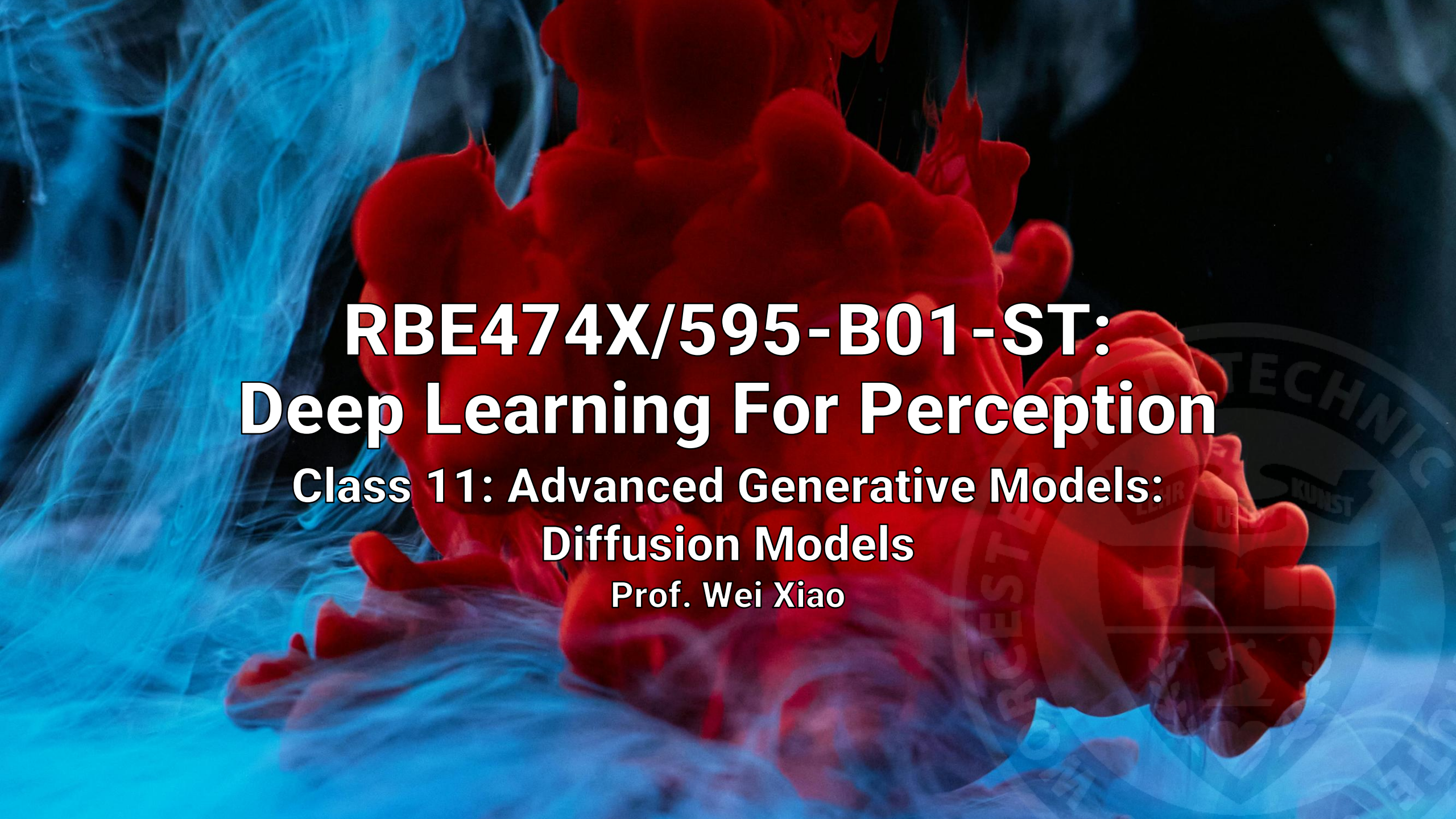
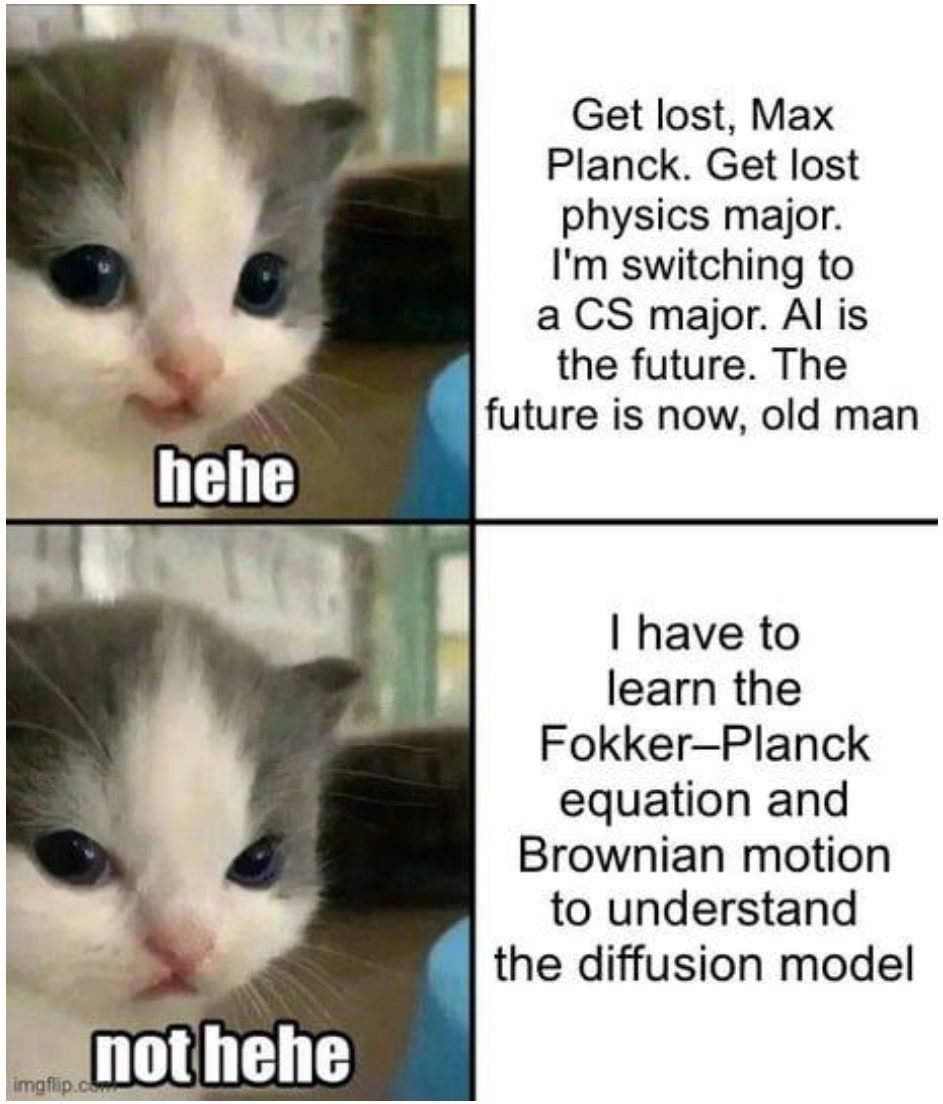


RBE474X/595-B01-ST: Deep Learning For Perception

Class 11: Advanced Generative Models: Diffusion Models

Prof. Wei Xiao

Please Ask Questions!



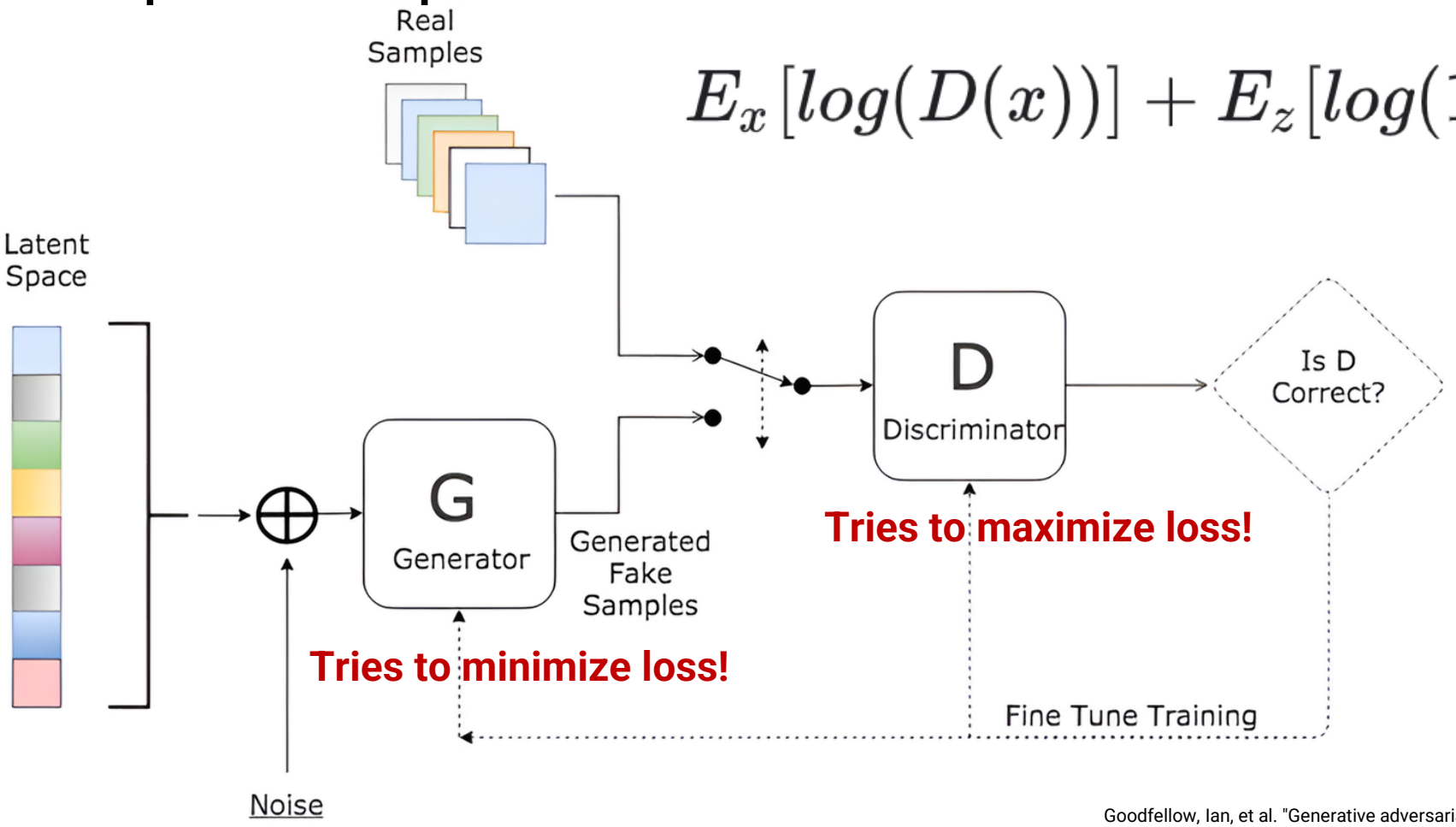
Today's lecture is Math heavy! Slow me down if needed!

Why diffusion?

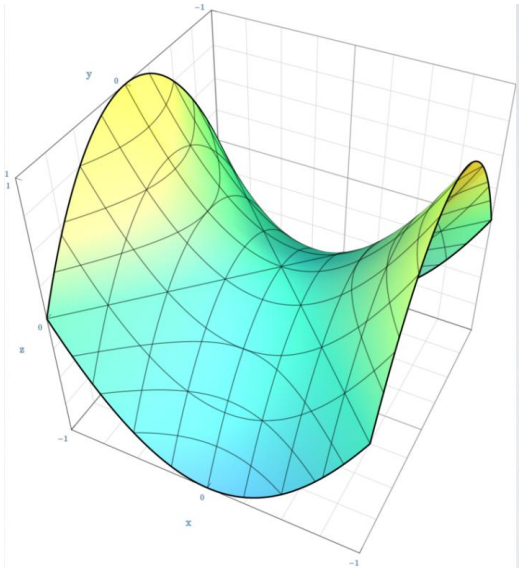
Recall GANs

Use A Competition Strategy!
Latent Space for interpolation

$$E_x[log(D(x))] + E_z[log(1 - D(G(z)))]$$



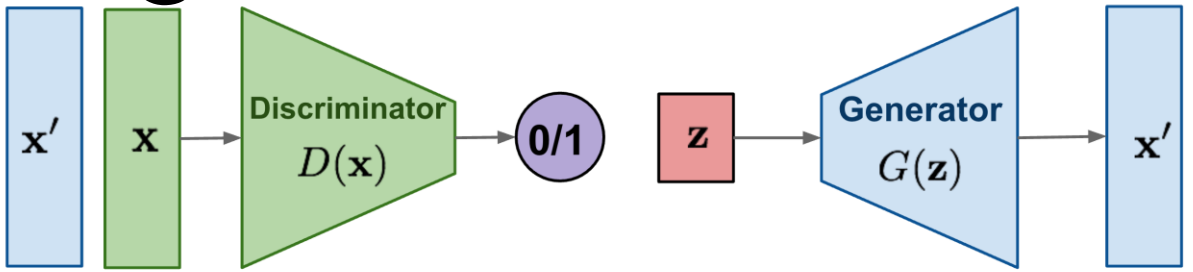
Saddle point problem!



Goodfellow, Ian, et al. "Generative adversarial networks." Communications of the ACM 63.11 (2020): 139-144.

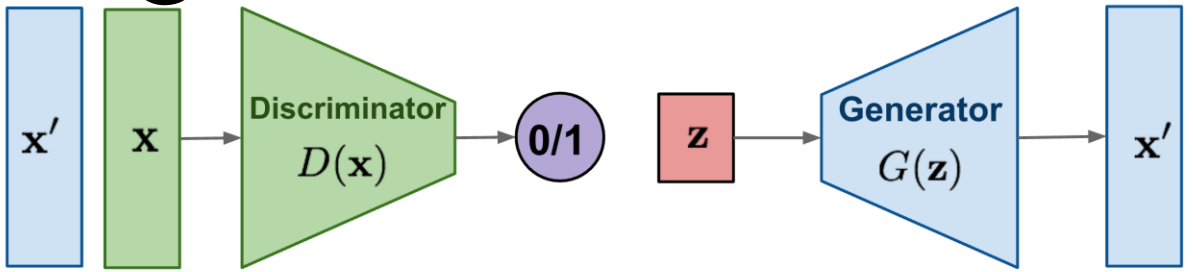
Generating Images

GAN: Adversarial training

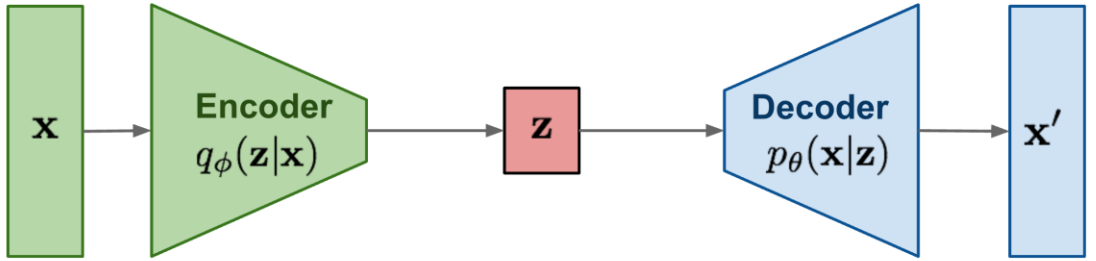


Generating Images

GAN: Adversarial training

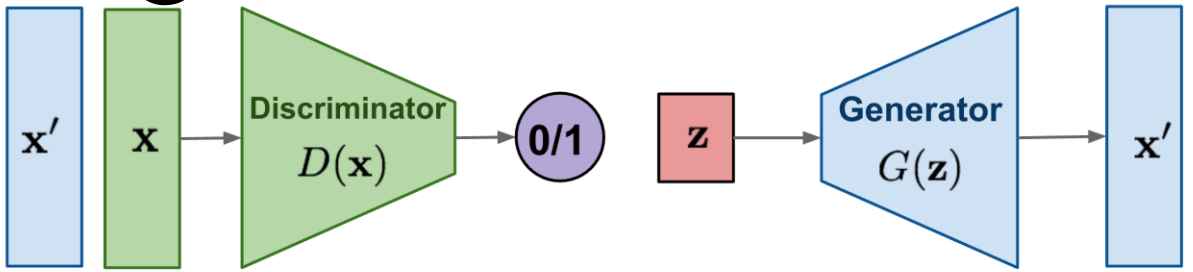


VAE: maximize variational lower bound

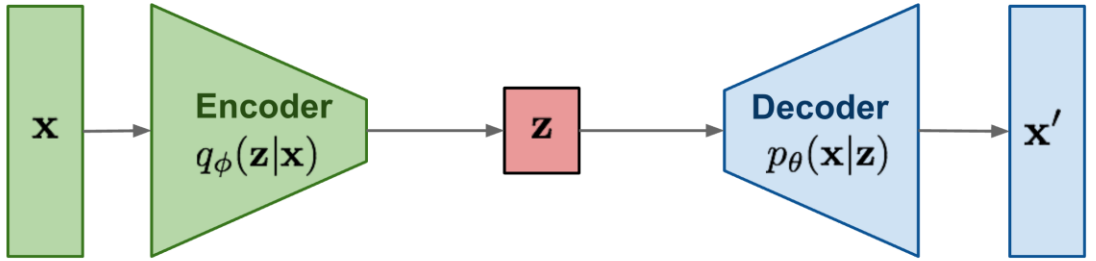


Generating Images

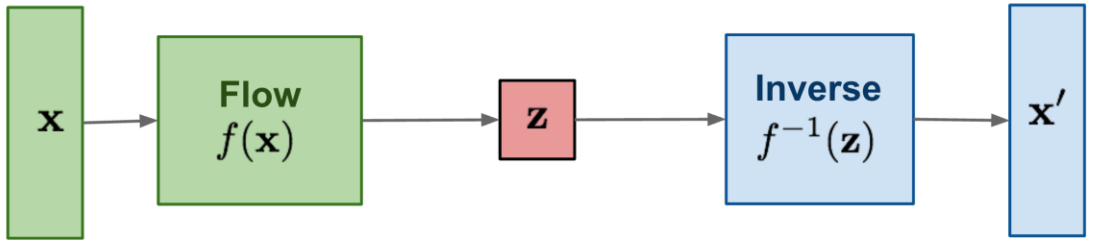
GAN: Adversarial training



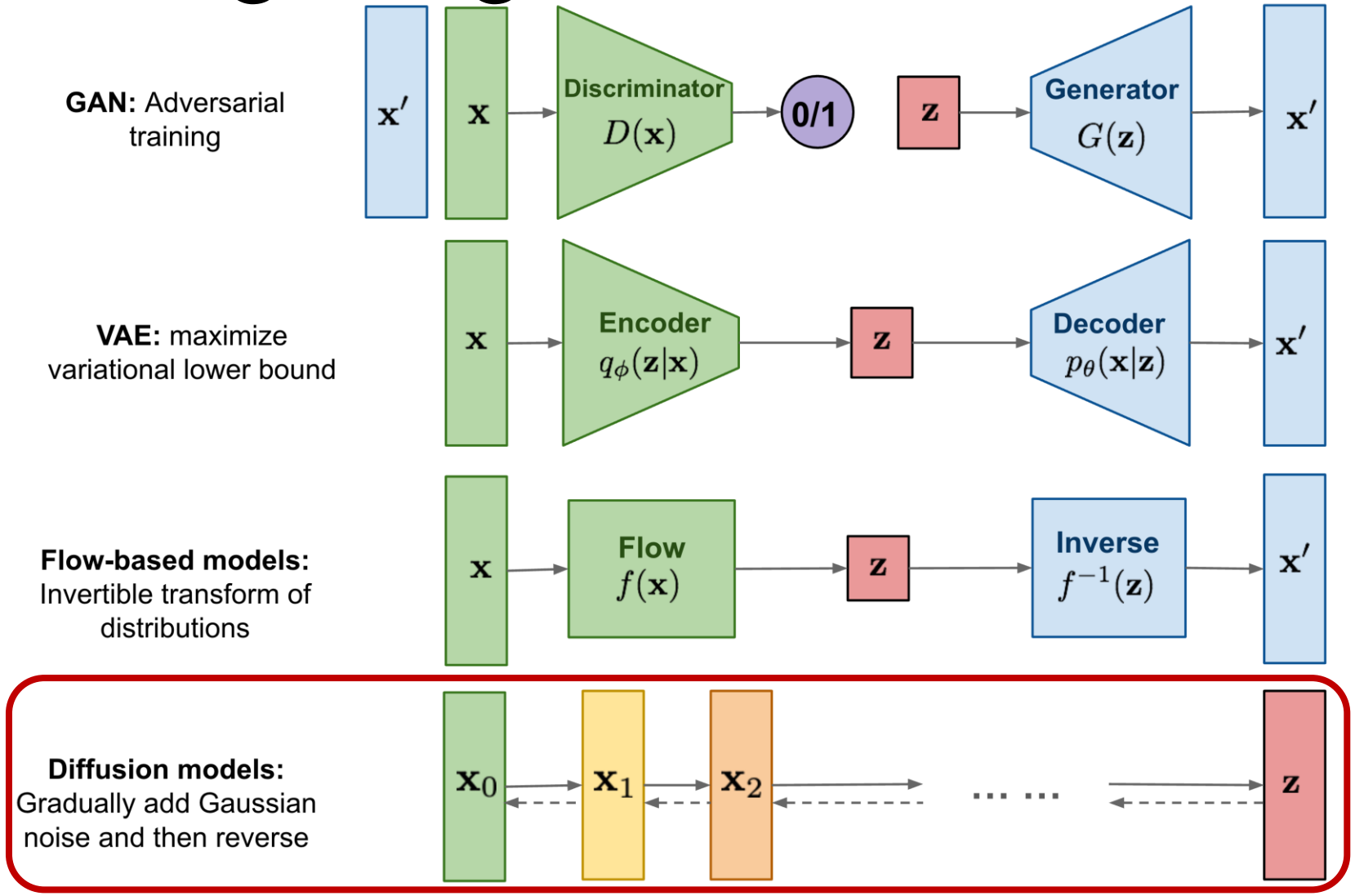
VAE: maximize variational lower bound



Flow-based models: Invertible transform of distributions

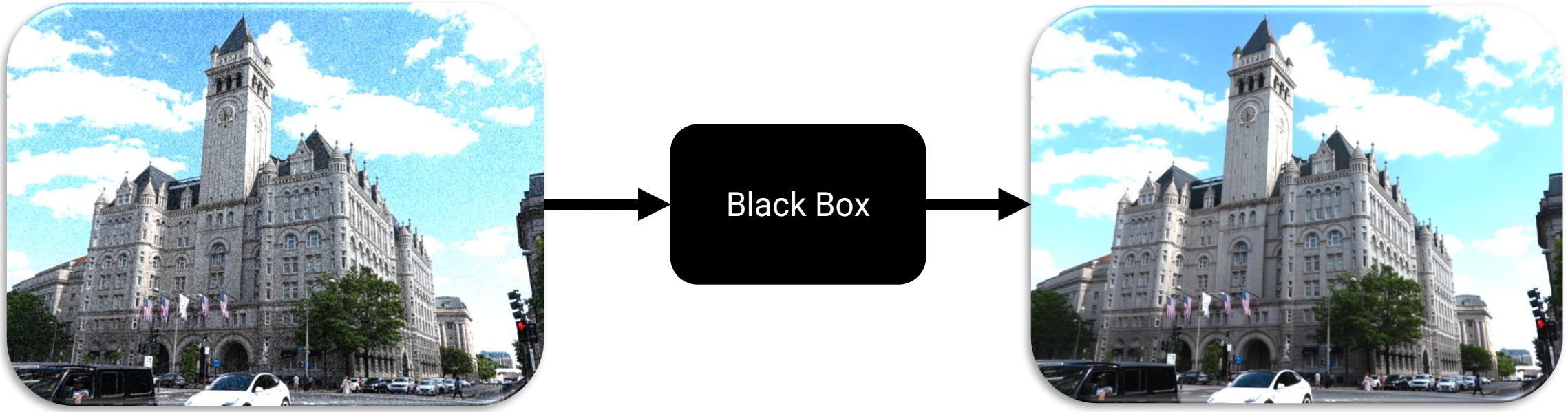


Generating Images

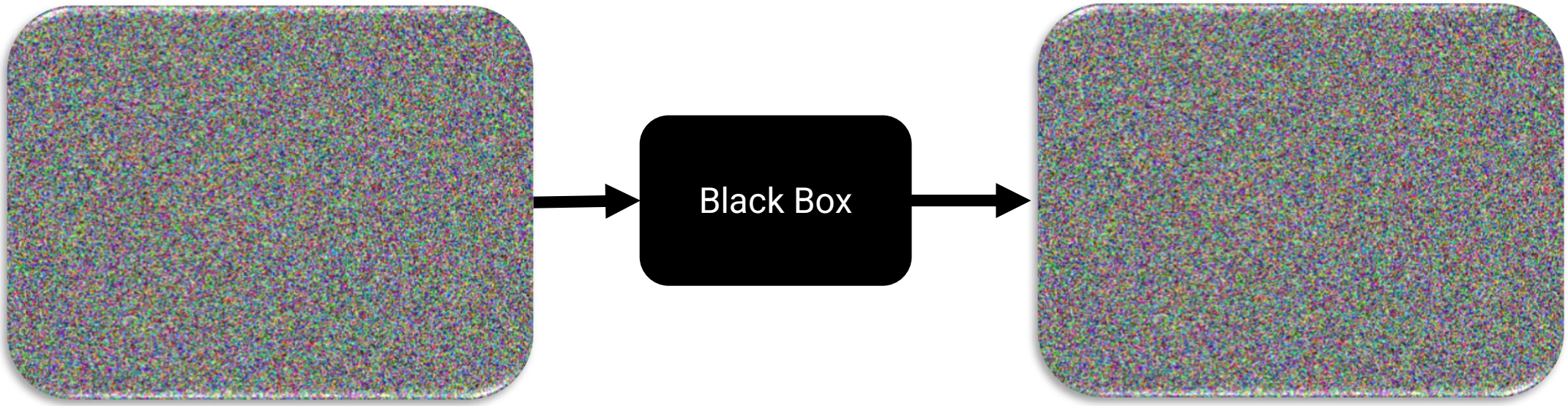


Intuition

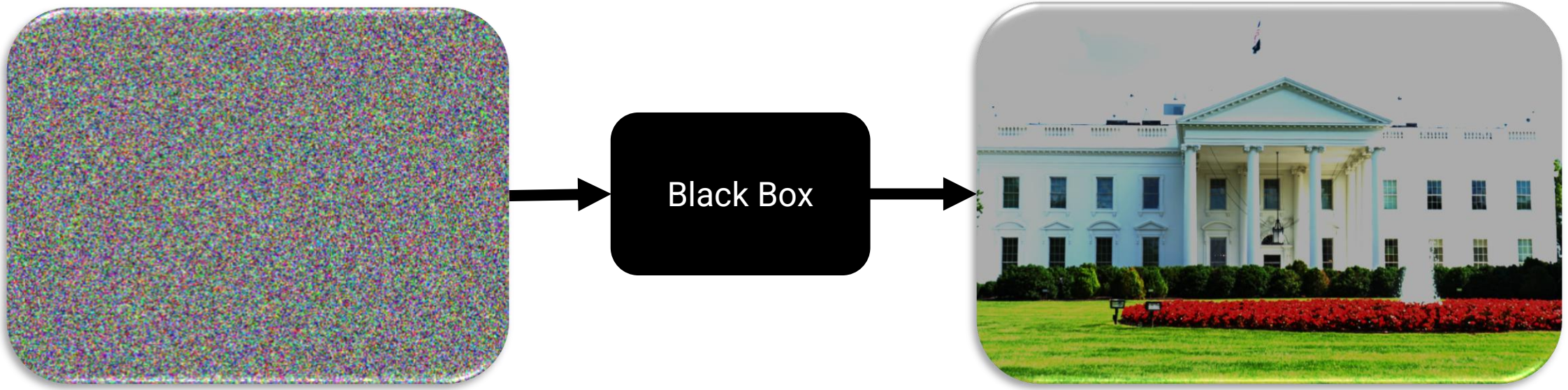
Image denoising



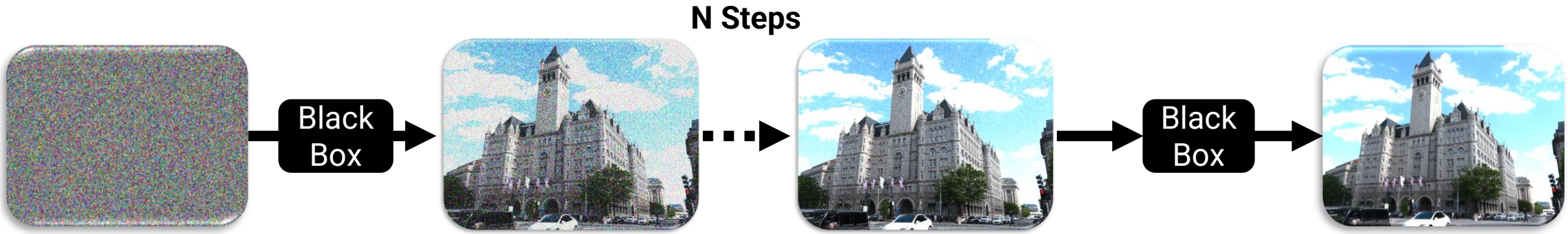
Noise as input?

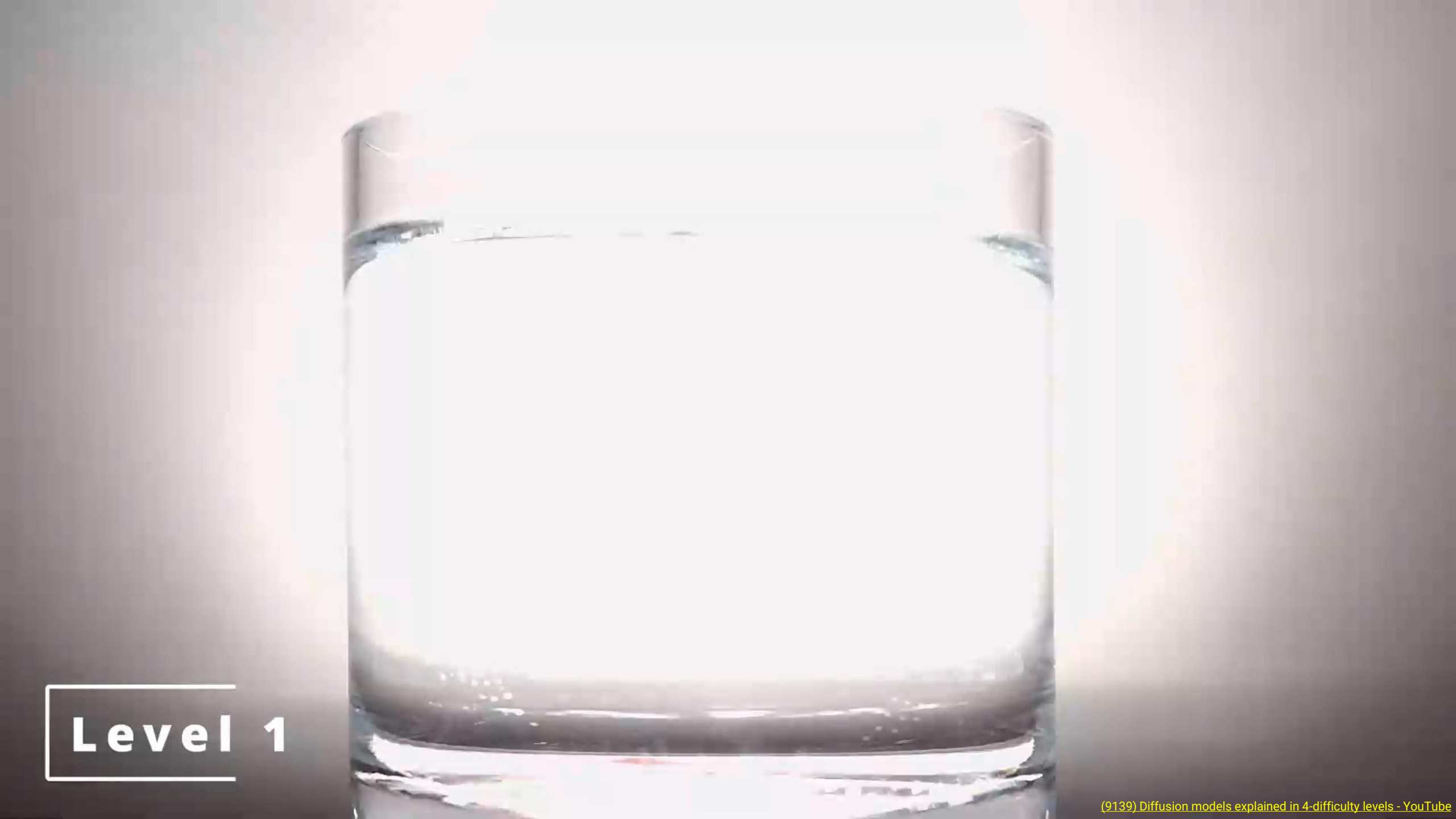


Noise as input?



Diffusion Model



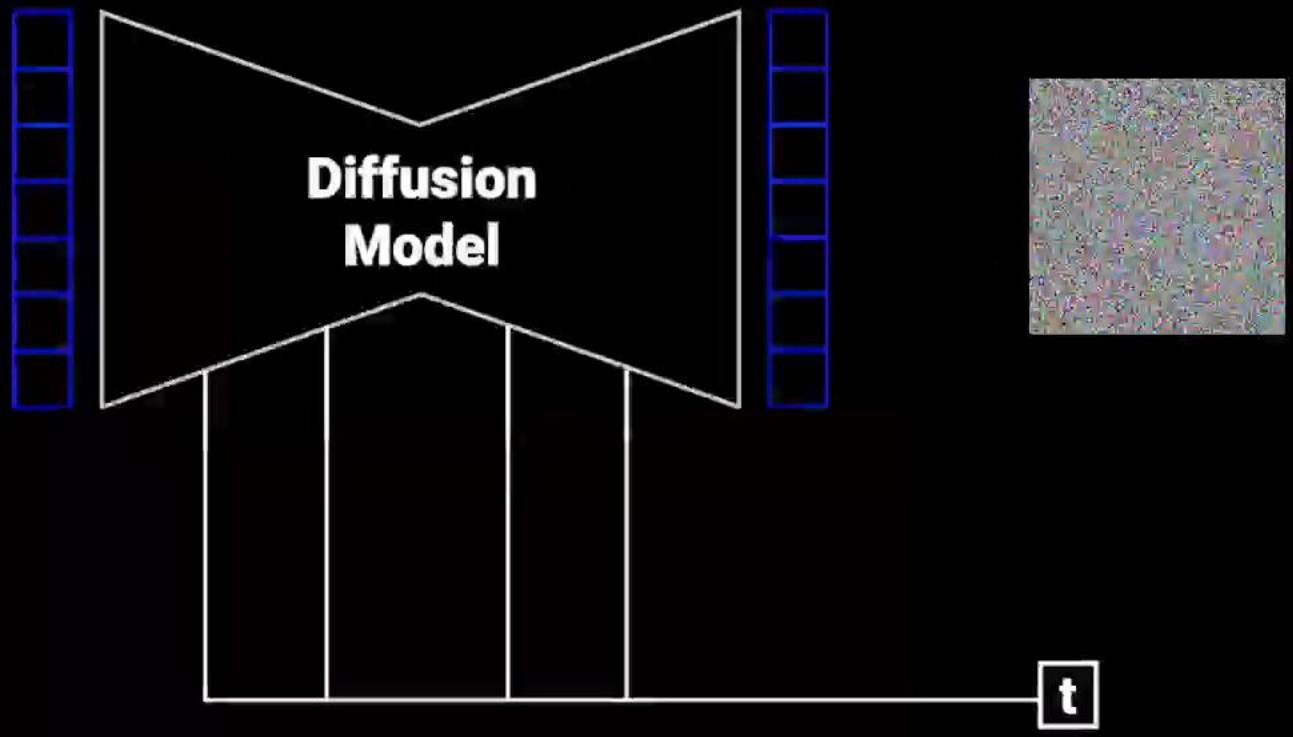

Level 1

Diffusion Model

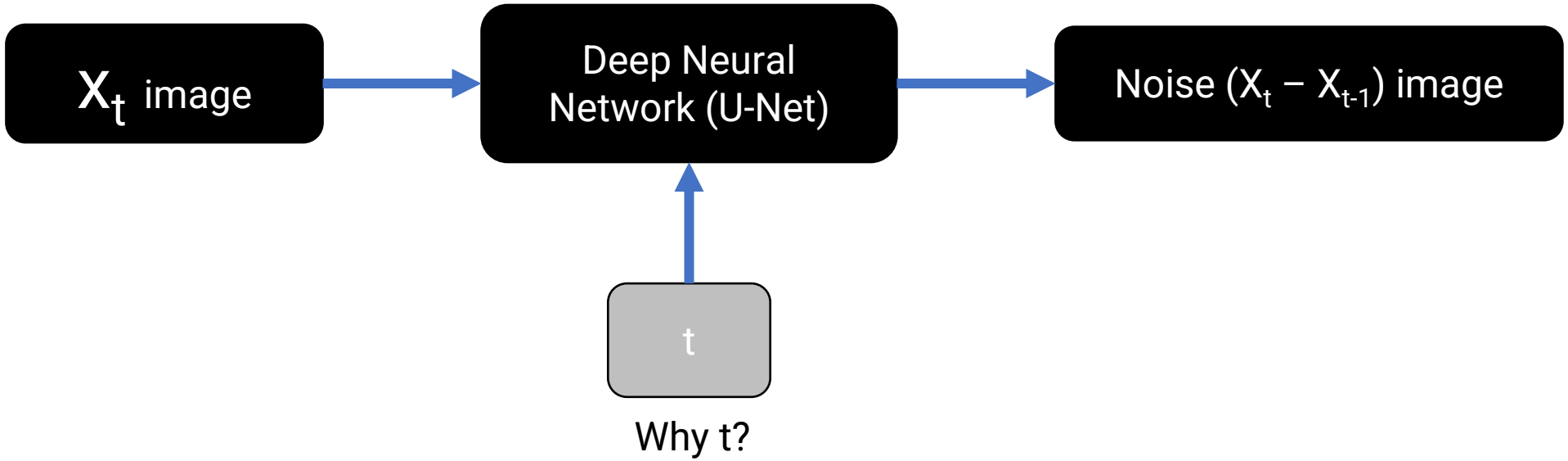
What We Want To Do!

How Do We Predict?

We do it recursively!



Diffusion model architecture



Lets’ get the data first

Forward Process

Pieter Abbeel

Forward Process

$q(x_t|x_{t-1})$

This is a Markov Chain just like we've seen before!

Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics." International Conference on Machine Learning. PMLR, 2015.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.

(9141) Diffusion Models | Paper Explanation | Math Explained - YouTube

A Drop Of Math

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

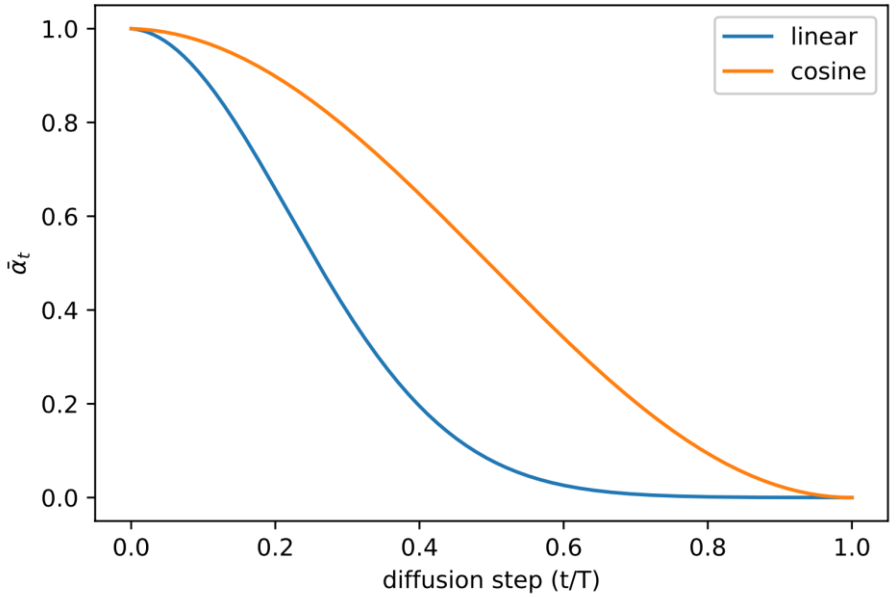
Forward Process

Normal Distribution

Output

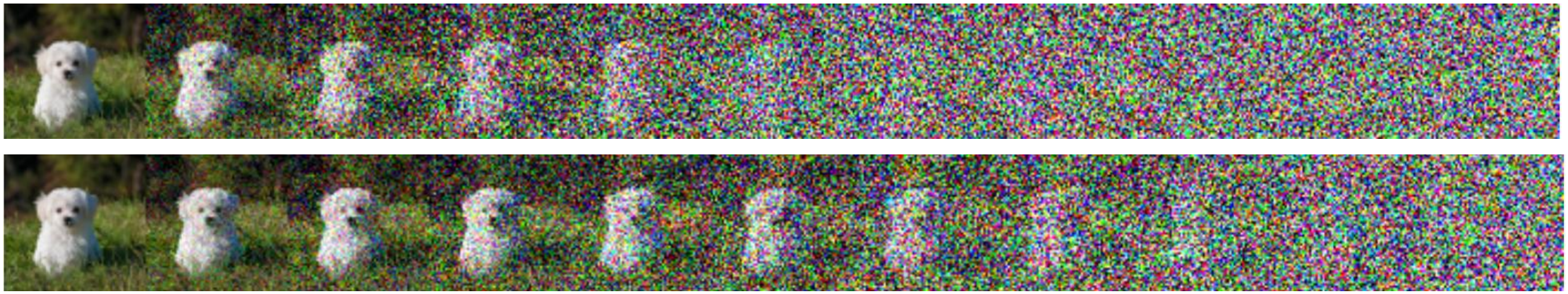
Mean

Variance



We are recursively adding Gaussian Noise

β_t follows a **Linear schedule**



β_t follows a better **Cosine schedule**

Nichol, Alexander Quinn, and Prafulla Dhariwal. "Improved denoising diffusion probabilistic models." International Conference on Machine Learning. PMLR, 2021.

Mean and Variance

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

using the notation $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$

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

beta<<1
This form ensures,

- At t = 0, the distribution is almost equal to original image
- At t = T, the distribution is almost the gaussian noise

Training

A Drop Of Math

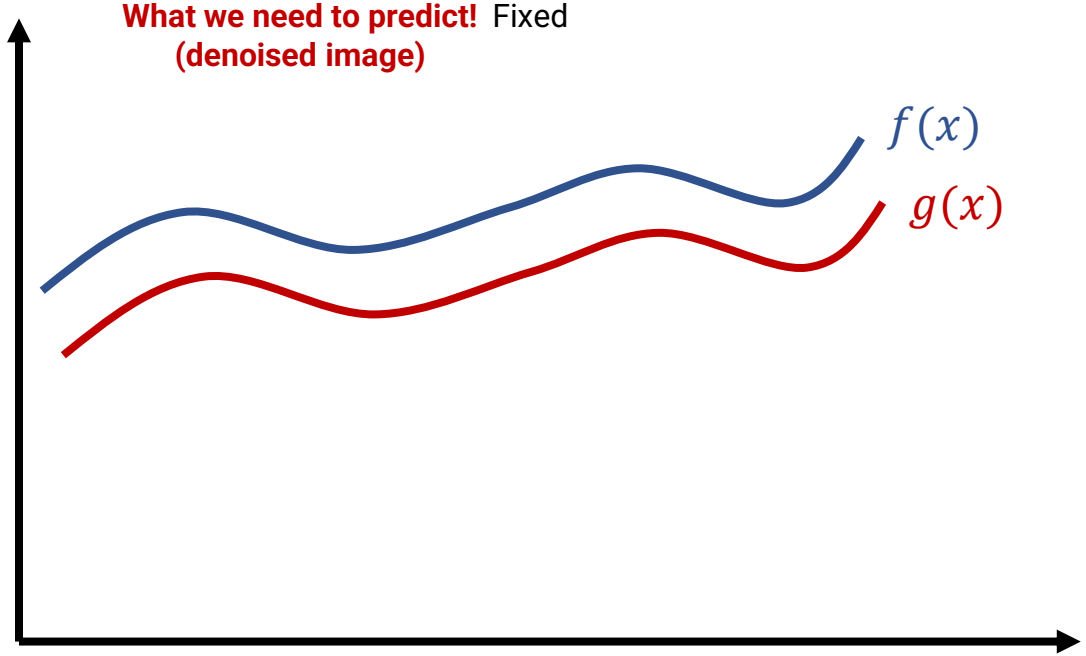
From The Ocean

A Drop Of Math

We want to predict a slightly denoised image given a noisy image!

$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$ **Reverse Process**

What we need to predict! (denoised image) Fixed



Loss function becomes:

$$\mathcal{L}_t = \frac{1}{2\sigma_t^2} \|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2$$

MSE between predicted and actual means (images)

Further simplifies to:

$$\mathcal{L}_t = \frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1 - \hat{\alpha}_t)} \|\epsilon - \epsilon_\theta(x_t, t)\|^2$$

Finally, in practice

$$\boldsymbol{\mathcal{L}_t = \|\epsilon - \epsilon_\theta(x_t, t)\|^2}$$

Loss between predicted and actual noise

I skipped over pages of math to get here!

Main trick: Optimize **Variational Lower Bound** rather than optimization function

Concept adapted from Variational Auto Encoders (VAEs)

Training Algorithm

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
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

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

Backward Process

Backward Process
$p(x_{t-1}|x_t)$

DDPM Algorithm

Algorithm 2 Sampling

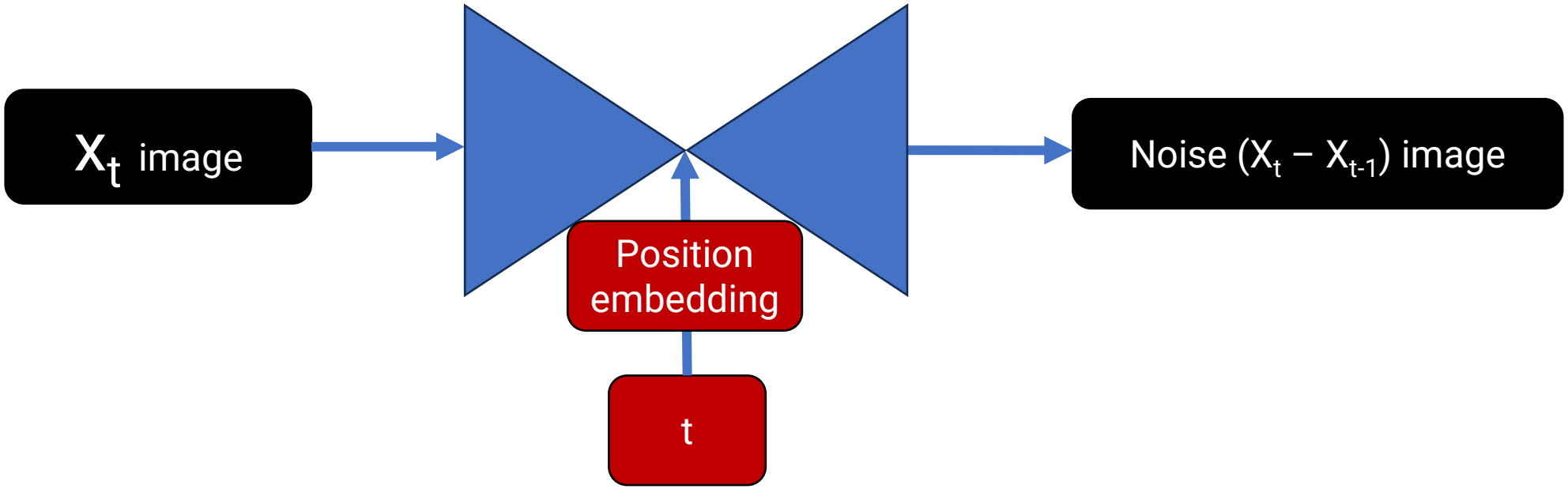
1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}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

What are Diffusion Models? | Lil'Log (lilianweng.github.io)

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.

Diffusion Network Embedding

Time embedding



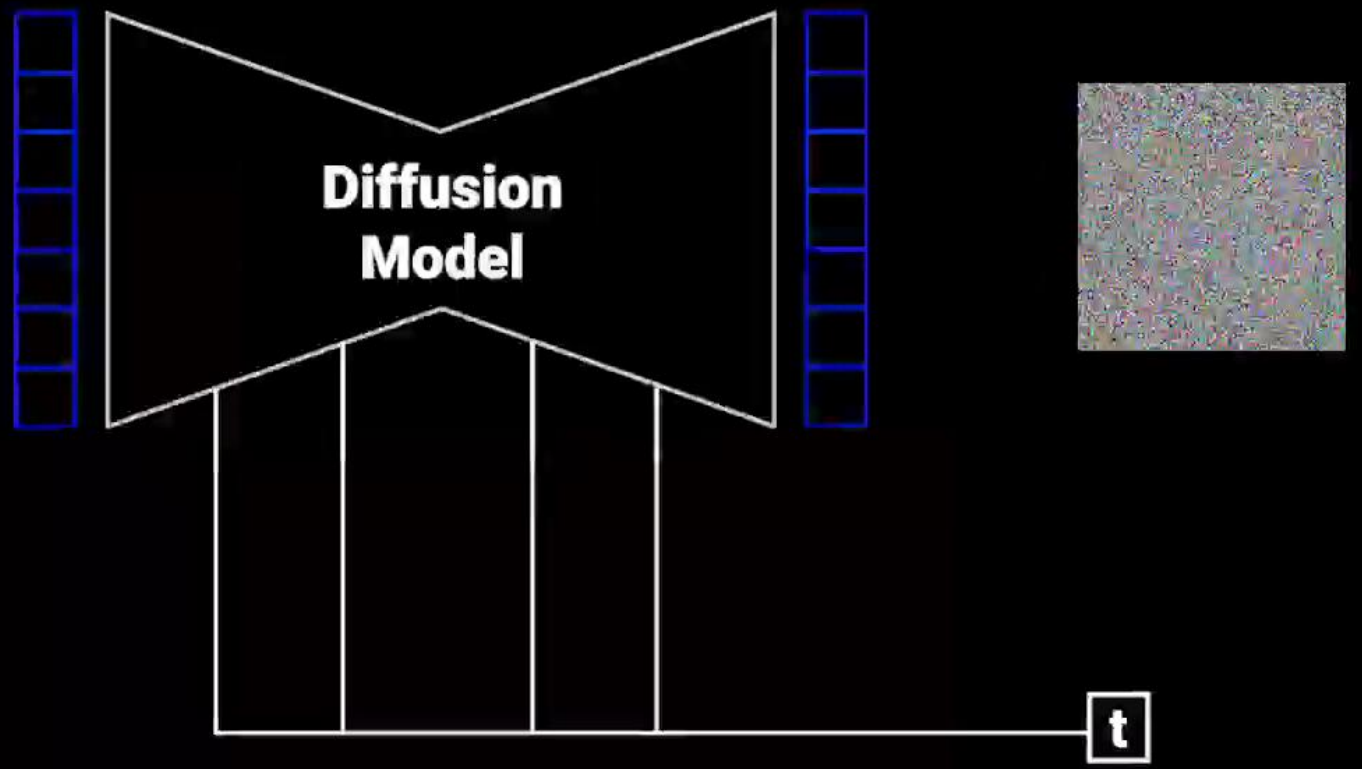
The network does not know how much noise is present in the image without timestep.

Position encoding

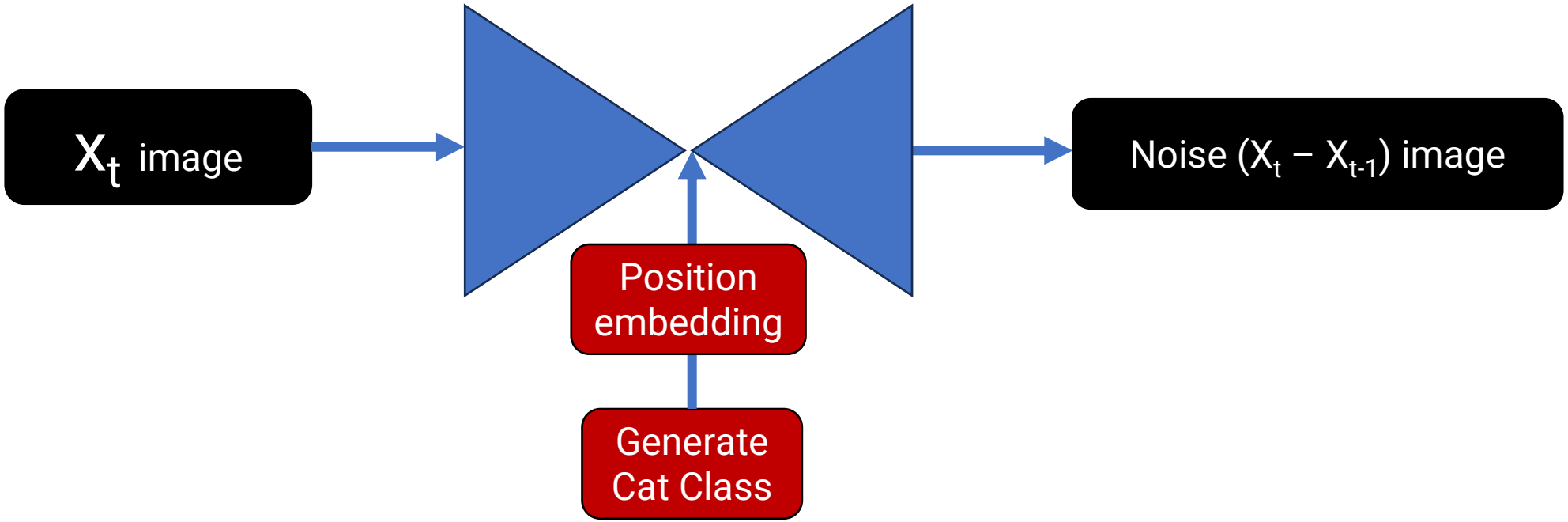
- Decompose the input into sine and cosine components of varying frequencies and add them to the latent space.
- This method is also used in NeRF and transformers.

How Do We Predict?

We do it recursively!



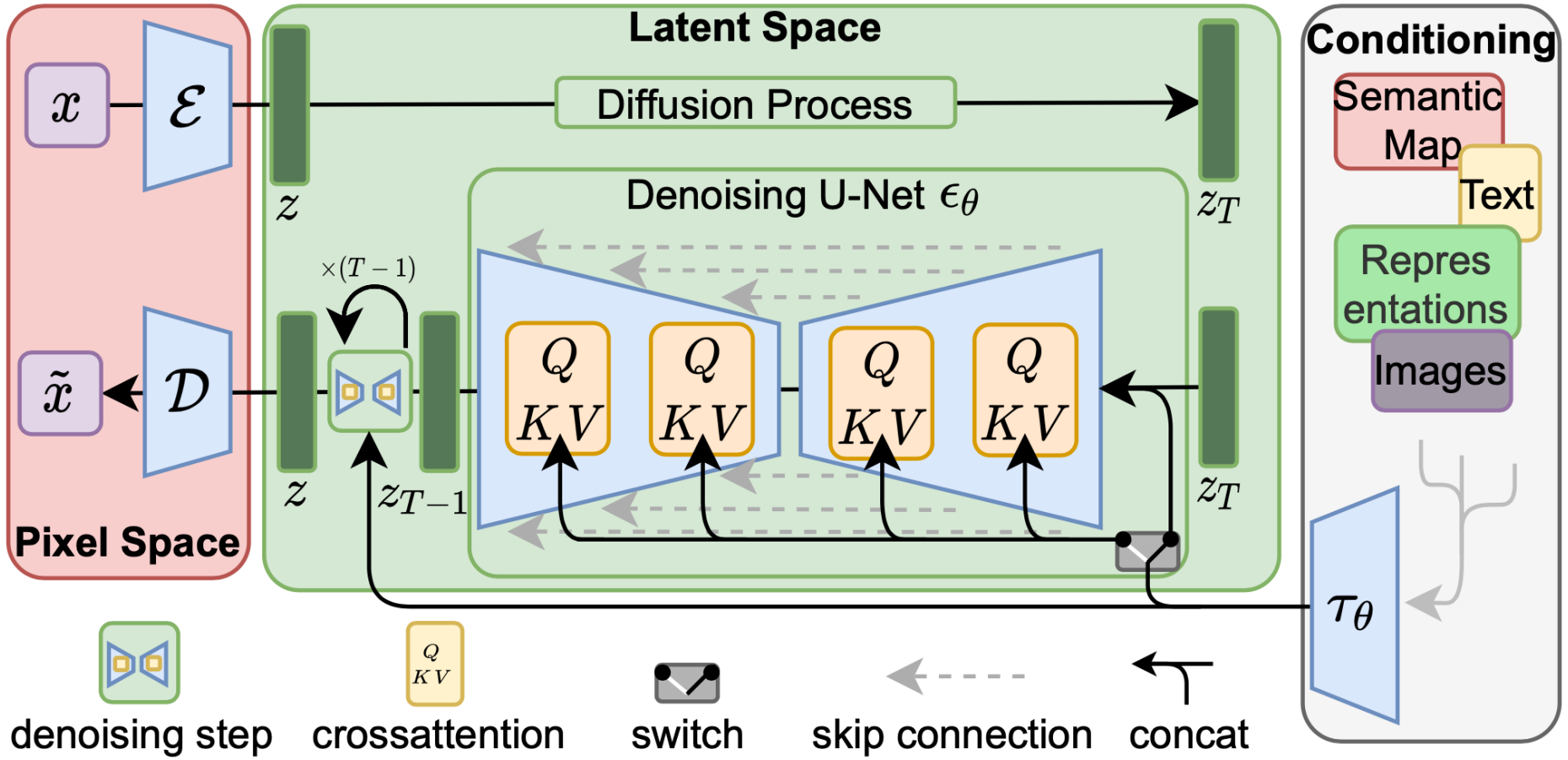
Condition embedding



What else can we embed?

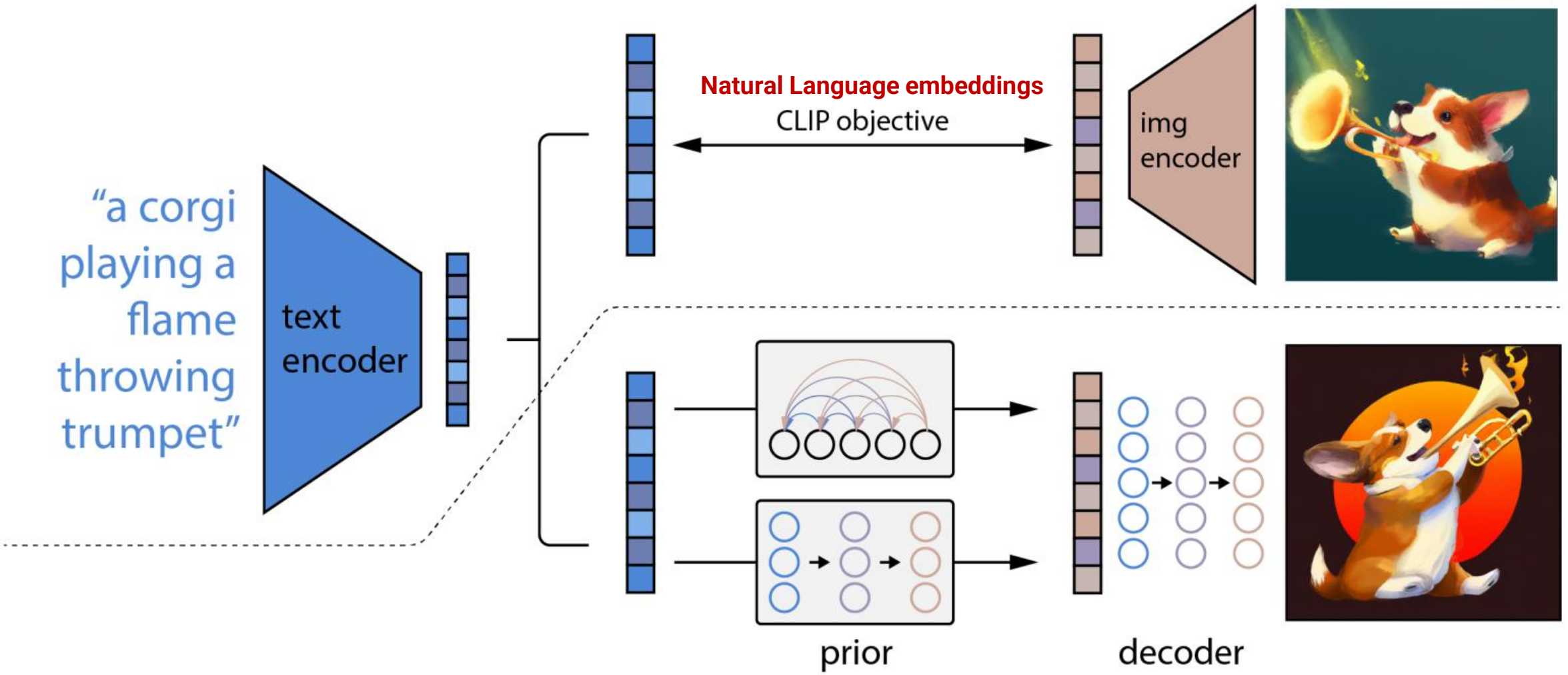
What else can we use?

- Semantics
- Text
- Other Images



Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Dall·E2



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

Works Well!

vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

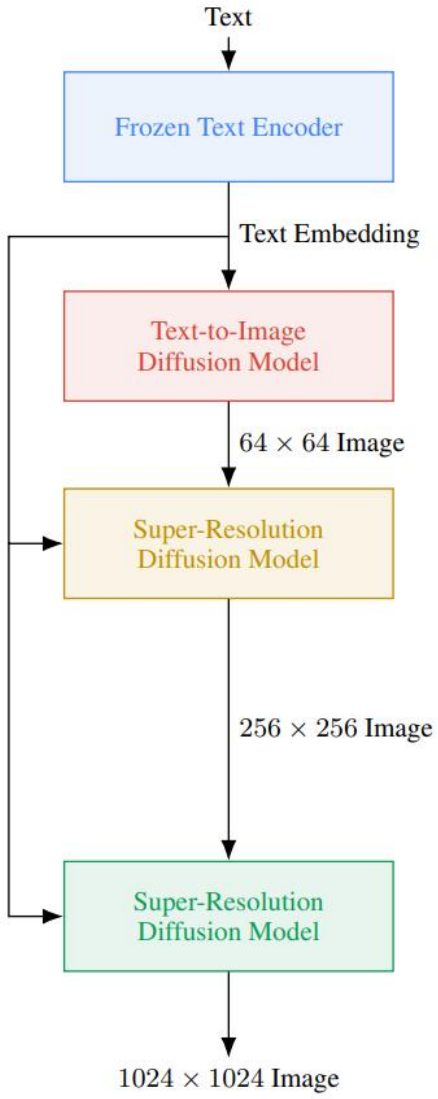
a close up of a handpalm with leaves growing from it

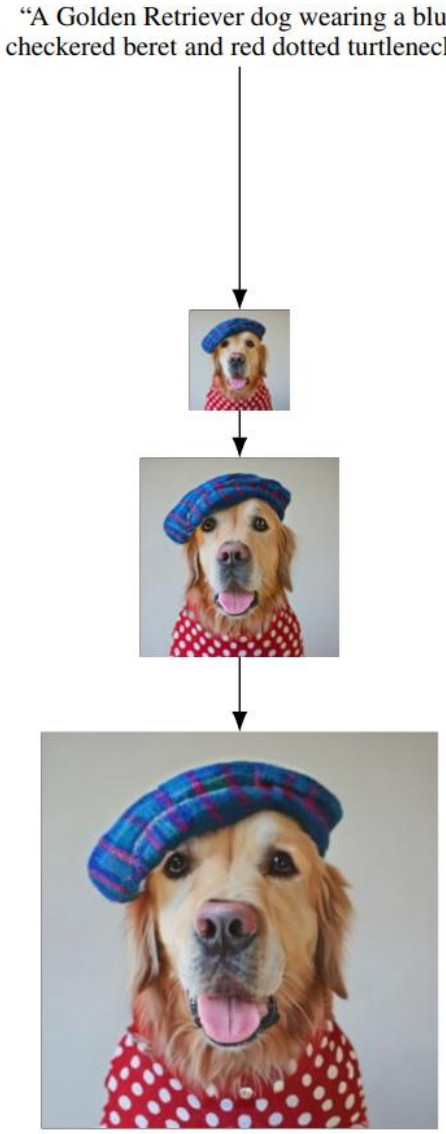
an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

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

Imagen





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

Works Well Too!

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

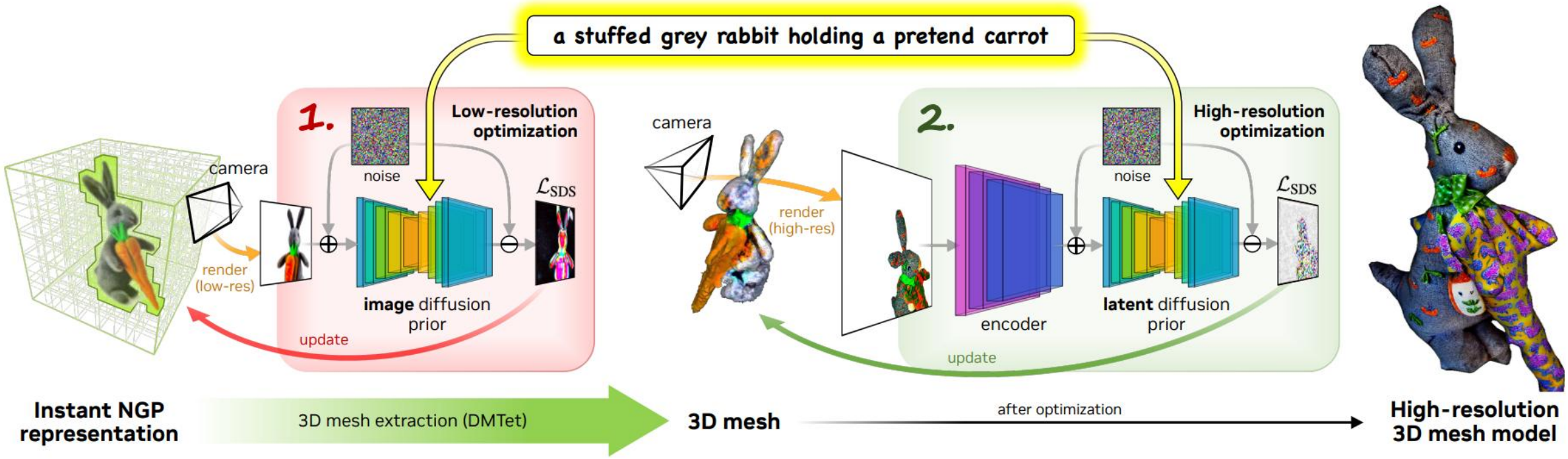
A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Teddy bears swimming at the Olympics 400m Butterfly event.

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

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

Can Do 3D Too!



Lin, Chen-Hsuan, et al. "Magic3D: High-Resolution Text-to-3D Content Creation." arXiv preprint arXiv:2211.10440 (2022).

3D Assets For Everyone!

a silver platter piled high with fruits

michelangelo style statue of an astronaut

a stuffed grey rabbit holding a pretend carrot

an iguana holding a balloon

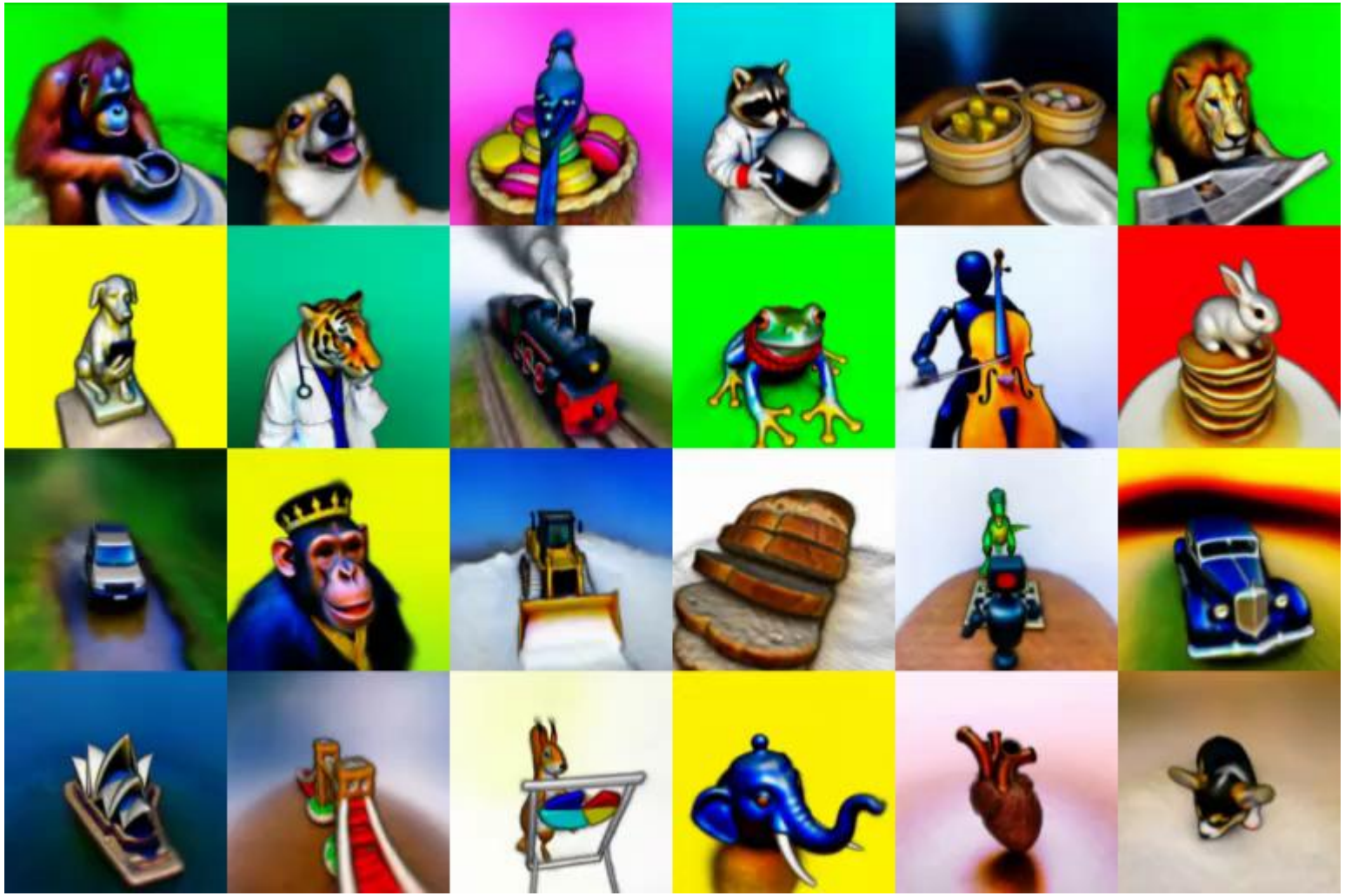
a beautiful dress made out of garbage bags

an imperial state crown of england

a blue poison-dart frog sitting on a water lily

neuschwanstein castle, aerial view

DreamFusion

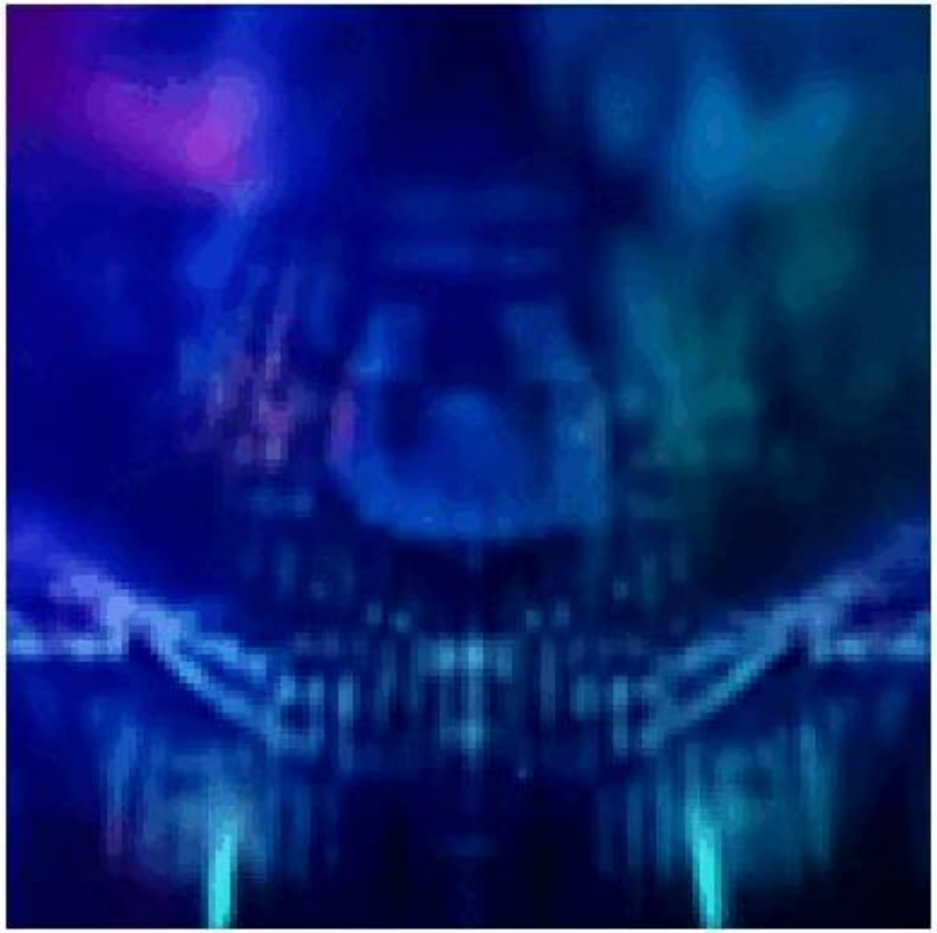


Poole, Ben, et al. "Dreamfusion: Text-to-3d using 2d diffusion." arXiv preprint arXiv:2209.14988 (2022).

Anything Weird About This Video?

Can Do Videos Too!

Abstract background

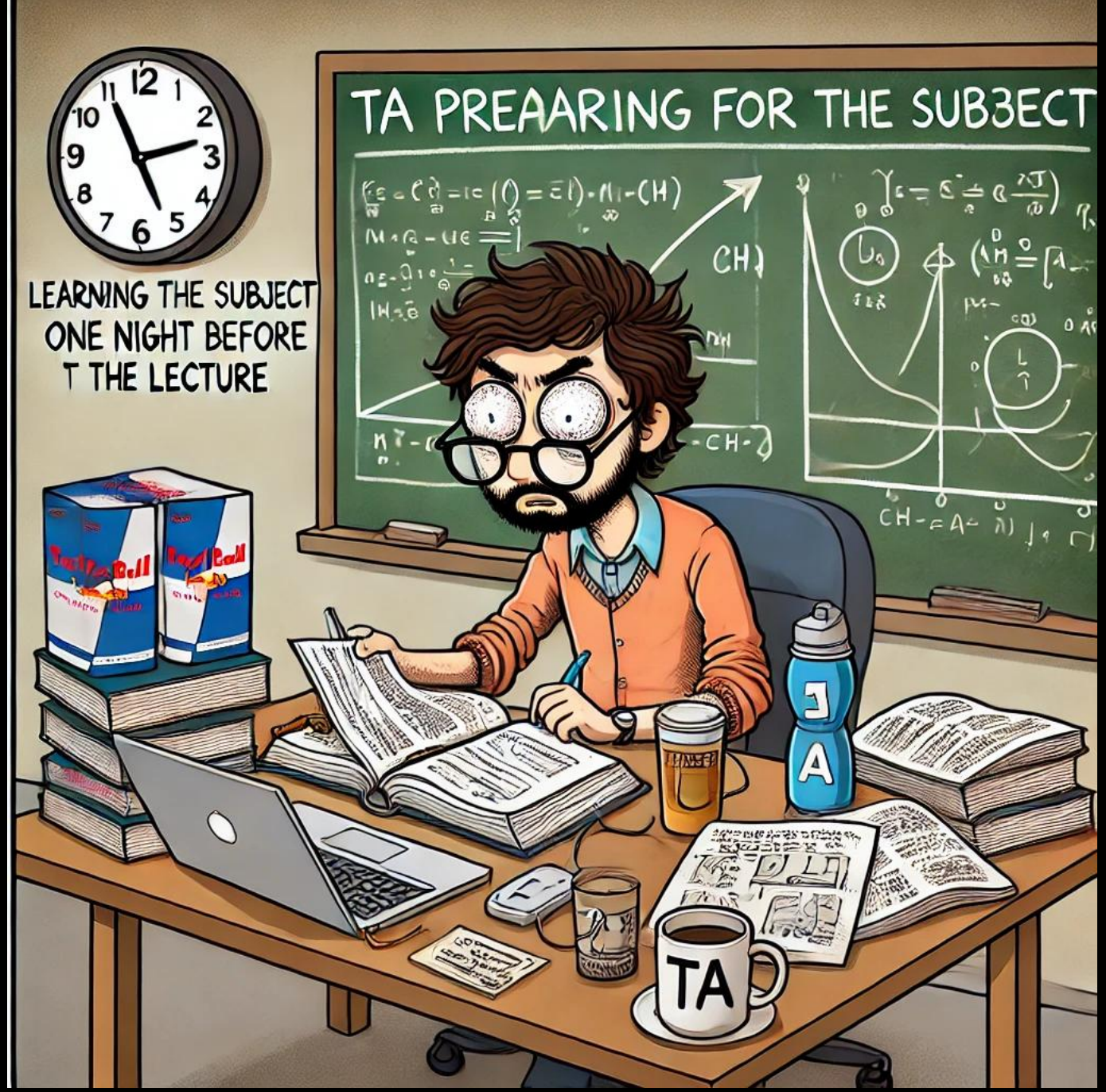


Ho, Jonathan, et al. "Video diffusion models." arXiv preprint arXiv:2204.03458 (2022).

Best Digital Art

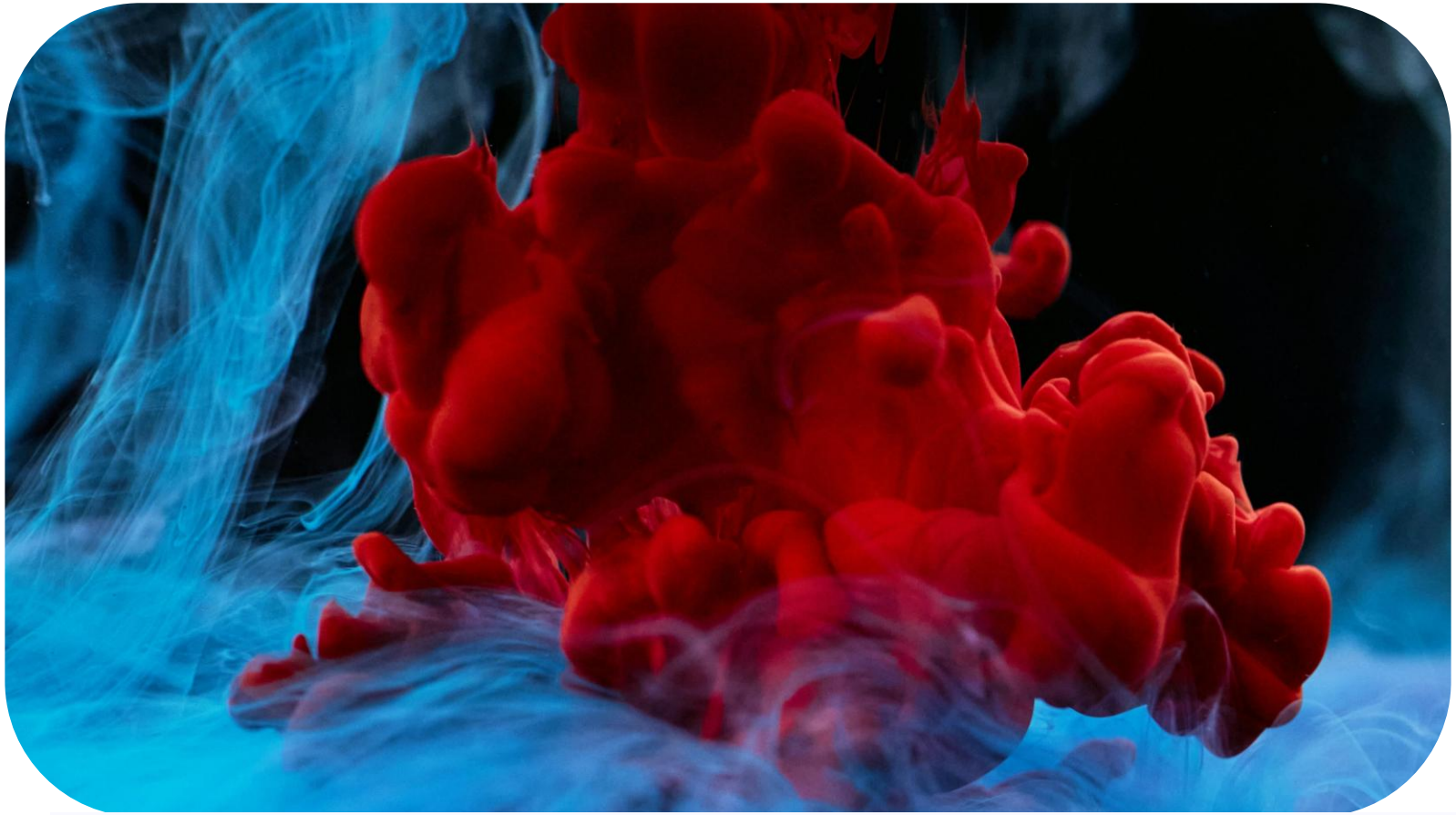
Catch was that it was AI generated!

People Have Begun to Sell Their Prompts for AI-Generated Artwork - ExtremeTech



Gen AI is not perfect

Next Class!



Advanced Generative Models++: Multi-modal Generative Deep Learning