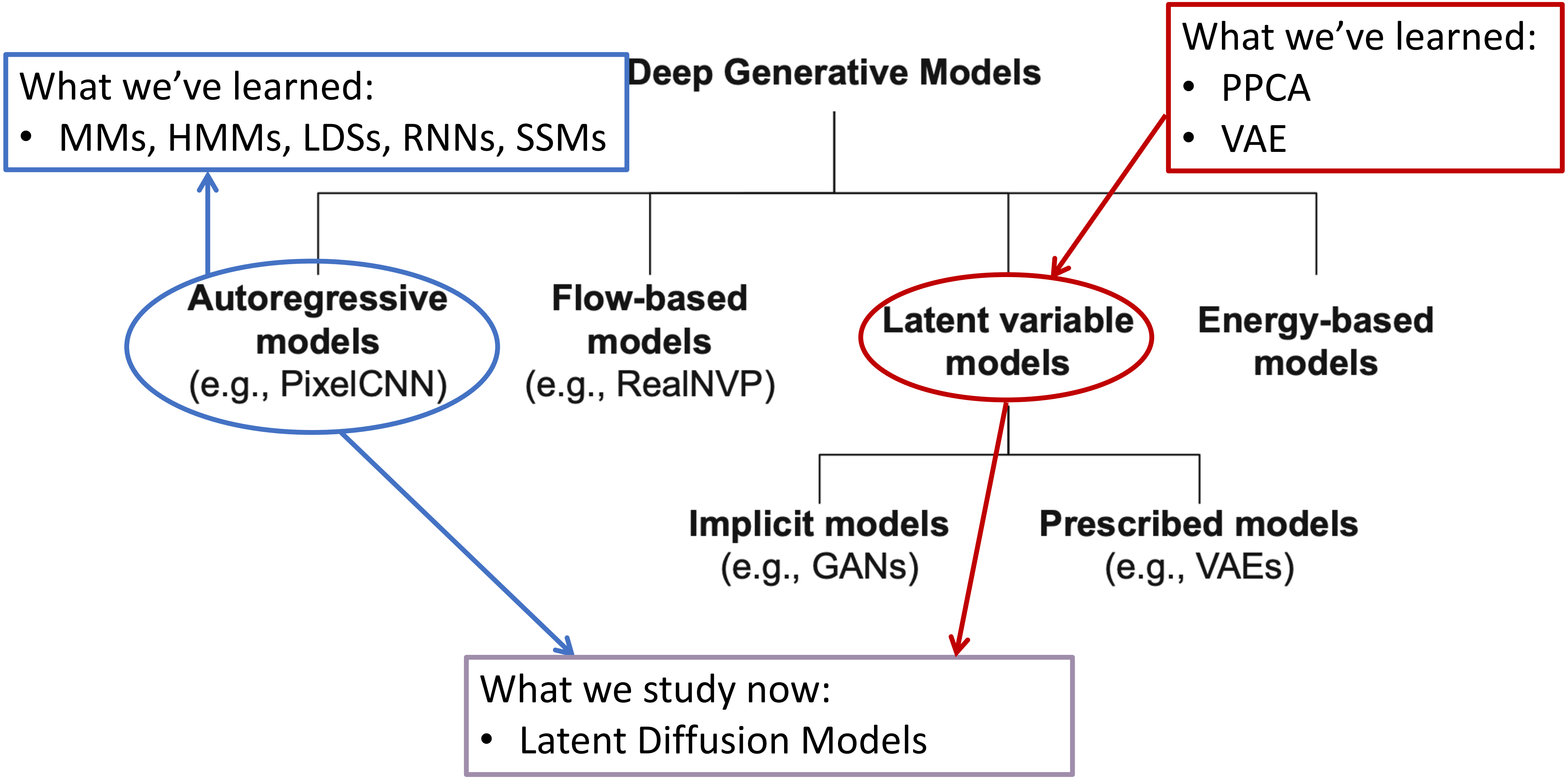
Deep Generative Models: Diffusion Models

Fall Semester 2025

René Vidal
Director of the Center for Innovation in Data Engineering and Science (IDEAS)
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE

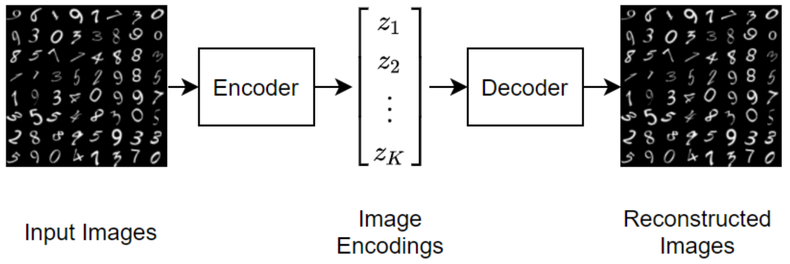
Penn
UNIVERSITY of PENNSYLVANIA

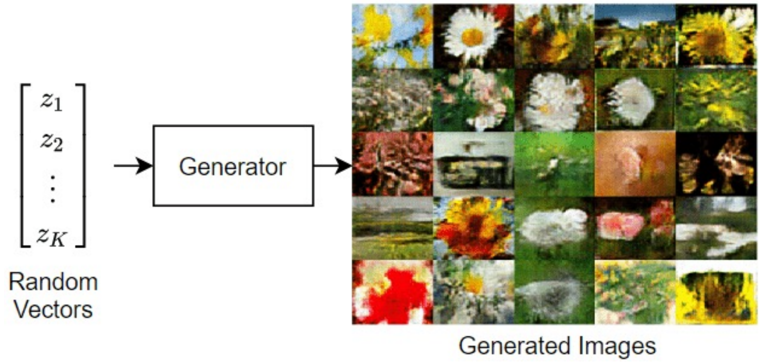
Taxonomy of Generative Models



Diffusion Models

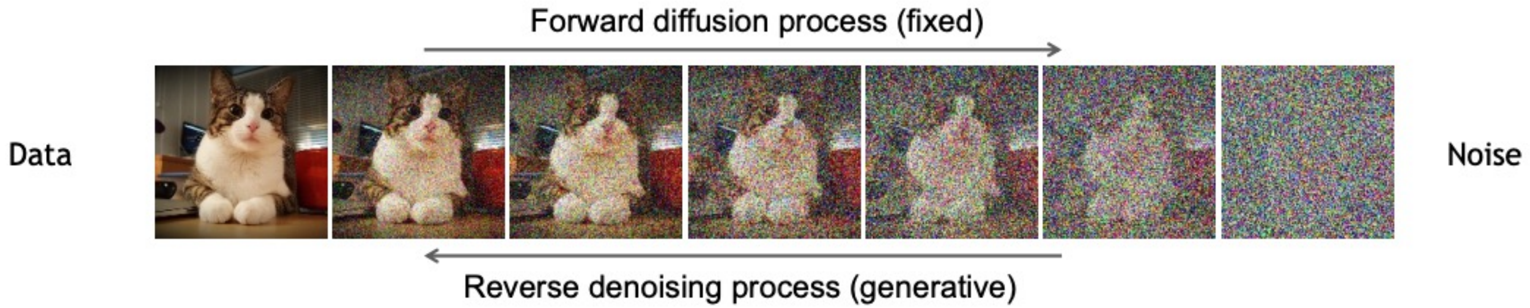
- The journey of generative models has evolved significantly in recent years.
- **Variational Autoencoders (VAEs)** introduce probabilistic modeling for latent representations but struggled with generating high-quality images.
- This led to the rise of **Generative Adversarial Networks (GANs)**, which leverage adversarial learning to produce high-quality, realistic outputs but suffered from issues like mode collapse and unstable training.
- The introduction of **Diffusion Models** achieve state-of-the-art results with superior stability and diversity in generated samples, particularly in multimodal image synthesis.

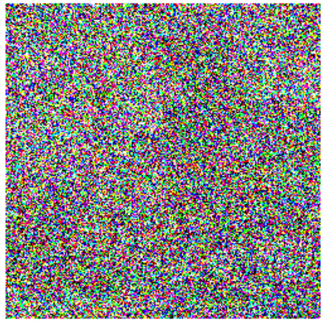


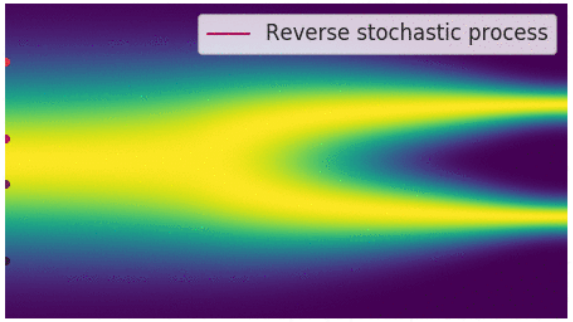


Diffusion Models

- A Latent Diffusion Model is a VAE with an autoregressive latent space
- The VAE encoder maps data to noise by gradually adding Gaussian noise to the input using a diffusion process.
- The VAE decoder maps noise to data by learning how to reverse the forward diffusion process. The reverse process predicts how to denoise.







Outline

- Markov Hierarchical Variational Auto Encoders (MHVAE)
  - Encoder and Decoder of a MHVAE
  - **Derivation of the ELBO of a MHVAE**
- Diffusion Models are MHVAEs with Linear Gaussian Autoregressive latent space
  - Forward Process
  - Conditional Distributions for the Forward Process
  - Reverse Process
  - ELBO for Diffusion Models is a particular case of ELBO for VAEs with extra structure
  - Implementation Details
- Application of Diffusion Models
  - Stable Diffusion: Text-Conditioned Diffusion Model
  - ControlNet: Multimodal Control for Consistent Synthesis

Recall the Variational Autoencoder (VAE)

Datapoint x

Encoder
$q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$

μ

$\log \sigma^2$

Sampler
$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z} \mid \mu_\phi(\boldsymbol{x}), \sigma_\phi^2(\boldsymbol{x}) I)$

Decoder
$p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$

$\mu_\theta(\boldsymbol{z})$

ELBO Objective

$$\mathbb{E}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) - KL\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p(\boldsymbol{z})\right)]$$

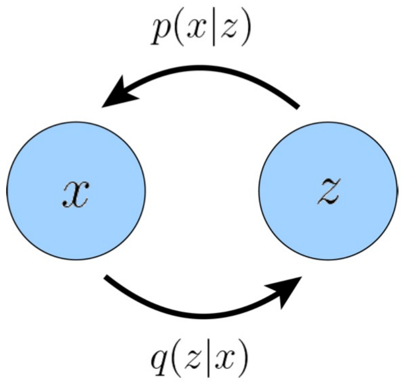
Recall the Evidence Lower Bound (ELBO)

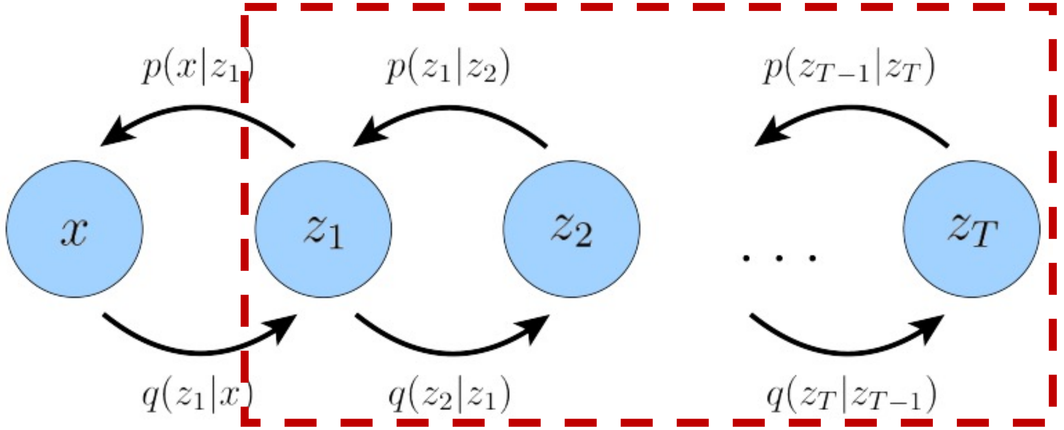
- The ELBO is the sum of a reconstruction term and a prior matching term

$$
\begin{aligned}
\log p_\theta(x) &\geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] + \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(z)}{q_\phi(z|x)}\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)]}_{\text{reconstruction term}} - \underbrace{\mathrm{D_{KL}}\left(q_\phi(z \mid x) \,||\, p(z)\right)}_{\text{prior matching term}}
\end{aligned}
$$

Latent Diffusion Models as “Autoregressive VAEs”

- A Latent Diffusion Model is as a Markovian Hierarchical Variational Autoencoder (MHVAE) with T hierarchical latents $z = z_{1:T} = \{z_t\}_{t=1}^{T}$ modeled by a Markov chain where each latent z_t is generated only from the previous latent z_{t+1}.



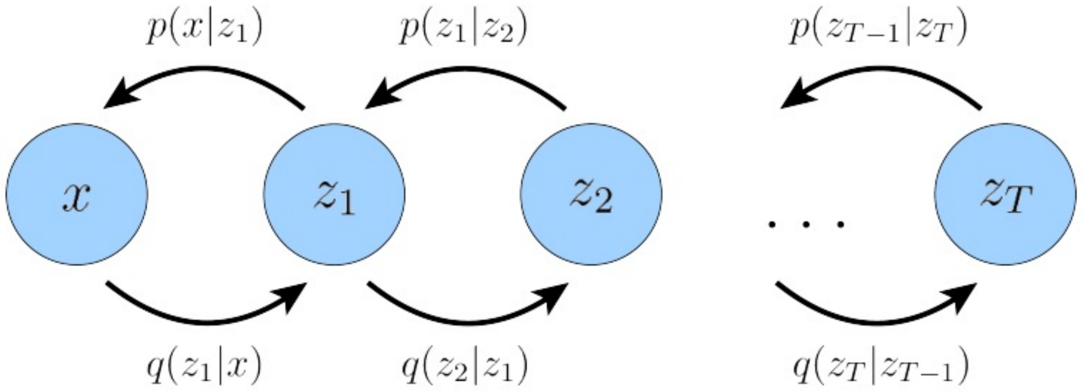


$z = z_{1:T} = \{z_t\}_{t=1}^{T}$

- What is the VAE encoder $q(z \mid x)$ of a Diffusion Model ?
- What is the VAE decoder $p(x \mid z)$ of a Diffusion Model ?
- What is the ELBO of a Diffusion Model ?

MHVAE Encoder, Decoder, and ELBO

- A MHVAE is a VAE whose encoder and decoder are autoregressive models:



$$p_\theta(x, z_{1:T}) = p_\theta(z_T) p_\theta(x \mid z_1) \prod_{t=2}^{T} p_\theta(z_{t-1} \mid z_t)$$

$$q_\phi(z_{1:T} \mid x) = q_\phi(z_1 \mid x) \prod_{t=2}^{T} q_\phi(z_t \mid z_{t-1})$$

- Given this joint distribution and posterior, we can rewrite the ELBO for MHVAE as:

$$\mathbb{E}_{q_\phi(z_{1:T}|x)}\left[\log \frac{p_\theta(x, z_{1:T})}{q_\phi(z_{1:T} \mid x)}\right] = \mathbb{E}_{q_\phi(z_{1:T}|x)}\left[\log \frac{p_\theta(z_T) p_\theta(x \mid z_1) \prod_{t=2}^{T} p_\theta(z_{t-1} \mid z_t)}{q_\phi(z_1 \mid x) \prod_{t=2}^{T} q_\phi(z_t \mid z_{t-1})}\right]$$

Decomposition of the ELBO for a MHVAE

- Let us make the change of variables $x \to x_0$ and $z_{1:T} \to x_{1:T}$.

- **Theorem**: The ELBO for a MHVAE can be written as

$$\mathbb{E}_{q_\phi(x_{1:T}|x)}\left[\log \frac{p_\theta(x_T)p_\theta(x_0 \mid x_1)\prod_{t=2}^{T} p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_1 \mid x_0)\prod_{t=2}^{T} q_\phi(x_t \mid x_{t-1})}\right] =$$

$$\underbrace{\mathbb{E}_{q_\phi(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(x_T \mid x_0) \,||\, p_\theta(x_T)\right)}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q_\phi(x_t|x_0)}\left[D_{\mathrm{KL}}\left(q_\phi(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\right)\right]}_{\text{score matching term}}$$

- Proof (1/3): Reversing q

$$q_\phi(x_t \mid x_{t-1}) = q_\phi(x_t \mid x_{t-1}, x_0) = \frac{q_\phi(x_{t-1}|x_t, x_0)q_\phi(x_t|x_0)}{q_\phi(x_{t-1}|x_0)}.$$

Decomposition of the ELBO for a MHVAE

- Proof (2/3): substituting q and using telescopic product to cancel factors

$$\log p(x) \geq \mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_T)p_\theta(x_0 \mid x_1)\prod_{t=2}^{T} p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_1 \mid x_0)\prod_{t=2}^{T} q_\phi(x_t \mid x_{t-1})}\right]$$

$$= \mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_T)p_\theta(x_0 \mid x_1)}{q_\phi(x_1 \mid x_0)} \prod_{t=2}^{T} \frac{p_\theta(x_{t-1} \mid x_t)}{\frac{q_\phi(x_{t-1} \mid x_t, x_0)q_\phi(x_t \mid x_0)}{q_\phi(x_{t-1} \mid x_0)}}\right]$$

$$= \mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_T)p_\theta(x_0 \mid x_1)}{q_\phi(x_1 \mid x_0)} \prod_{t=2}^{T} \frac{p_\theta(x_{t-1} \mid x_t)q_\phi(x_1 \mid x_0)}{q_\phi(x_{t-1} \mid x_t, x_0)q_\phi(x_T \mid x_0)}\right]$$

$$= \mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_T)p_\theta(x_0 \mid x_1)q_\phi(x_1 \mid x_0)}{q_\phi(x_1 \mid x_0)q_\phi(x_T \mid x_0)} \prod_{t=2}^{T} \frac{p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_{t-1} \mid x_t, x_0)}\right]$$

Decomposition of the ELBO for a MHVAE

- Proof (3/3): expanding into three terms and simplifying expectations

$$\log p(x) \geq \mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_T)p_\theta(x_0 \mid x_1)}{q_\phi(x_T \mid x_0)} \prod_{t=2}^{T} \frac{p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_{t-1} \mid x_t, x_0)}\right]$$

$$= \mathbb{E}_{q_\phi(x_{1:T}|x_0)}[\log p_\theta(x_0 \mid x_1)] + \mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_T)}{q_\phi(x_T \mid x_0)}\right] + \sum_{t=2}^{T} \mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_{t-1} \mid x_t, x_0)}\right]$$

$$= \mathbb{E}_{q_\phi(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)] + \mathbb{E}_{q_\phi(x_T|x_0)}\left[\log \frac{p_\theta(x_T)}{q_\phi(x_T \mid x_0)}\right] + \sum_{t=2}^{T} \mathbb{E}_{q_\phi(x_t|x_0)}\left[\log \frac{p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_{t-1} \mid x_t, x_0)}\right]$$

$$= \underbrace{\mathbb{E}_{q_\phi(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(x_T \mid x_0) | p_\theta(x_T)\right)}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q_\phi(x_t|x_0)}\left[D_{\mathrm{KL}}\left(q_\phi(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\right)\right]}_{\text{score matching term}}$$

Why can we simplify expectations?

$$\int p(x_{2:T} \mid x_1) dx_{2:T} = 1$$

- For the first term:

$$\mathbb{E}_{q_\phi(x_{1:T}|x_0)}[\log(p_\theta(x_0 \mid x_1))] = \int \log(p_\theta(x_0 \mid x_1))\, q_\phi(x_{1:T} \mid x_0) dx_{1:T}$$

$$= \int \log(p_\theta(x_0 \mid x_1))\, q_\phi(x_1 \mid x_0) q_\phi(x_{2:T} \mid x_1) dx_1 dx_{2:T}.$$

$$= \int \log p_\theta(x_0 \mid x_1) q_\phi(x_1 \mid x_0) dx_1 = \mathbb{E}_{q_\phi(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)]$$

- Similarly for the second term:

$$\mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_T)}{q_\phi(x_T \mid x_0)}\right] = \int \log\left(\frac{p_\theta(x_T)}{q_\phi(x_T \mid x_0)}\right) q_\phi(x_{1:T} \mid x_0) dx_{1:T} = \int \log\left(\frac{p_\theta(x_T)}{q_\phi(x_T \mid x_0)}\right) q_\phi(x_T \mid x_0) q_\phi(x_{1:T-1} \mid x_T, x_0) dx_{1:T}$$

ELBO for MHVAE

$$\int p(x_{2:T})dx_{2:T} = 1$$

- Why can we simplify expectations? (1/2)

$$\mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log\left(\frac{p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_{t-1} \mid x_t, x_0)}\right)\right] = \int \log\left(\frac{p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_{t-1} \mid x_t, x_0)}\right) q_\phi(x_{1:T} \mid x_0) dx_{1:T}$$

$$= \int \log\left(\frac{p_\theta(x_{t-1}|x_t)}{q_\phi(x_{t-1}|x_t, x_0)}\right) \prod_{\tau=1}^{T} q_\phi(x_\tau \mid x_{\tau-1})\, dx_{1:T} \qquad \text{(Markov property)}$$

$$= \int \boxed{\log\left(\frac{p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_{t-1} \mid x_t, x_0)}\right)} q_\phi(x_t \mid x_{t-1}) \prod_{\tau \neq t}^{T} q_\phi(x_\tau \mid x_{\tau-1})\, dx_{1:T}$$

The log term depends only on x_t, x_{t-1}, x_0 and thus marginalizing q_ϕ gives

$$\boxed{\int \prod_{\tau=1}^{t-2} q_\phi(x_\tau \mid x_{\tau-1})\, dx_{1:(t-2)} = q_\phi(x_{t-1} \mid x_0)} \quad \boxed{\int \prod_{\tau=t+1}^{T} q_\phi(x_\tau \mid x_{\tau-1})\, dx_{(t+1):T} = 1}$$

ELBO for MHVAE

$$\boxed{\int p(x_{2:T})dx_{2:T} = 1}$$

- Why can we simplify expectations? (2/2)

$$= \int \log\left(\frac{p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_{t-1} \mid x_t, x_0)}\right) q_\phi(x_t \mid x_{t-1}) q_\phi(x_{t-1} \mid x_0) dx_t dx_{t-1}$$

Using the fact that $q_\phi(x_t \mid x_{t-1}) q_\phi(x_{t-1} \mid x_0) = q_\phi(x_t \mid x_0) q_\phi(x_{t-1} \mid x_t, x_0)$, it holds

$$= \int \log\left(\frac{p_\theta(x_{t-1} \mid x_t)}{q_\phi(x_{t-1} \mid x_t, x_0)}\right) q_\phi(x_t \mid x_0) q_\phi(x_{t-1} \mid x_t, x_0) dx_{t-1} dx_t$$

$$= \int \log\left(\frac{p_\theta(x_{t-1} | x_t)}{q_\phi(x_{t-1} | x_t, x_0)}\right) q_\phi(x_t \mid x_0) \mathbb{E}_{x_{t-1} \sim q_\phi(\cdot | x_t, x_0)}\left[\log \frac{p_\theta(x_{t-1} | x_t)}{q_\phi(x_{t-1} | x_t, x_0)}\right] dx_t$$

$$= \mathbb{E}_{q_\phi(x_t | x_0)}\left[\mathbb{E}_{q_\phi(x_{t-1} | x_t, x_0)}\left[\log\left(\frac{p_\theta(x_{t-1} | x_t)}{q_\phi(x_{t-1} | x_t, x_0)}\right)\right]\right]$$

$$= \boxed{\mathbb{E}_{q_\phi(x_t | x_0)}\left[D_{\mathrm{KL}}\left(q_\phi(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\right)\right]}$$

Interpretation of the ELBO for MHVAE

$$\log p\,(x)$$
$$\geq \underbrace{\mathbb{E}_{q_\phi(x_1|x_0)}[\log p_\theta\,(x_0 \mid x_1)]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(x_T \mid x_0)|p_\theta(x_T)\right)}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q_\phi(x_t|x_0)}\left[D_{\mathrm{KL}}\left(q_\phi(x_{t-1} \mid x_t, x_0)\,||\,p_\theta(x_{t-1} \mid x_t)\right)\right]}_{\text{score matching term}}$$

$\mathbb{E}_{q(x_1|x_0)}[\log p_\theta\,(x_0 \mid x_1)]$ can be interpreted as a reconstruction term; like its analogue in the ELBO of a vanilla VAE. This term can be approximated and optimized using a Monte Carlo estimate.

$D_{\mathrm{KL}}\left(q_\phi(x_T \mid x_0)|p_\theta(x_T)\right)$ represents how close the distribution of the final latent distribution is to the standard Gaussian prior.

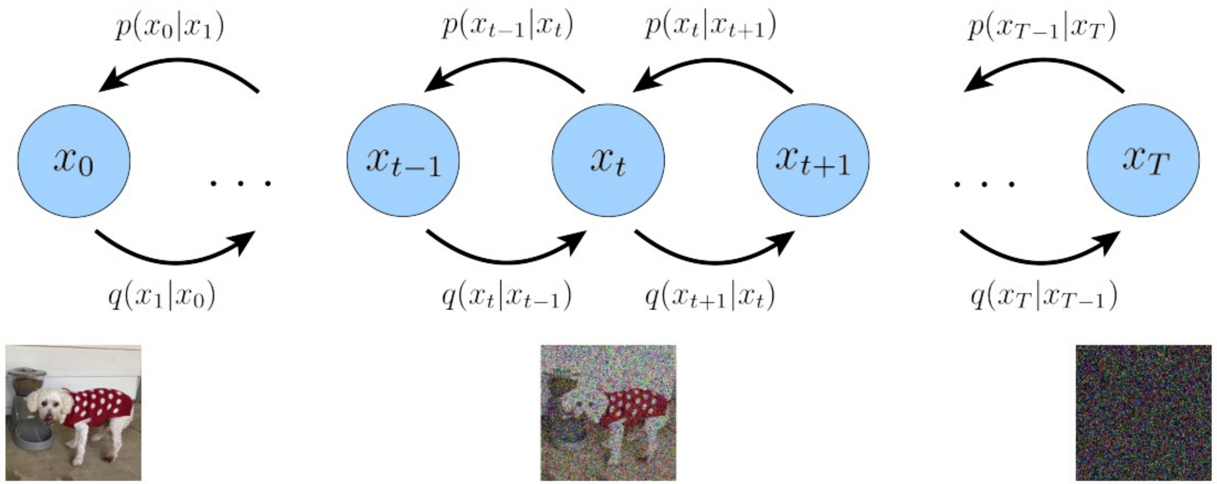
$\mathbb{E}_{q_\phi(x_t|x_0)}\left[D_{\mathrm{KL}}\left(q_\phi(x_{t-1} \mid x_t, x_0)\,||\,p_\theta(x_{t-1} \mid x_t)\right)\right]$ is a score matching term.
As we will show, the diffusion model learns the denoising transition step $p_\theta(x_{t-1} \mid x_t)$ as an approximation to the tractable, ground-truth denoising transition step $q_\phi(x_{t-1} \mid x_t, x_0)$.

Understand Diffusion Model from VAE Perspective

- **A Diffusion Model is an MHVAE**: $x_0 = x$ is the data and $x_{1:T} = z_{1:T}$ is the latent variable
- All latent variables have the same dimension as the dimension of the data
- The structure of the encoder $q_\phi(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q_\phi(x_t \mid x_{t-1})$ is not learned, but it is pre-specified as a linear Gaussian model

$$q_\phi(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t) I)$$

- The parameter α_t is chosen such that $x_T \sim \mathcal{N}(x_T; 0, I)$ is a standard Gaussian



The Forward Process of Diffusion Model

Given the formulation of a single noising step:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t}\, \epsilon, \ \ \epsilon \sim \mathcal{N}(\epsilon; 0, I),$$

we can recursively derive the closed form for arbitrary noising steps:

$$\begin{aligned} \mathbb{E}[\, x_t \mid x_0 \,] &= \mathbb{E}\Big[\sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \varepsilon_t \mid x_0 \Big] \\ &= \sqrt{\alpha_t} \mathrm{E}[\, x_{t-1} \mid x_0 \,] + \sqrt{1 - \alpha_t} \mathrm{E}[\varepsilon_t] \\ &= \sqrt{\alpha_t} \mathbb{E}[\, x_{t-1} \mid x_0 \,] \end{aligned}$$

That is:

$$\mathrm{E}[\, x_t \mid x_0 \,] \ = \sqrt{\alpha_t} \sqrt{\alpha_{t-1}} \mathrm{E}[\, x_{t-2} \mid x_0 \,] = \cdots = \sqrt{\bar{\alpha}_t} x_0$$

The Forward Process of Diffusion Model

The variance is given by

$$\begin{aligned}\mathrm{Var}(\mathrm{x_t} \mid \mathrm{x_0}) &= \mathrm{Var}\left(\sqrt{\alpha_t}\mathrm{x_{t-1}} + \sqrt{1-\alpha_t}\varepsilon_t \mid \mathrm{x_0}\right) \\ &= \alpha_t \mathrm{Var}(\mathrm{x_{t-1}} \mid \mathrm{x_0}) + (1-\alpha_t)\mathrm{Var}(\varepsilon_t) \\ &= \alpha_t \mathrm{Var}(\mathrm{x_{t-1}} \mid \mathrm{x_0}) + (1-\alpha_t)\mathrm{I}\end{aligned}$$

That is:

$$\begin{aligned}\mathrm{Var}(x_t \mid x_0) &= (1-\alpha_t)I + \alpha_t(1-\alpha_{t-1})I + \alpha_t\alpha_{t-1}(1-\alpha_{t-2})I + \ \ldots \\ &= \left(1 - \prod_{s=1}^{t} \alpha_s\right)I \\ &= (1-\overline{\alpha_t})I\end{aligned}$$

The Forward Process of Diffusion Model

To summarize: $x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1 - \overline{\alpha_t}}\,\epsilon, \epsilon \sim \mathcal{N}(\epsilon; 0, I)$

That is: $x_0 = \frac{x_t - \sqrt{1-\overline{\alpha_t}}\epsilon_0}{\sqrt{\overline{\alpha_t}}}$

The forward diffusion process can be seen as a paradigm that x_t is a linear Gaussian transformation of x_0 with scheduled randomness from a standard normal distribution.

We will use this for the reparameterization trick later.

ELBO for Diffusion Model: Score Matching Term

- To compute the third term, we need $q(x_{t-1} \mid x_t, x_0) = \dfrac{q(x_t \mid x_{t-1}, x_0)q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)}$
- Letting $\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$, recall that $q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I\right)$$

- Therefore

$$\begin{aligned} \mathrm{q}(\mathrm{x}_{t-1} \mid \mathrm{x}_t, \mathrm{x}_0) &= \frac{\mathrm{q}(\mathrm{x}_t \mid \mathrm{x}_{t-1}, \mathrm{x}_0)\mathrm{q}(\mathrm{x}_{t-1} \mid \mathrm{x}_0)}{\mathrm{q}(\mathrm{x}_t \mid \mathrm{x}_0)} \\ &= \frac{\mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}}x_{t-1}, (1-\alpha_t)I\right)\mathcal{N}\left(x_{t-1}; \sqrt{\overline{\alpha_{t-1}}}x_0, (1-\overline{\alpha_{t-1}})I\right)}{\mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I\right)} \\ &\propto \mathrm{N}\Bigg(x_{t-1}; \underbrace{\frac{\sqrt{\overline{\alpha_t}}(1-\overline{\alpha_{t-1}})x_t + \overline{\overline{\alpha}}_{t-1}(1-\alpha_t)x_0}{1-\overline{\alpha_t}}}_{\mu_q(x_t, x_0)}, \underbrace{\frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}I}_{\Sigma_q(t)}\Bigg) \end{aligned}$$

ELBO for Diffusion Model: Matching the Mean

- Recall KL divergence for Gaussians

$$D_{\mathrm{KL}}\left(\mathcal{N}(\mathrm{x};\mu_{\mathrm{x}},\Sigma_{\mathrm{x}})|\mathcal{N}(\mathrm{y};\mu_{\mathrm{y}},\Sigma_{\mathrm{y}})\right)=\frac{1}{2}\left[\cancel{log\frac{|\Sigma_{\mathrm{y}}|}{|\Sigma_{\mathrm{x}}|}}-\mathrm{d}+\cancel{\mathrm{tr}}(\Sigma_{\mathrm{y}}^{-1}\Sigma_{\mathrm{x}})+(\mu_{\mathrm{y}}-\mu_{\mathrm{x}})^{\mathrm{T}}\Sigma_{\mathrm{y}}^{-1}(\mu_{\mathrm{y}}-\mu_{\mathrm{x}})\right]$$

- Choose variance of p to match exactly variance of q

$$\sigma_q^2(t)=\frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}$$

$$D_{\mathrm{KL}}(q(x_{t-1}\mid x_t,x_0)|p_\theta(x_{t-1}\mid x_t))$$

$$=D_{\mathrm{KL}}\left(\mathcal{N}\left(x_{t-1};\mu_q,\Sigma_q(t)\right)|\mathcal{N}\left(x_{t-1};\mu_\theta,\Sigma_q(t)\right)\right)$$

$$=\frac{1}{2\sigma_q^2(t)}\left[|\mu_\theta-\mu_q|_2^2\right]=\boxed{\frac{1}{2\sigma_q^2(t)}\frac{\overline{\alpha_{t-1}}(1-\alpha_t)^2}{(1-\overline{\alpha_t})^2}\left[|\widehat{x_\theta}(x_t,t)-x_0|_2^2\right]}$$

- Choose mean of p to match form of mean of q

$$\mu_\theta(x_t,t)=\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})x_t+\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)\widehat{x_\theta}(x_t,t)}{1-\overline{\alpha_t}}\qquad \mu_q(x_t,x_0)=\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})x_t+\sqrt{\overline{\alpha_{t-1}}}(1-\alpha_t)x_0}{1-\overline{\alpha_t}}$$

Reparameterization as an Alternative Form for ELBO

- Plugging our previous finding $x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_0}{\sqrt{\bar{\alpha}_t}}$ into the denoising transition mean $\mu_q(x_t, x_0)$, we have:

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}$$

$$= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\dfrac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}}{1-\bar{\alpha}_t}$$

$$= \frac{1-\bar{\alpha}_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0$$

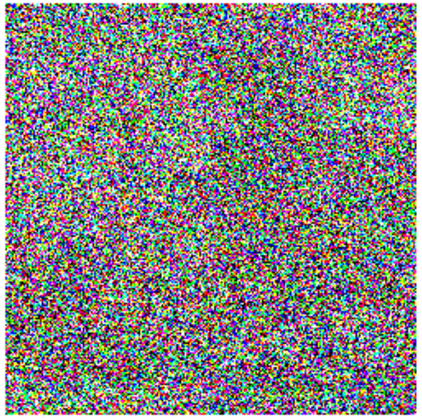
$$= \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \boxed{\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}}\boldsymbol{\epsilon}_0$$

- This inspires us to approximate the denoising transition mean as choosing the mean of p to match q: $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \boxed{\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}$

Progressive Denoising or Direct Reconstruction?

- The model predicts the noise to be removed in each step (i.e., denoising) by optimizing score matching term. This reduces to minimizing the difference between the predicted noise and the ground-truth schedule noise:

$$
\begin{aligned}
&\underset{\boldsymbol{\theta}}{\operatorname{argmin}} D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\right) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} D_{\mathrm{KL}}\left(\mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)\right) \parallel \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_q(t)\right)\right) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} D_{\mathrm{KL}} \frac{1}{2\sigma_q^2(t)}\left[\left\| \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t + \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \right\|_2^2\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}\left[\left\|\boldsymbol{\epsilon}_0 - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right\|_2^2\right]
\end{aligned}
$$



- Predicting $\boldsymbol{x}_0$ from a highly noisy $\boldsymbol{x}_t$ in one step is complex because the signal is buried under significant noise, especially at large t. By predicting the noise at each step, the model progressively refines $\boldsymbol{x}_t$ towards $\boldsymbol{x}_0$, which makes the learning task more manageable (e.g., converges better / requires smaller network capacity).

Training and Sampling from Diffusion Model

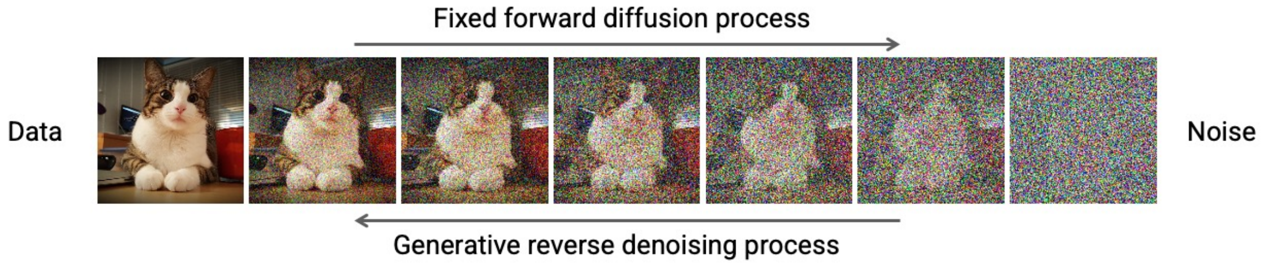
- [Ho et al., 2020] (DDPM) chooses to build the training procedure by performing SGD on the set of training images over timesteps.
- The sampling procedure iteratively executes the denoising process from a Gaussian initialization x_T.

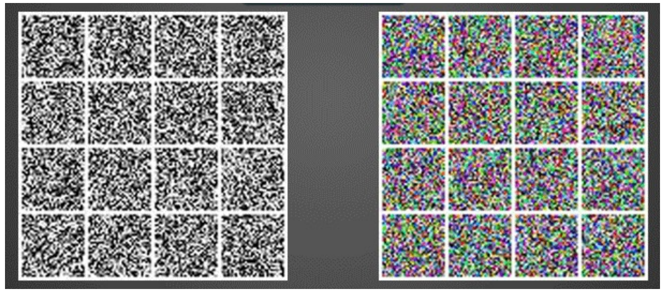
Algorithm 1 Training

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

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: $\quad \mathbf{x}_{t-1} = \underbrace{\boxed{\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)}}_{\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)} + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

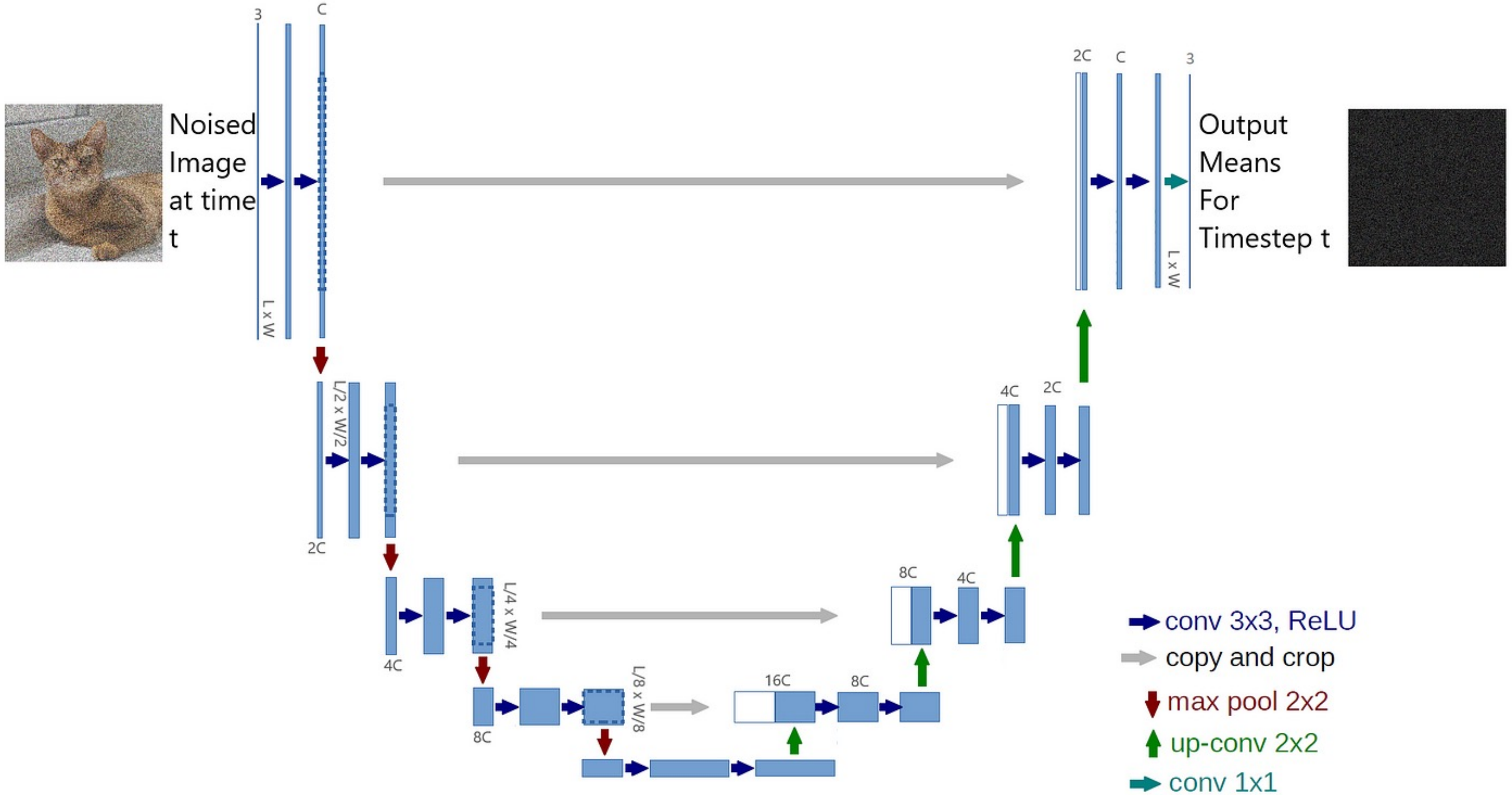




Implementation (DDPM)

DDPM uses U-Net with residual connection and self-attention layers to represent $\epsilon_\theta(\mathbf{x}_t, t)$.

The time representation is conditioned in the U-Net as sinusoidal positional embeddings or Fourier features.



Implementation (DDPM)

Scheduler for beta (β_t) indicates a predefined sequence of noise variances for each timestep t.

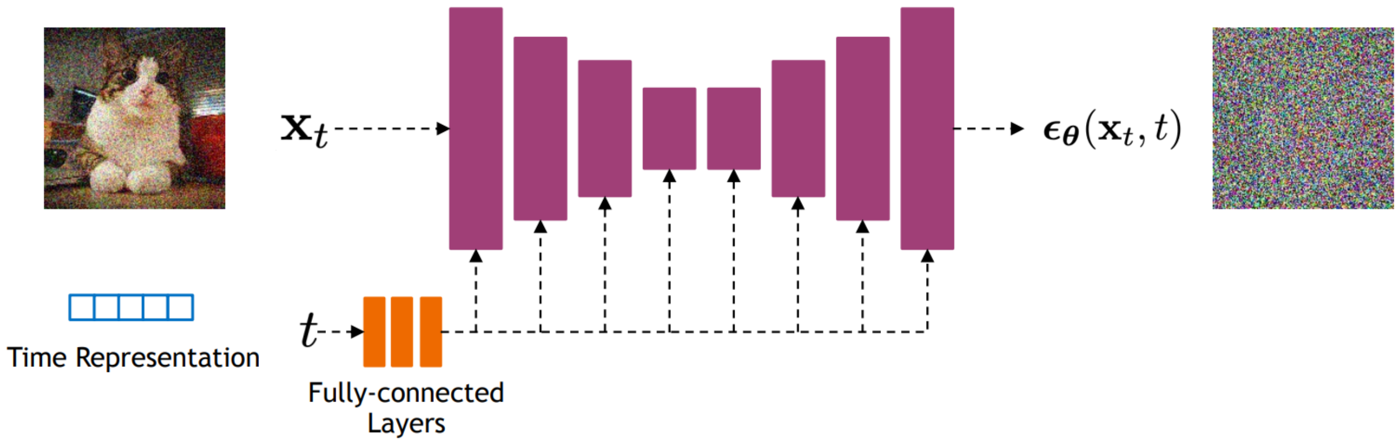
- Linear Schedule: β_t increases linearly from a small initial value to a maximum value.
- Cosine Schedule: Uses a cosine function to define β_t for smoother transitions.

Alpha Terms (α_t and $\bar{\alpha}_t$) is then derived from the beta:

- $\alpha_t = 1 - \beta_t$
- $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$

Creating Training Data as the forward diffusion (noising) process is simulated by adding Gaussian noise to images according to the noise schedule. For each training image x_0 and timestep t, we generate a noisy image x_t using the closed-form equation:

$$x_t = \sqrt{\overline{\alpha_t}} x_0 + \sqrt{1 - \overline{\alpha_t}}\, \epsilon, \epsilon \sim \mathcal{N}(\epsilon; 0, I)$$



Implementation

- Samples of DDPM