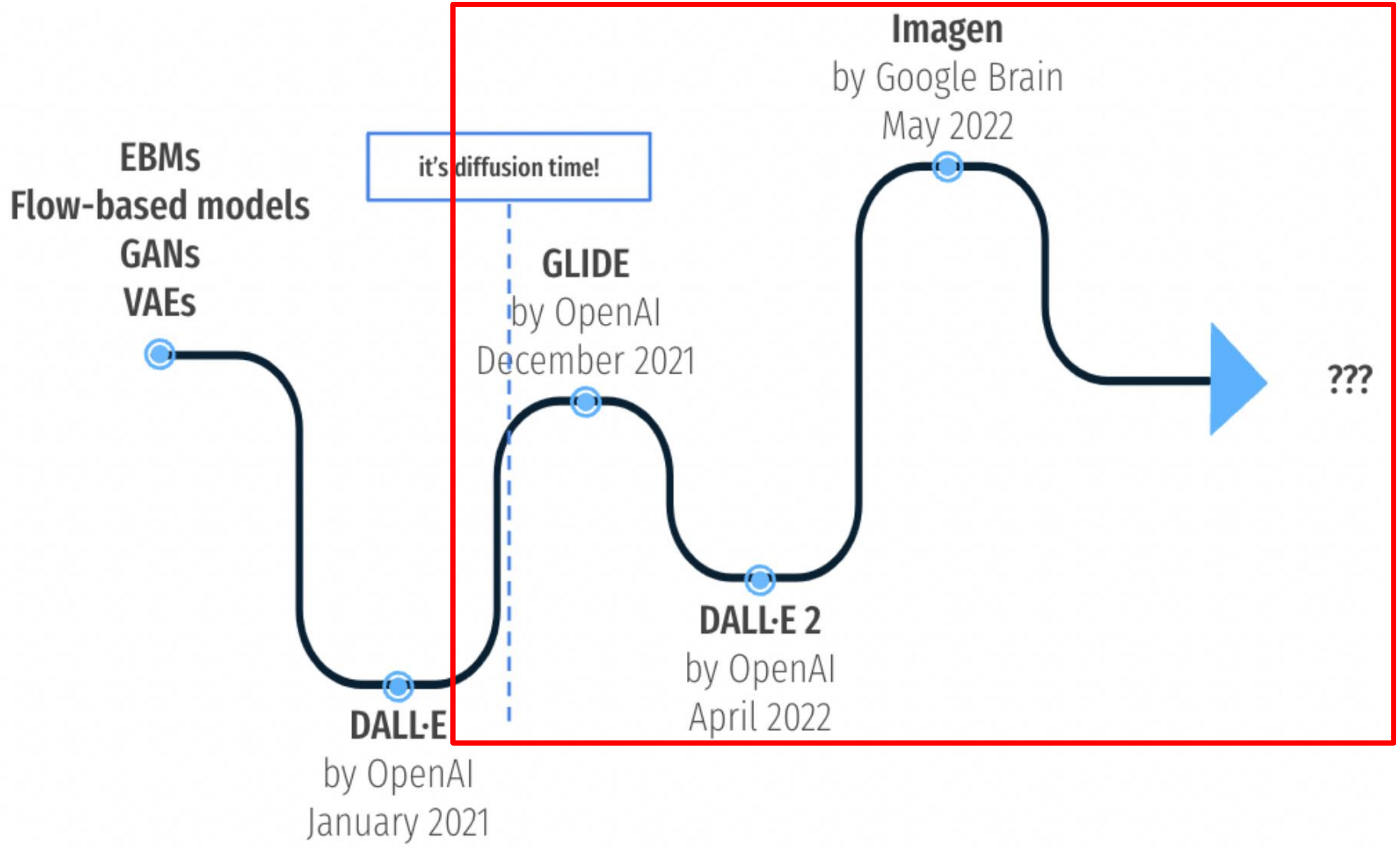
Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Google Research, Brain Team

Xinjie Li

03/15/2023

Timeline



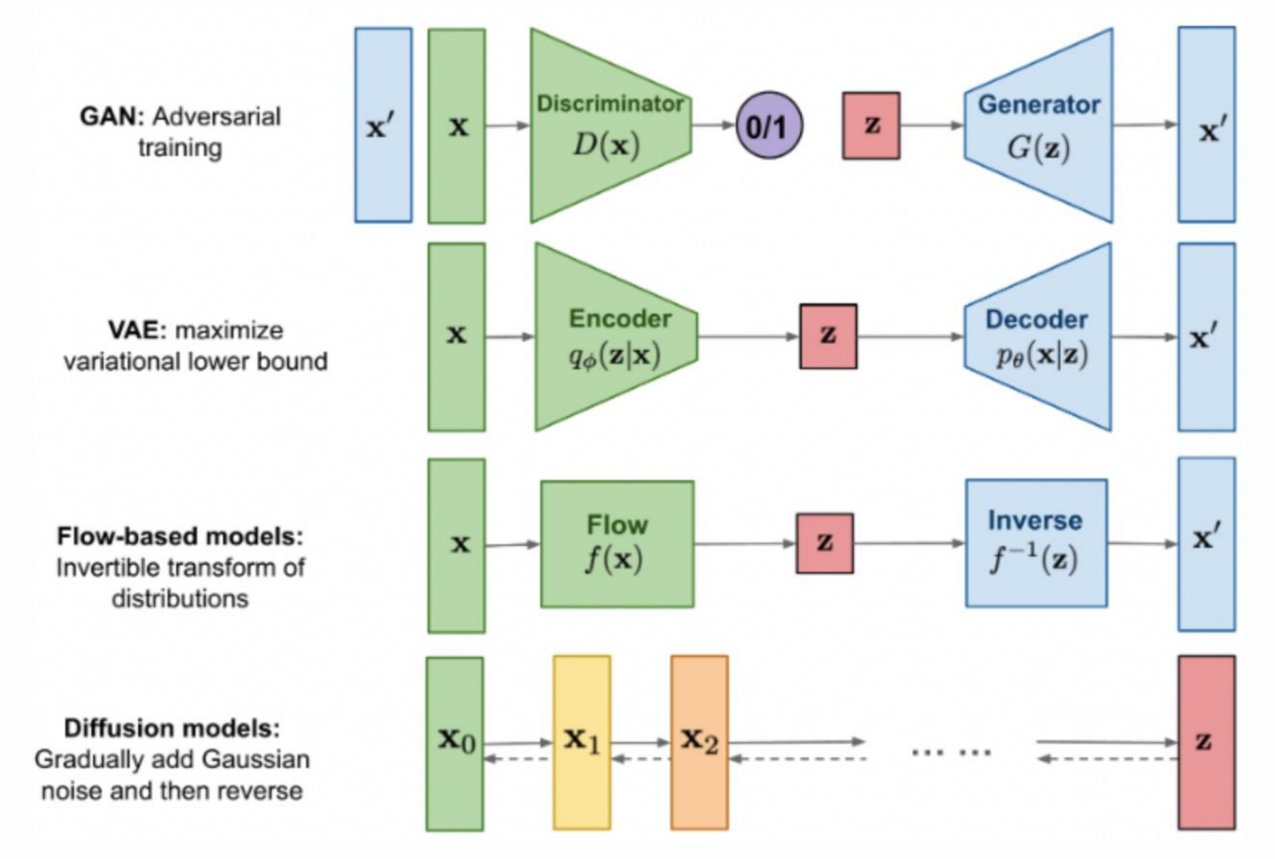
https://maciejdomagala.github.io/generative_models/2022/06/06/The-recent-rise-of-diffusion-based-models.html

Diffusion

Objective:

Generate Image **X** from latent space **Z**

Z -> X



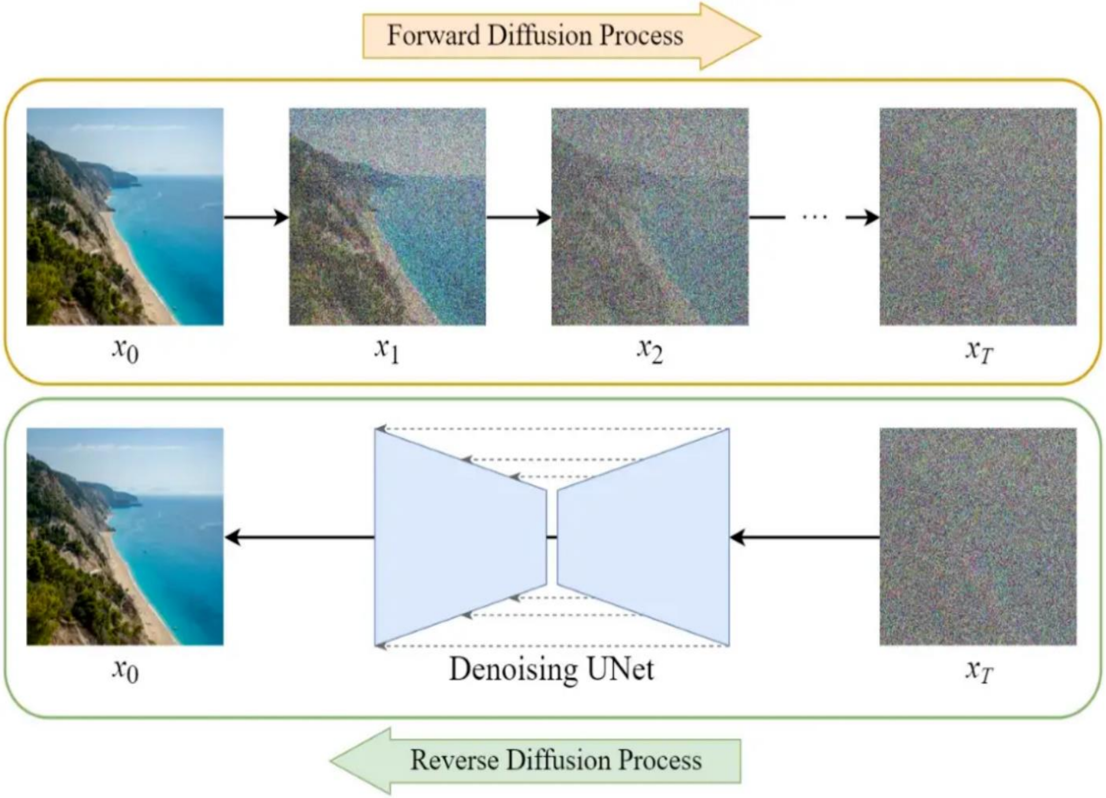
What are Diffusion Models? Lilian Weng. 2021

Diffusion

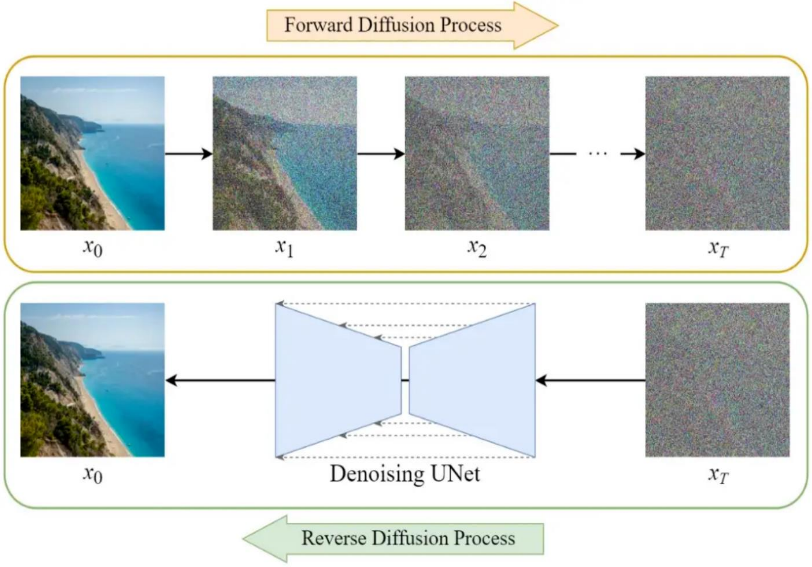
X_T == Z

Forward:
Get the ground truth

Backward:
Z -> X (generation)



Diffusion



Backward:

X_T -> X_0

How?

We remove some noise in X_T to get X_T-1
Step by step, we remove all the noise and get X_0

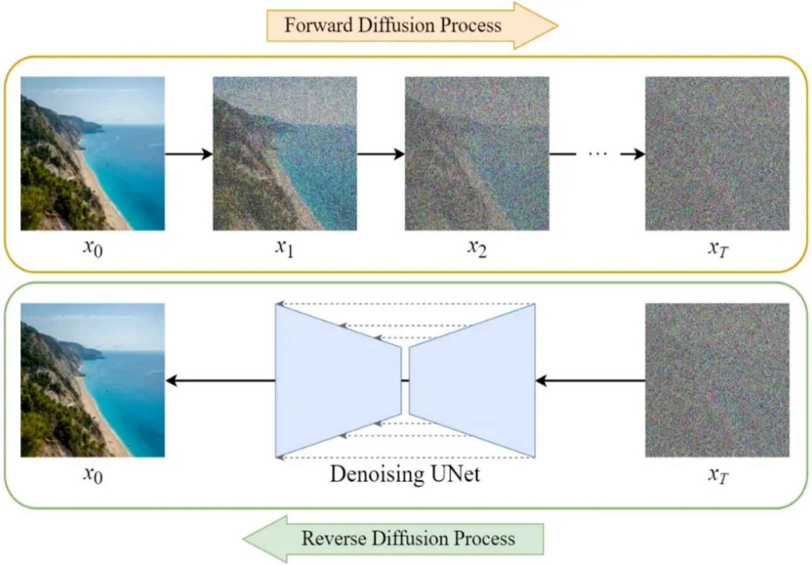
So?
We have X_T now, all we need is the noise.

So?
U-Net: input-> X_T output->noise.

Diffusion

Forward:

X_0 -> X_T



How?

Start distribution: $q(\mathbf{x_0})$

An image sample from $q(\mathbf{x_0})$: $\mathbf{x_0}$

Aim: $\mathbf{x_0} \to \mathbf{x_1} \to \mathbf{\ldots} \to \mathbf{x}_T$

Define this process: $q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)=\mathcal{N}\left(\sqrt{1-\beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}\right).$ *noising schedule* $\{\beta_t\}_{t=1}^T$

😵 we don't like step by step: re-parametrization

$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0,\left(1-\bar{\alpha}_t\right) \mathbf{I}\right)=\sqrt{\bar{\alpha}_t} \mathbf{x}_0+\sqrt{1-\bar{\alpha}_t} \cdot \epsilon$

ϵ represents Gaussian noise

$\alpha_t:=1-\beta_t$, $\bar{\alpha}_t:=\prod_{k=0}^t \alpha_k$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$

Diffusion

Backward:

X_T -> X_0

How?

We remove some noise in X_T to get X_T-1
Step by step, we remove all the noise and get X_0

So?
We have X_T now, all we need is the noise.

So?
U-Net: input-> X_T output->noise.

😵 we don't like step by step: re-parametrization

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right)=\mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0,\left(1-\bar{\alpha}_t\right)\mathbf{I}\right)=\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\cdot\epsilon$$

How to train?

Output: $\epsilon_\theta(x_t, t)$

Ground Truth: $\epsilon \sim \mathcal{N}(0, \mathbf{I})$

Loss:

$$||\epsilon-\epsilon_\theta(x_t, t)||^2=\left|\left|\epsilon-\epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\cdot\epsilon, t)\right|\right|^2$$

Diffusion

Training and inference

Algorithm 1 Training

repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Take gradient descent step on
    $\nabla_\theta \left\| \boldsymbol{\epsilon} - \mathbf{z}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
until converged

Algorithm 2 Sampling

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \mathbf{z}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

Denoising Diffusion Probabilistic Models. Jonathan Ho, Ajay Jain, Pieter Abbeel. 2020

GLIDE

How to add the additional information to Diffusion? : Guided Diffusion

Backward process of DDPM $p_\theta(x_{t-1}|x_t)$

Classifier-guided Diffusion $p_{\theta,\phi}(x_{t-1}|x_t, y) = Z \cdot p_\theta(x_{t-1}|x_t) \cdot p_\phi(y|x_{t-1})$

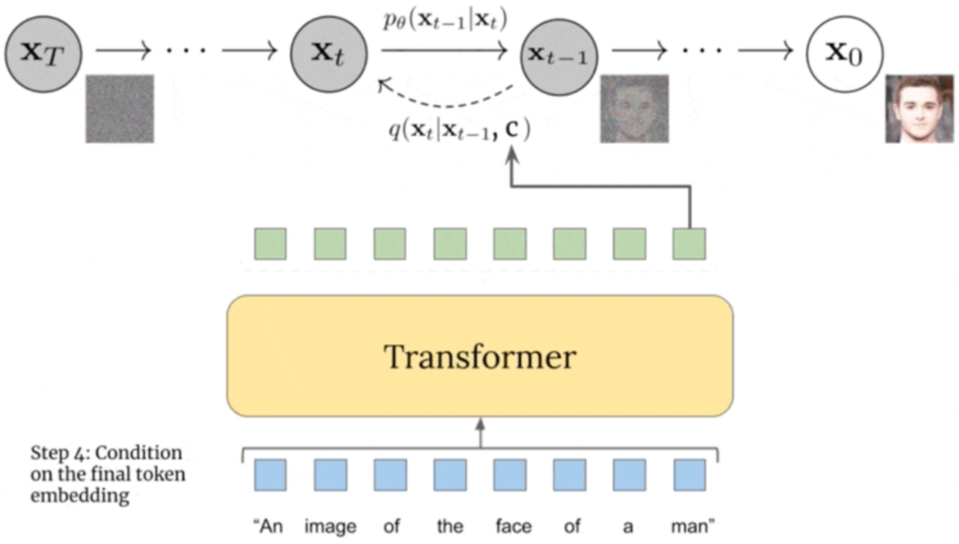
Classifier-free Diffusion $\hat{\epsilon}_\theta(\mathbf{x}_t, t \mid y) = \epsilon_\theta(\mathbf{x}_t, t \mid \emptyset) + s \cdot (\epsilon_\theta(\mathbf{x}_t, t \mid y) - \epsilon_\theta(\mathbf{x}_t, t \mid \emptyset))$

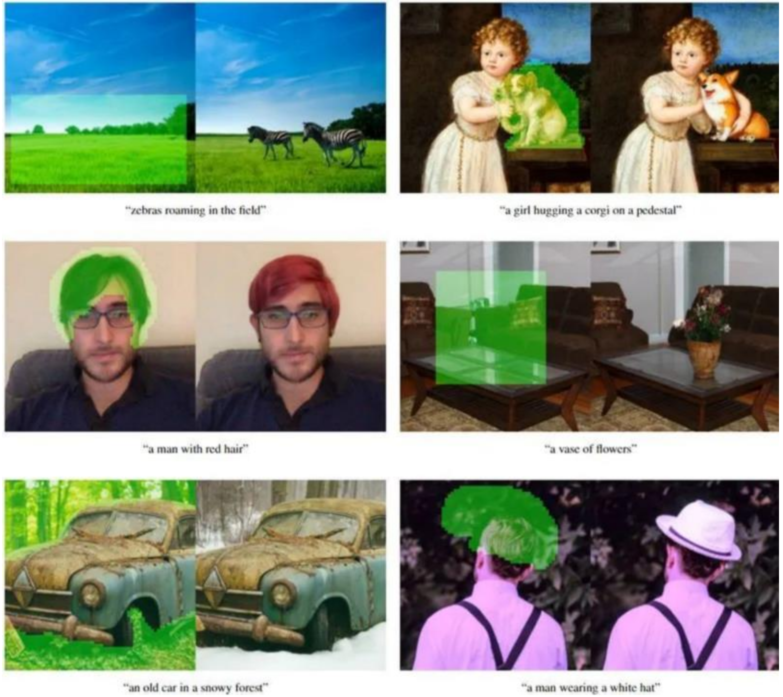
GLIDE $\hat{\epsilon}_\theta(x_t|Caption) = \epsilon_\theta(x_t) + s \cdot (\epsilon_\theta(x_t, Caption) - \epsilon_\theta(x_t))$

(more data; larger model; more GPUs)

Diffusion models beat gans on image synthesis. Dhariwal, P. and Nichol, A. 2021.
GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. Alex Nichol et al. 2021

GLIDE

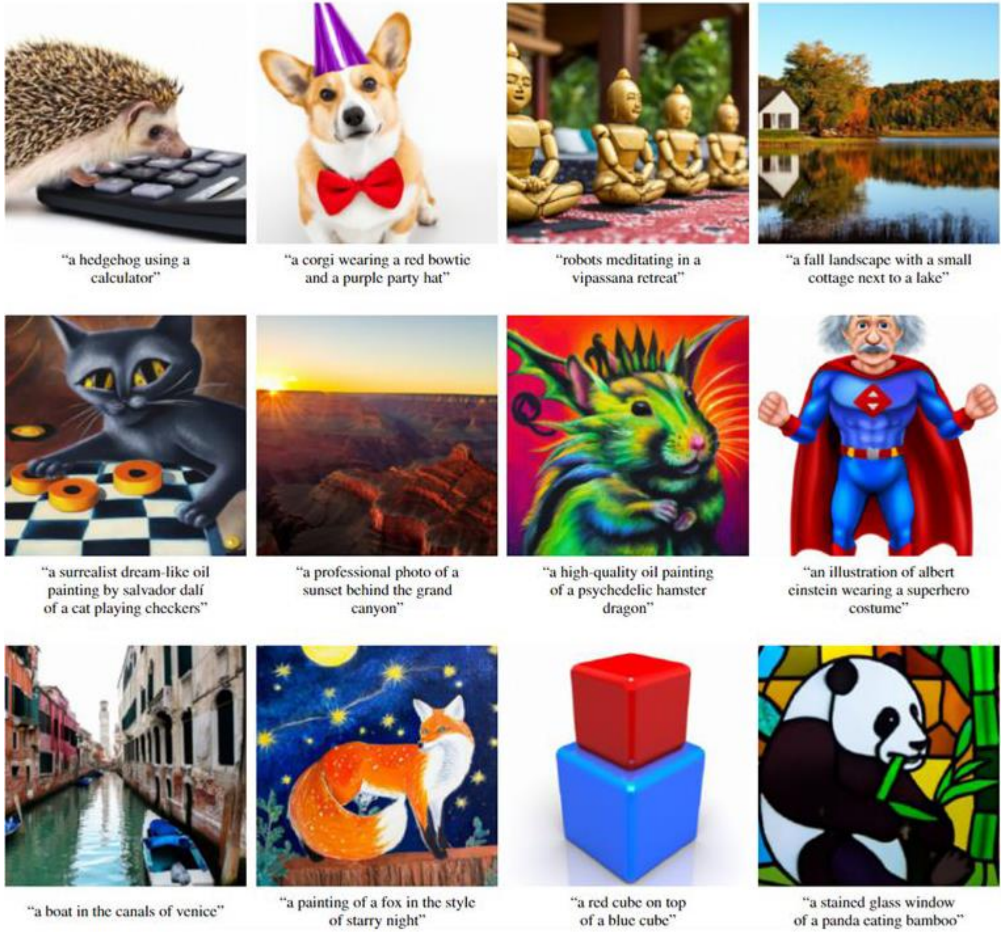




caption + mask area + super resolution

Diffusion models beat gans on image synthesis. Dhariwal, P. and Nichol, A. 2021.
GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. Alex Nichol et al. 2021

GLIDE

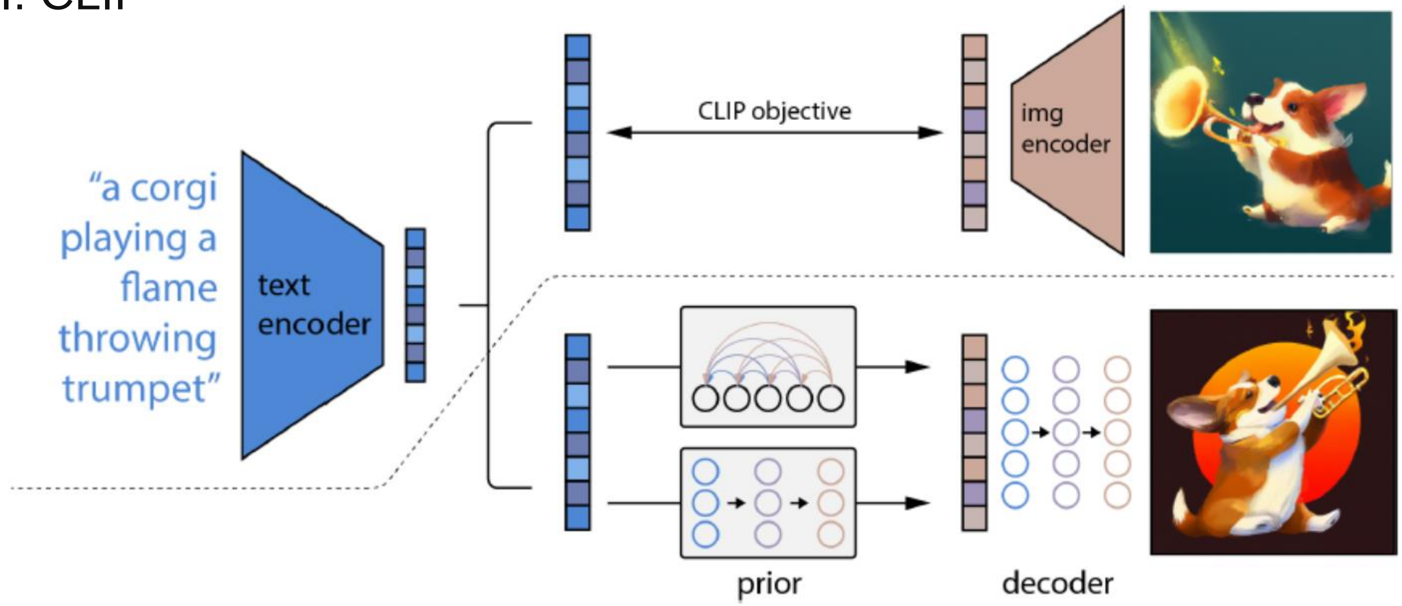


Diffusion models beat gans on image synthesis. Dhariwal, P. and Nichol, A. 2021.
GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. Alex Nichol et al. 2021

DALL·E 2

CLIP + GLIDE

Training stage I: CLIP



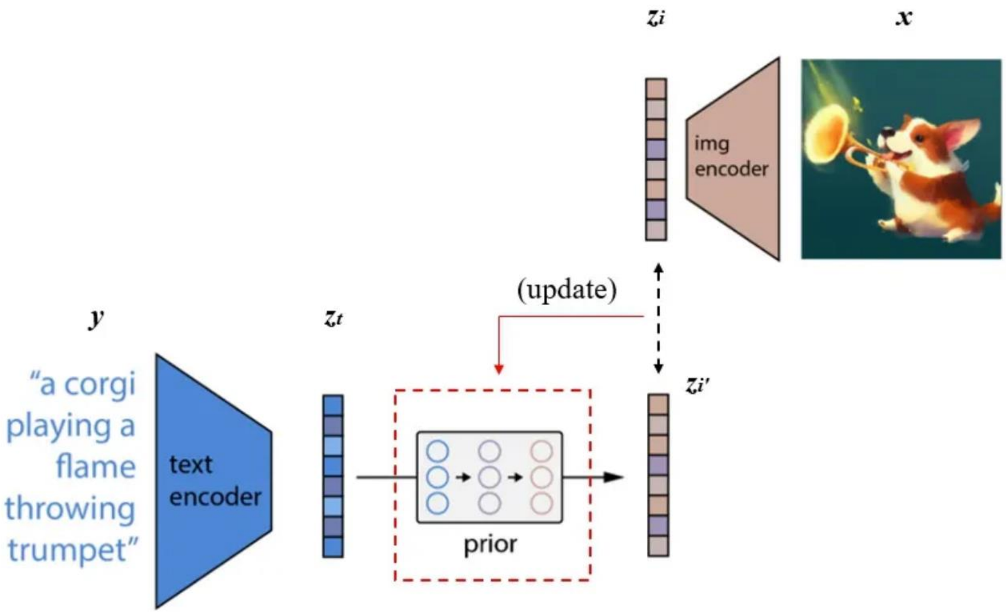
Hierarchical Text-Conditional Image Generation with CLIP Latents. Aditya Ramesh et al. 2022

DALL·E 2

CLIP + GLIDE

Training stage II:

Prior (latent space)



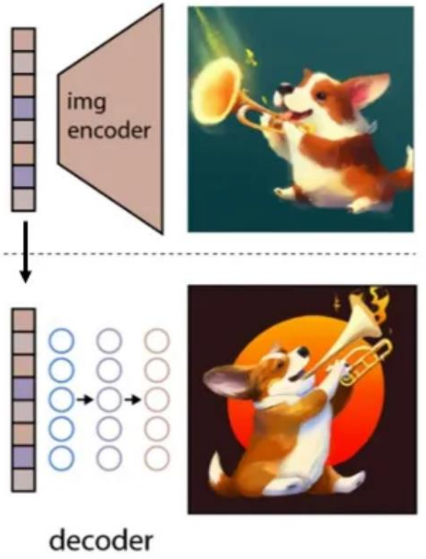
Hierarchical Text-Conditional Image Generation with CLIP Latents. Aditya Ramesh et al. 2022

DALL·E 2

CLIP + GLIDE

Training stage III:

Decoder (GLIDE)

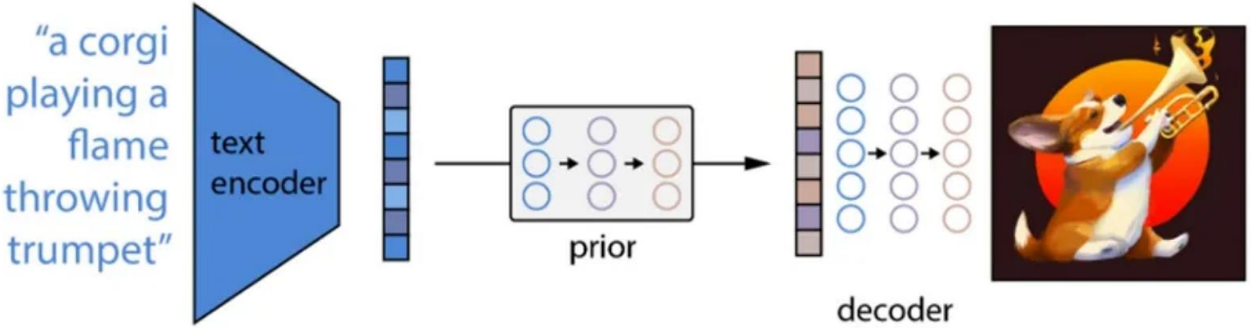


Hierarchical Text-Conditional Image Generation with CLIP Latents. Aditya Ramesh et al. 2022

DALL·E 2

CLIP + GLIDE

Inference



Hierarchical Text-Conditional Image Generation with CLIP Latents. Aditya Ramesh et al. 2022

DALL·E 2

Hierarchical Text-Conditional Image Generation with CLIP Latents. Aditya Ramesh et al. 2022

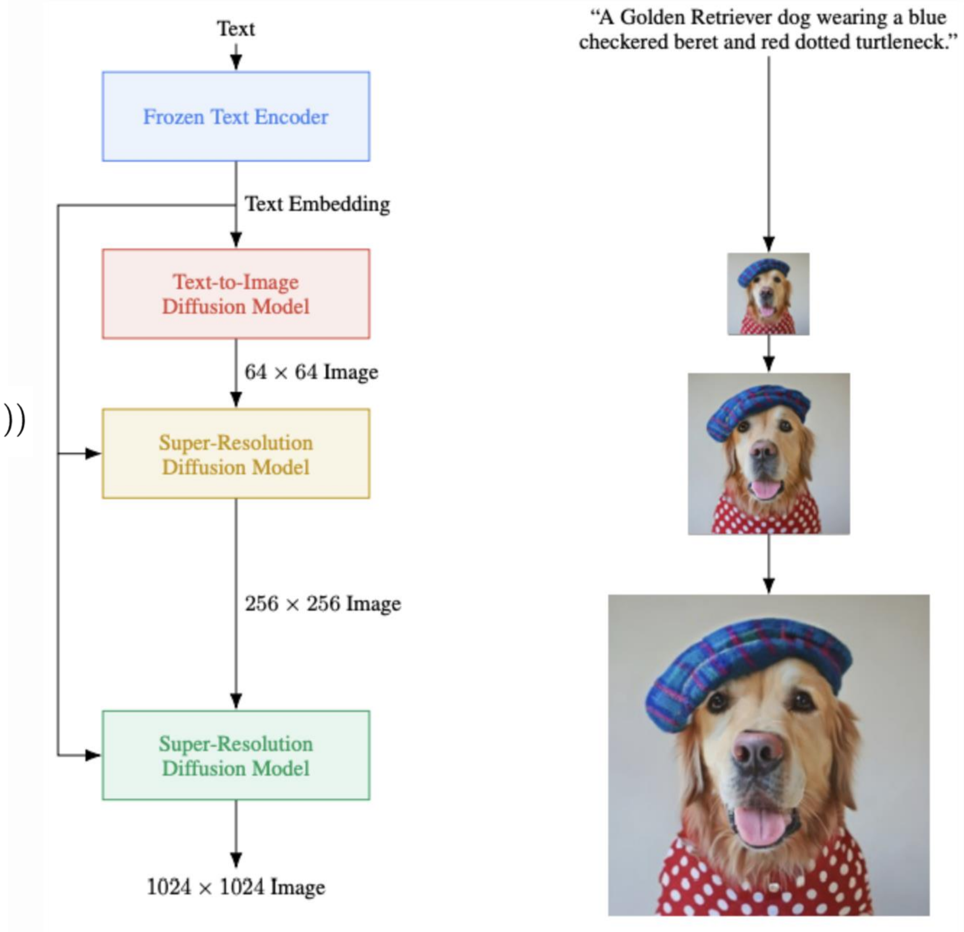
Imagen

A pretrained, frozen T5-XXL model

GLIDE (dynamic)

$$\hat{\epsilon}_\theta(x_t|Caption) = \epsilon_\theta(x_t) + s \cdot (\epsilon_\theta(x_t, Caption) - \epsilon_\theta(x_t))$$

Efficient U-net



Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Chitwan Saharia et al. 2022

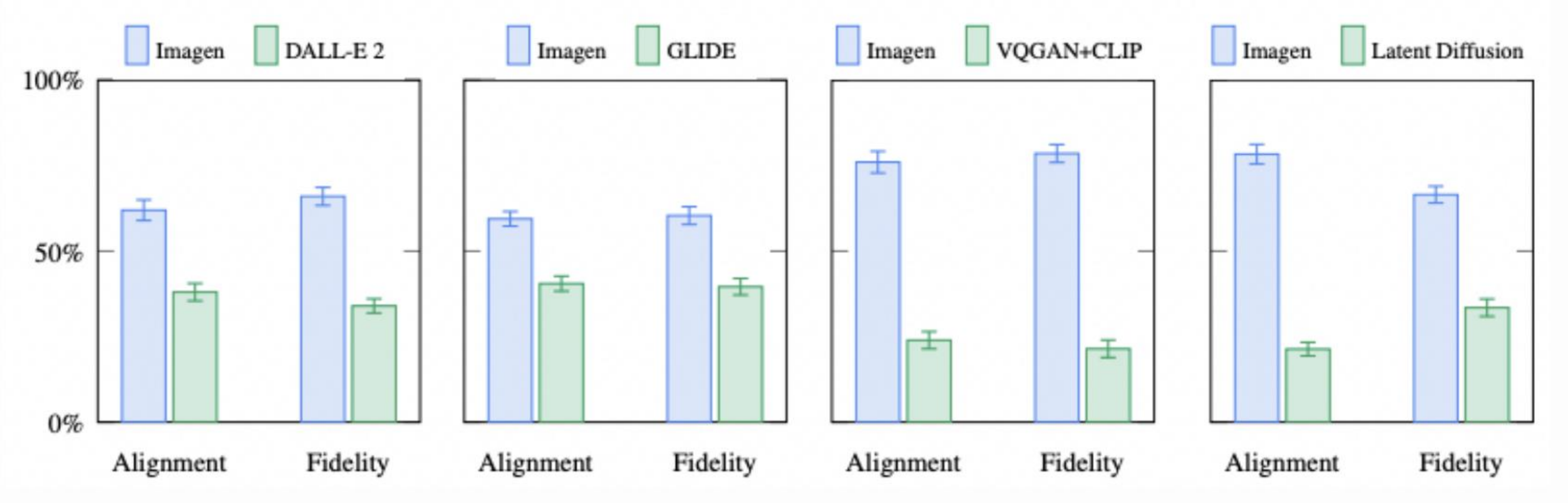
Imagen

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**

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Chitwan Saharia et al. 2022

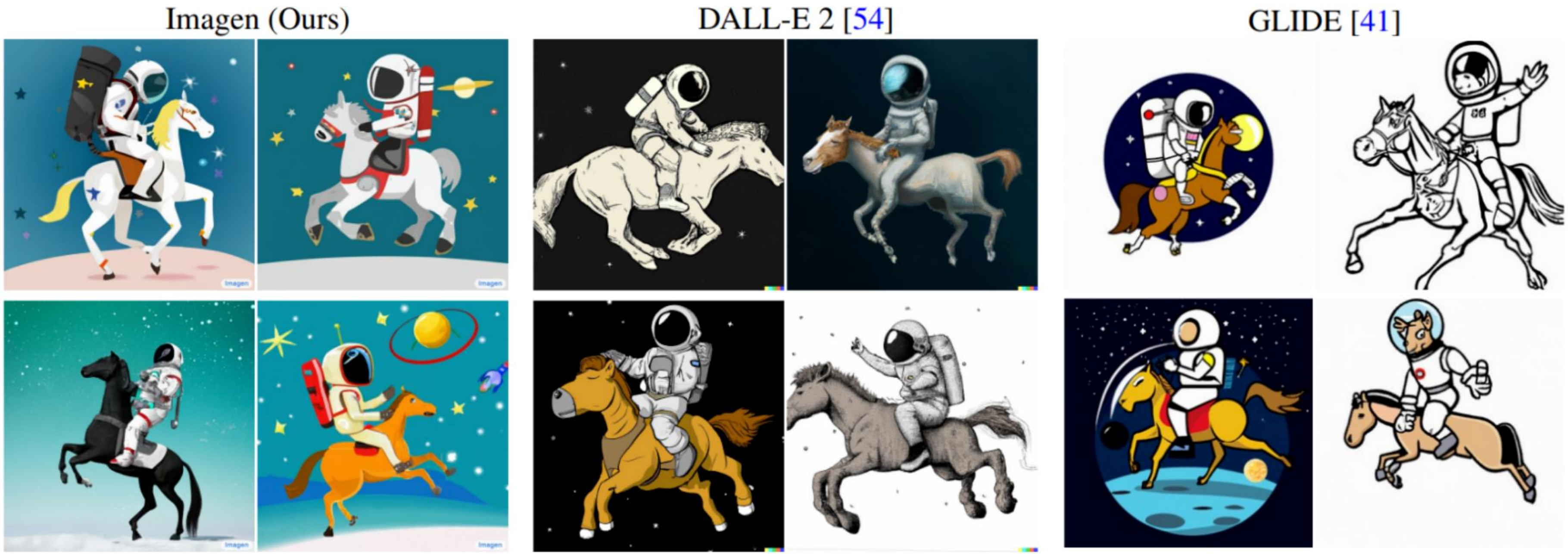
Imagen

DrawBench



Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Chitwan Saharia et al. 2022

Imagen



A horse riding an astronaut.

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Chitwan Saharia et al. 2022

Future work

Latent Space: Stable Diffusion

Fine-grained Control: ControlNet, T2I-Adapter and Composer

Inversion: DreamBooth

Applications: Make-A-Video; Make-A-Story; Magic3D

Text-to-image Diffusion Model in Generative AI: A Survey
https://github.com/heejkoo/Awesome-Diffusion-Models