

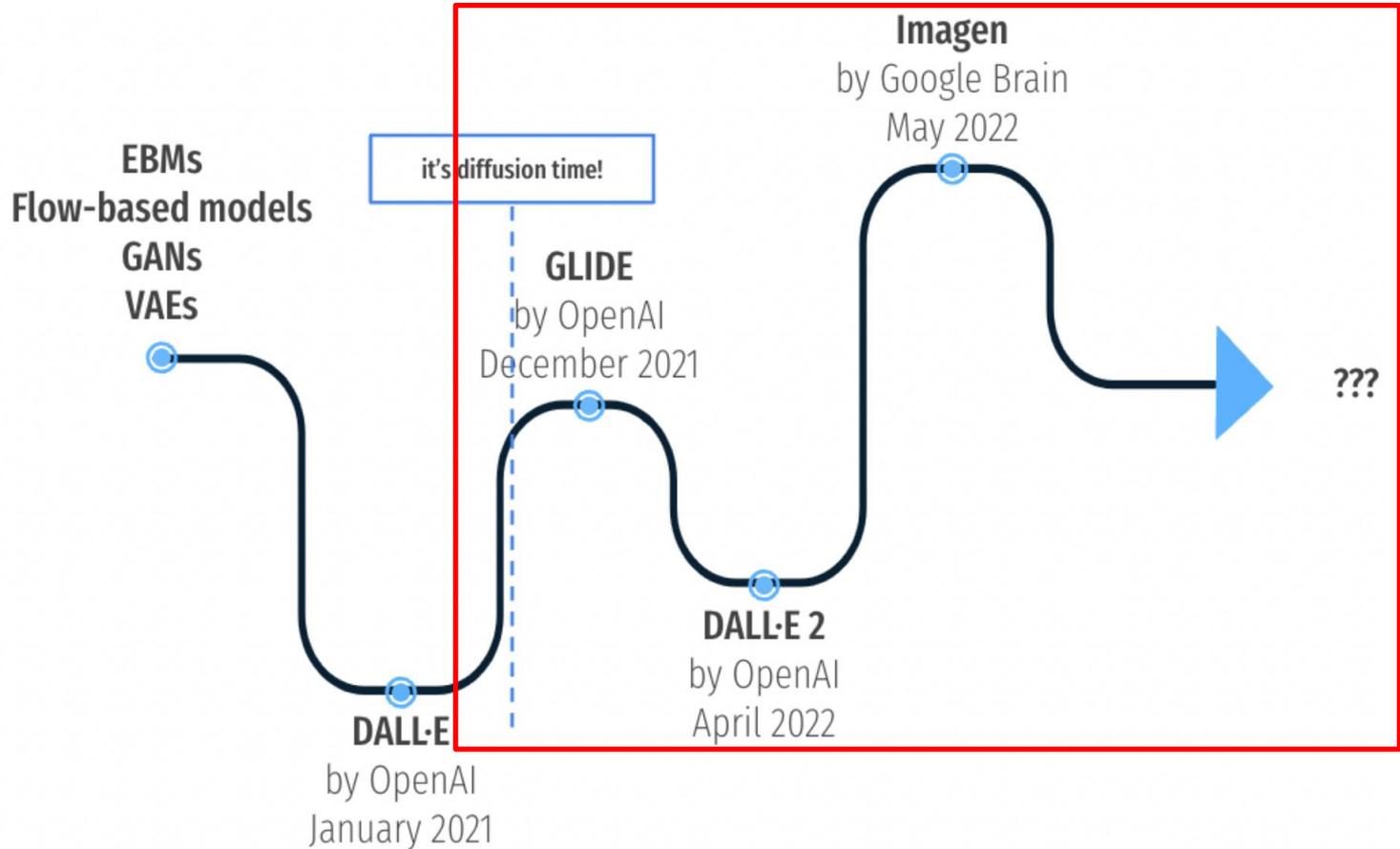
Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Google Research, Brain Team

Xinjie Li

03/15/2023

Timeline

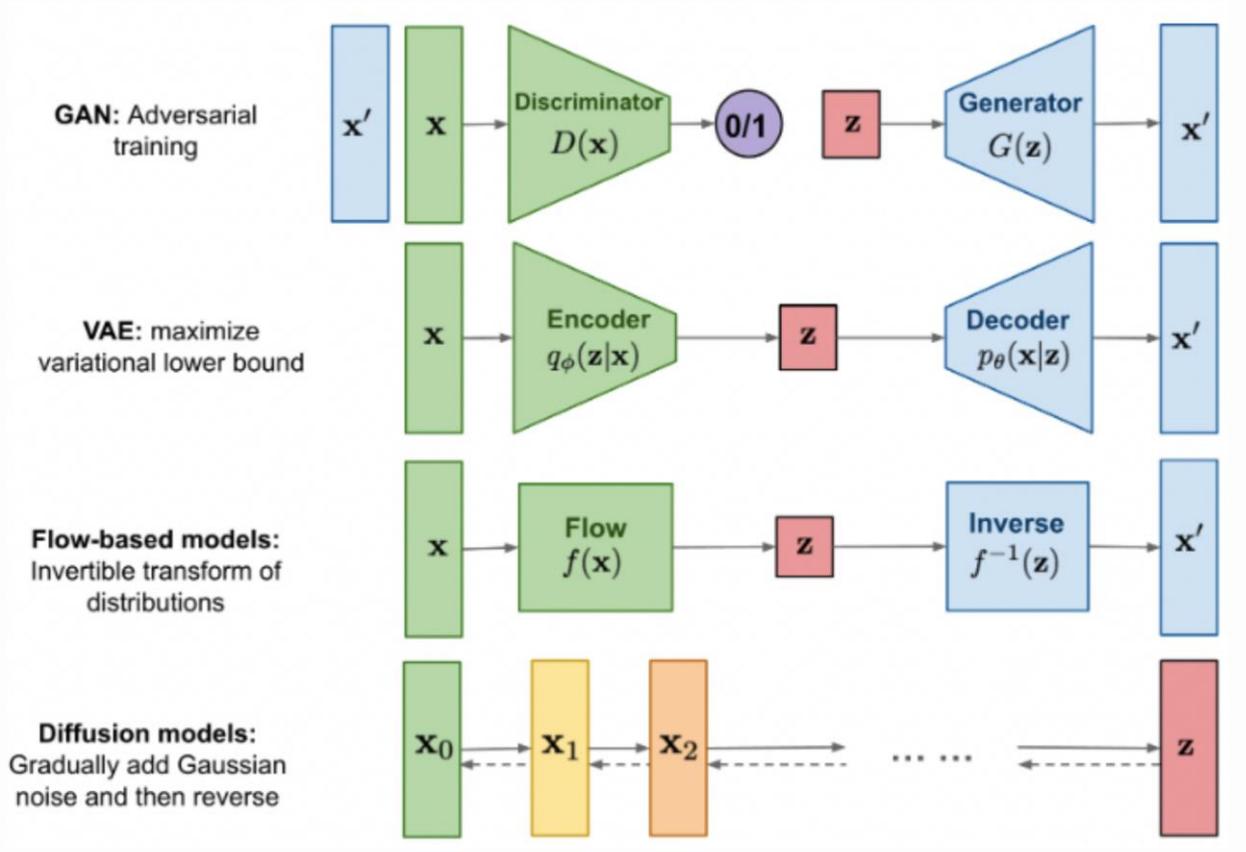


Diffusion

Objective:

Generate Image X
from latent space Z

$Z \rightarrow X$

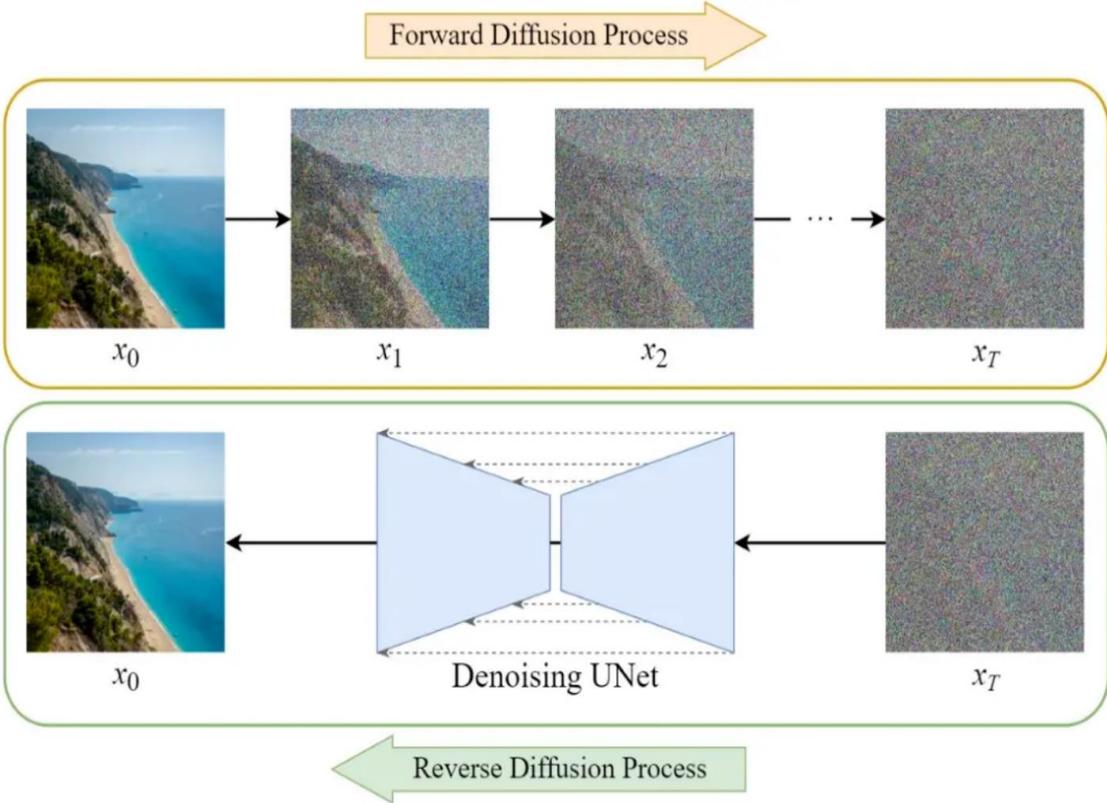


Diffusion

$$X_T == Z$$

Forward:
Get the ground truth

Backward:
 $Z \rightarrow X$ (generation)



Diffusion

Backward:

$X_T \rightarrow X_0$

How?

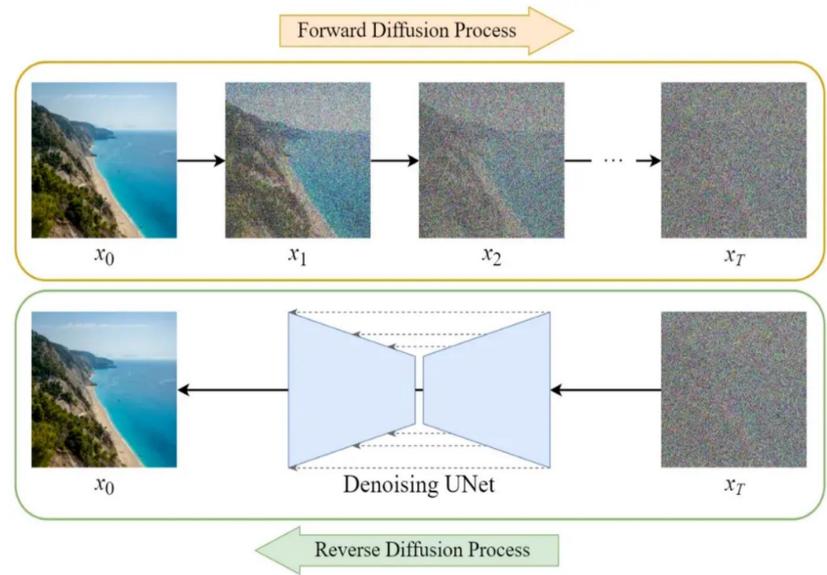
We remove some noise in X_T to get X_{T-1}
Step by step, we remove all the noise and get X_0

So?

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

So?

U-Net: input $\rightarrow X_T$ output \rightarrow noise.



Diffusion

Forward:

$x_0 \rightarrow x_T$

How?

Start distribution: $q(\mathbf{x}_0)$

An image sample from $q(\mathbf{x}_0) : \mathbf{x}_0$

Aim: $\mathbf{x}_0 \rightarrow \mathbf{x}_1 \rightarrow \dots \rightarrow \mathbf{x}_T$

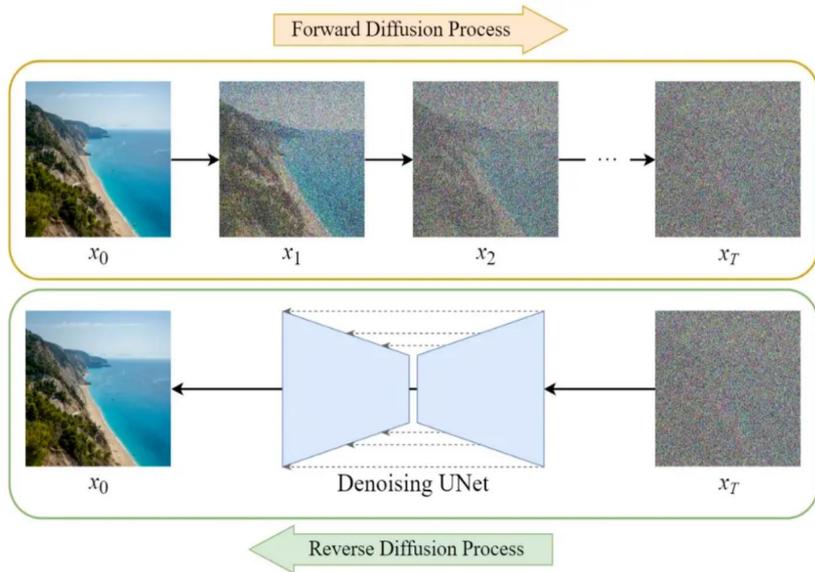
Define this process: $q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I})$. *noising schedule* $\{\beta_t\}_{t=1}^T$

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

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

ϵ represents Gaussian noise

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



Diffusion

Backward:

$X_T \rightarrow X_0$

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 $\rightarrow X_T$ output \rightarrow noise.

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

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

How to train?

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

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

Loss:

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

Diffusion

Training and inference

Algorithm 1 Training

- 1: **repeat**
 - 2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
 - 3: $t \sim \text{Uniform}(\{1, \dots, T\})$
 - 4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
 - 5: Take gradient descent step on
$$\nabla_{\theta} \|\boldsymbol{\epsilon} - \mathbf{z}_{\theta}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t)\|^2$$
 - 6: **until** converged
-

Algorithm 2 Sampling

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

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

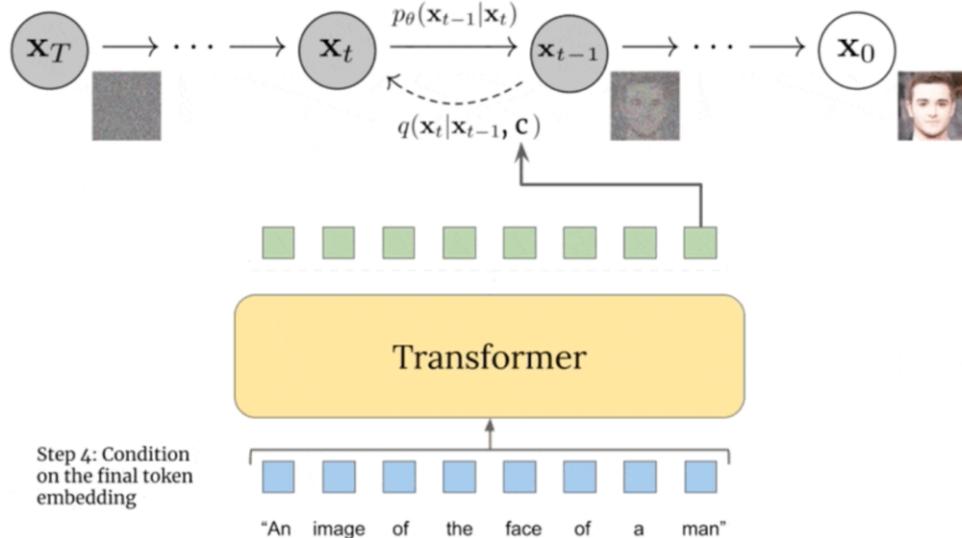
GLIDE $\hat{\epsilon}_{\theta}(x_t | \text{Caption}) = \epsilon_{\theta}(x_t) + s \cdot (\epsilon_{\theta}(x_t, \text{Caption}) - \epsilon_{\theta}(x_t))$

(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



"zebras roaming in the field"



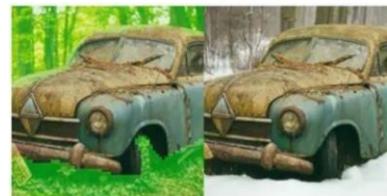
"a girl hugging a corgi on a pedestal"



"a man with red hair"



"a vase of flowers"



"an old car in a snowy forest"



"a man wearing a white hat"

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



"a hedgehog using a calculator"



"a corgi wearing a red bowtie and a purple party hat"



"robots meditating in a vipassana retreat"



"a fall landscape with a small cottage next to a lake"



"a surreal dream-like oil painting by salvador dali of a cat playing checkers"



"a professional photo of a sunset behind the grand canyon"



"a high-quality oil painting of a psychedelic hamster dragon"



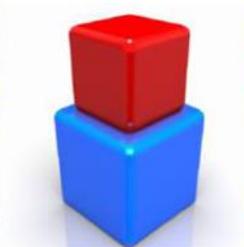
"an illustration of albert einstein wearing a superhero costume"



"a boat in the canals of venice"



"a painting of a fox in the style of starry night"



"a red cube on top of a blue cube"



"a stained glass window of a panda eating bamboo"

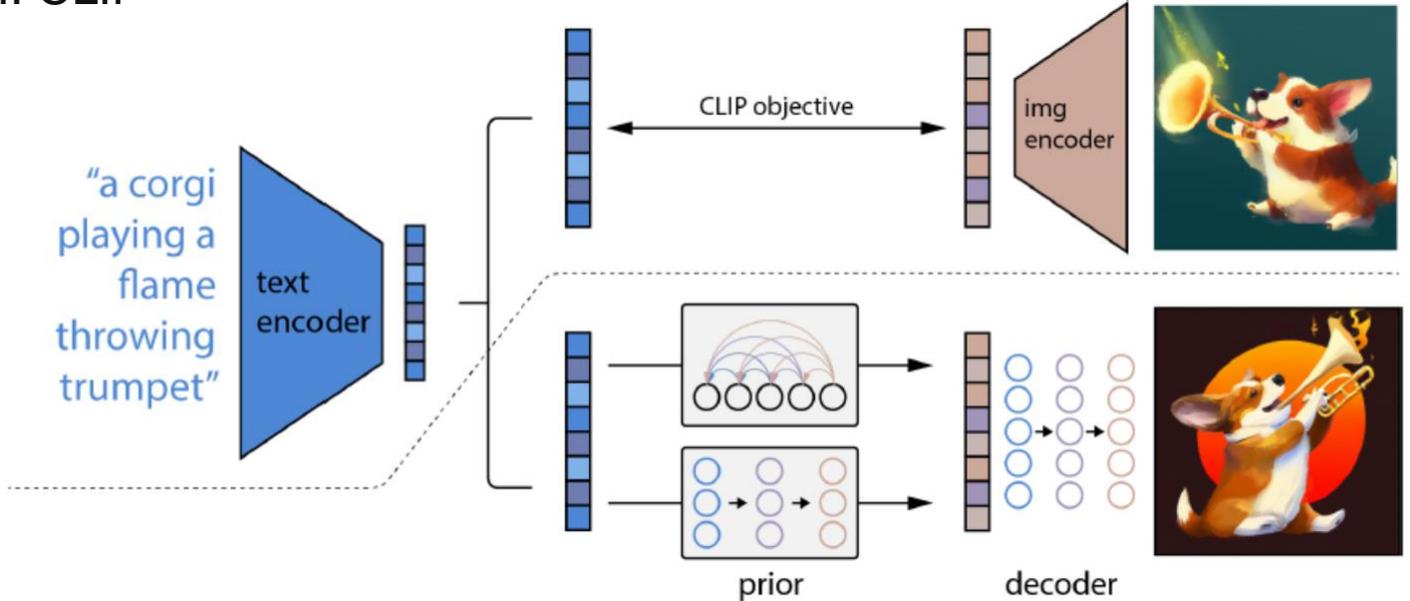
Diffusion models beat gans on image synthesis. Dhariwal, P. and Nichol, A. 2021.

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

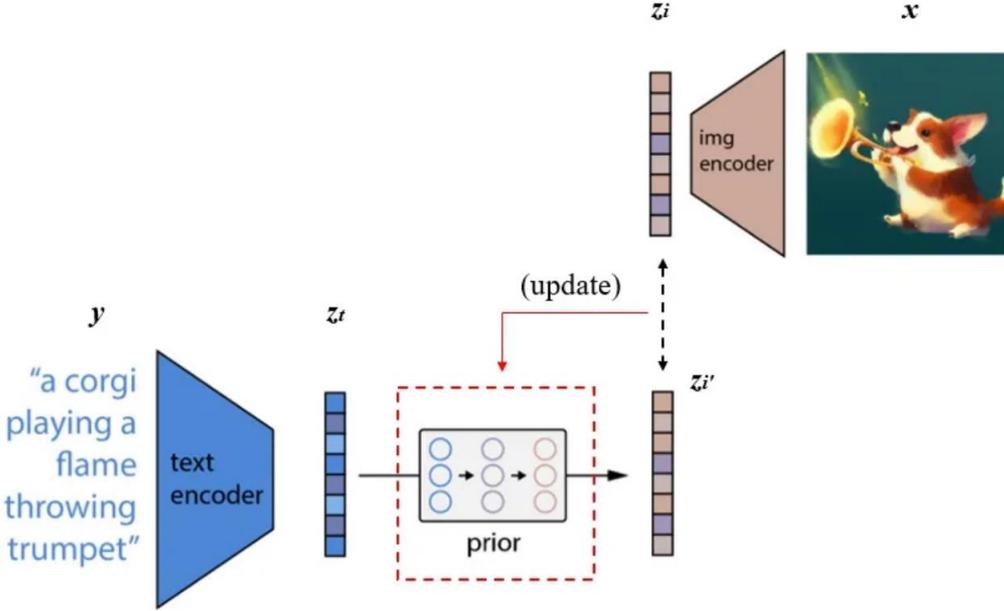


DALL-E 2

CLIP + GLIDE

Training stage II:

Prior (latent space)

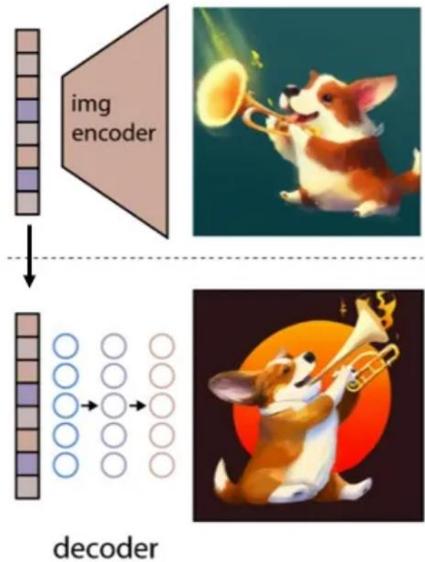


DALL-E 2

CLIP + GLIDE

Training stage III:

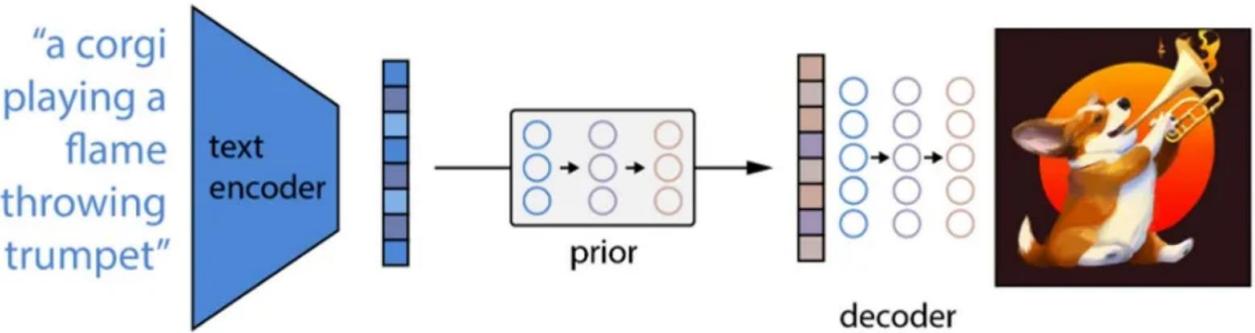
Decoder (GLIDE)



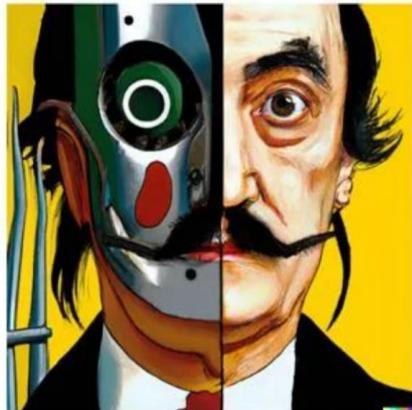
DALL·E 2

CLIP + GLIDE

Inference



DALL-E 2



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 hand holding a small green plant 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

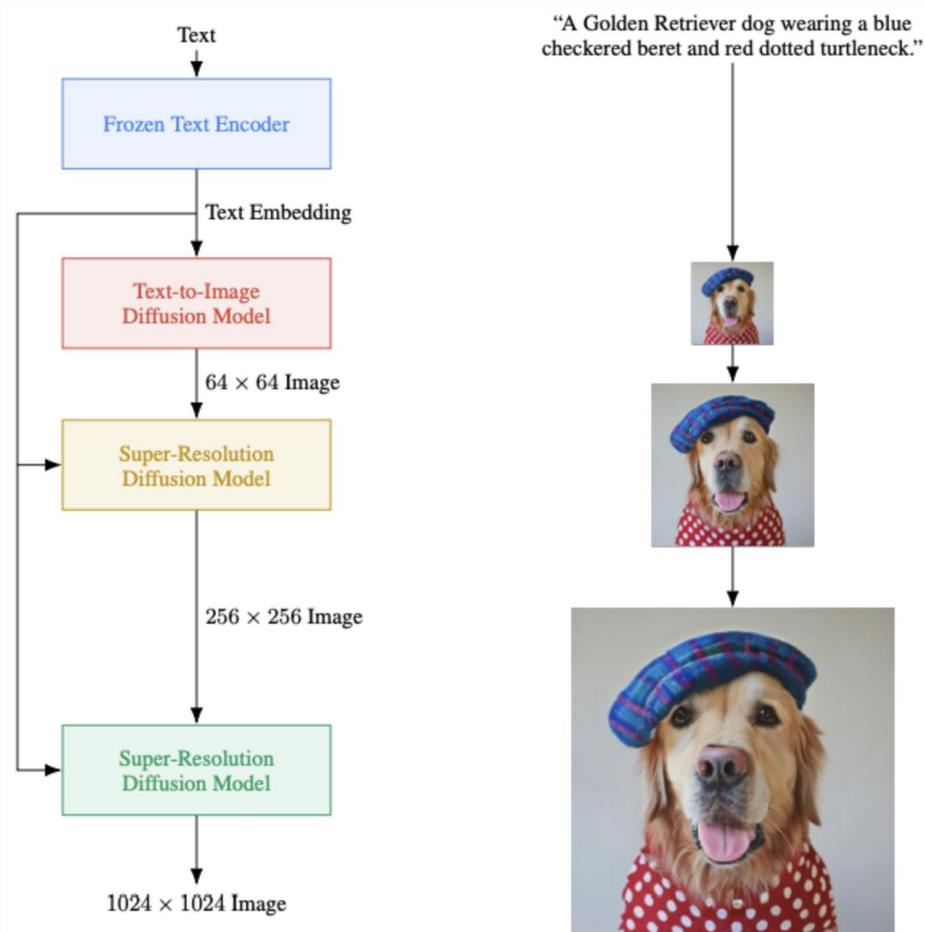
Imagen

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

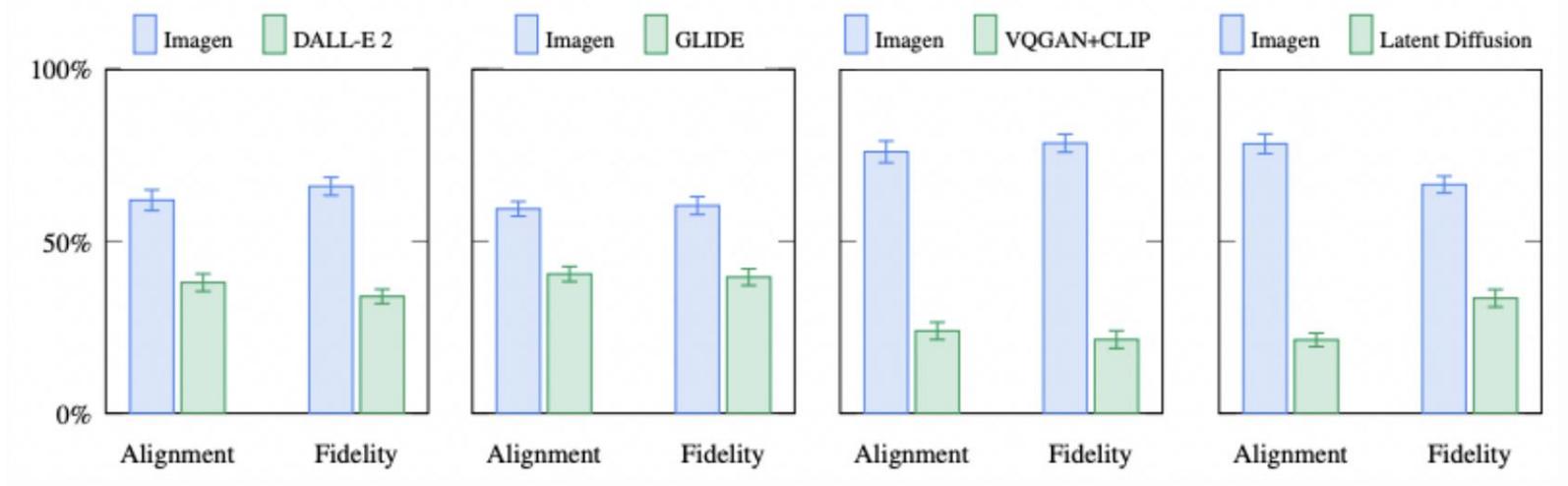


Imagen

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

Imagen

DrawBench

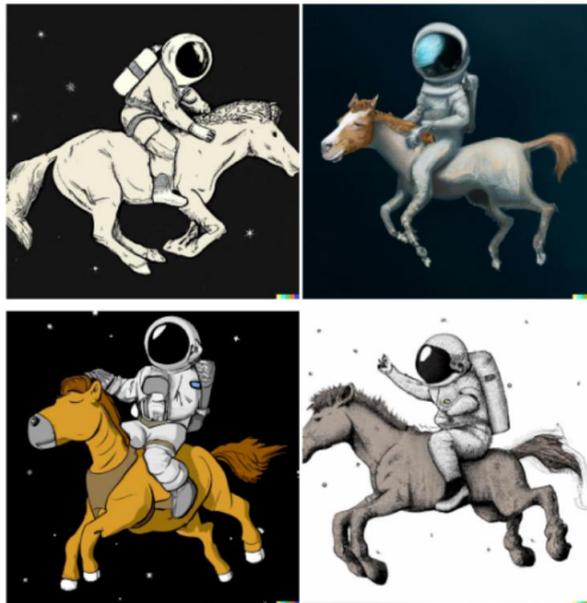


Imagen

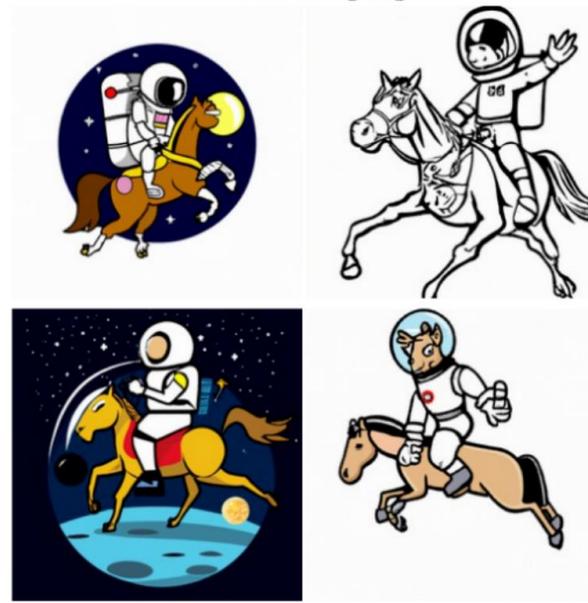
Imagen (Ours)



DALL-E 2 [54]



GLIDE [41]



A horse riding an astronaut.

Future work

Latent Space: Stable Diffusion

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

Inversion: DreamBooth

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