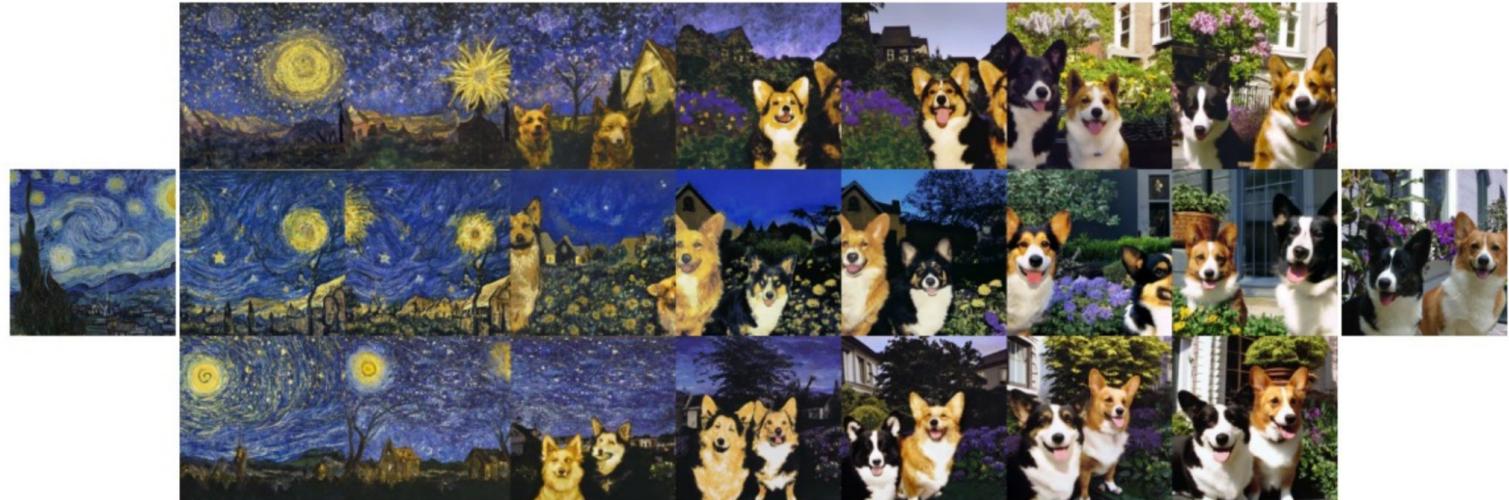
Decoder: produces images conditioned on CLIP image embeddings (and text caption)

The model is trained as cascaded DMs 64->256->1024

It is observed that classifier-free guidance works better for sample quality here.



DALL.E 2 | Open AI



Interpolate CLIP embeddings to generate different interpolation trajectories

DALL.E 2 | Open AI



a photo of a cat → an anime drawing of a super saiyan cat, artstation



a photo of a victorian house → a photo of a modern house



a photo of an adult lion → a photo of lion cub

Change the image CLIP embedding towards the difference of the text CLIP embeddings of two prompts. Note that decoder latent is kept as a constant.

Imagen | Google Research

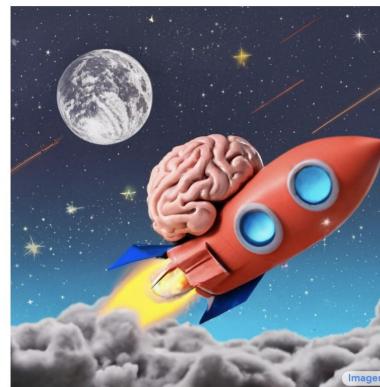
- Generates 1kx1k images
- Exceptional photo-realism
- Extremely simple parameterization
- SOTA on quantitative and qualitative benchmarks
- Proposes a new qualitative benchmark (drawbench)



Teddy bears swimming at the Olympics 400m Butterfly event.



A cute corgi lives in a house made out of sushi.



A brain riding a rocketship heading towards the moon.



A dragon fruit wearing karate belt in the snow.

Imagen | Google Research

Model details

- Cascaded DMs
64 -> 256 -> 1024
- Classifier-free guidance and dynamic thresholding
- Frozen large pretrained language models as text encoders (T5-XXL)

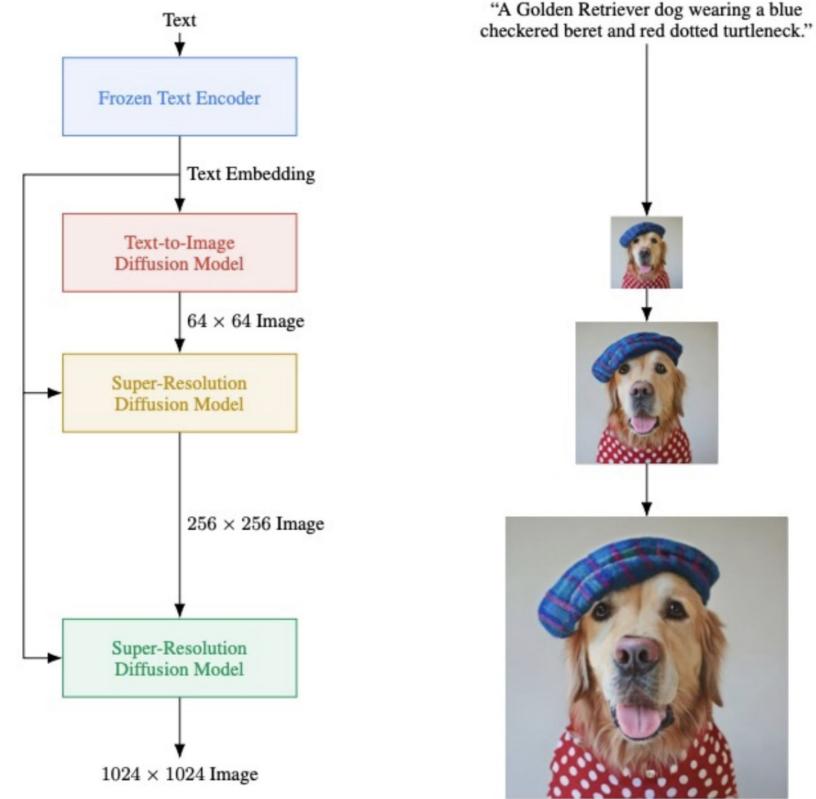


Imagen | Google Research

Main discoveries

- Better text-conditioning signal is important, i.e. large frozen text-encoders are used, eg. T5-XXL
- Stronger classifier-free guidance leads to better text-alignment but worse image quality

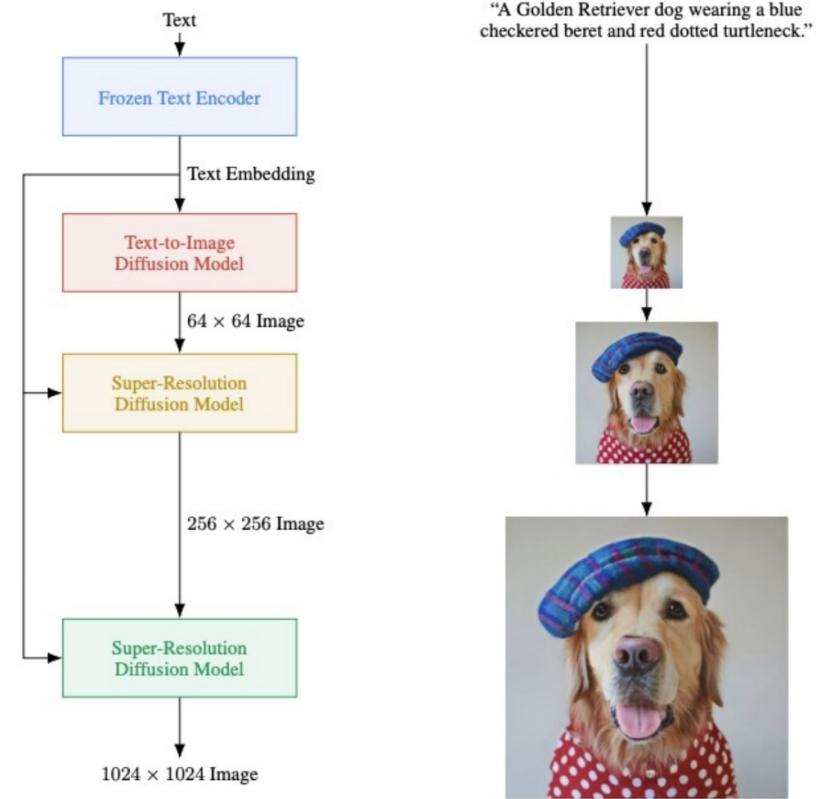


Imagen | Google Research

The paper also proposes a new benchmark called the “drawbench”

- Collection of 200 prompts that test semantic understanding and image diversity.



A brown bird and a blue bear.



One cat and two dogs sitting on the grass.



A sign that says 'NeurIPS'.



A small blue book sitting on a large red book.



A blue coloured pizza.



A wine glass on top of a dog.



A pear cut into seven pieces arranged in a ring.



A photo of a confused grizzly bear in calculus class.



A small vessel propelled on water by oars, sails, or an engine.

Imagen | Google Research

Model	FID-30K	Zero-shot FID-30K
AttnGAN [76]	35.49	
DM-GAN [83]	32.64	
DF-GAN [69]	21.42	
DM-GAN + CL [78]	20.79	
XMC-GAN [81]	9.33	
LAFITE [82]	8.12	
Make-A-Scene [22]	7.55	
DALL-E [53]		17.89
LAFITE [82]		26.94
GLIDE [41]		12.24
DALL-E 2 [54]		10.39
Imagen (Our Work)		7.27

Imagen achieves SOTA using auto-evaluation scores on COCO dataset

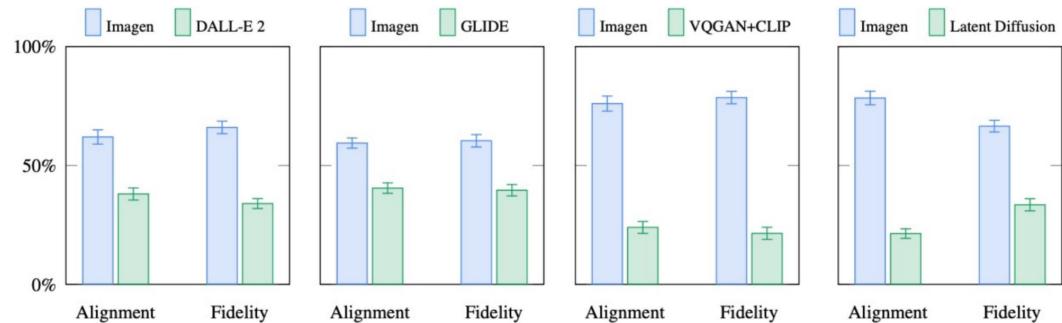


Imagen is preferred over recent work by human raters in sample quality & image-text alignment on DrawBench