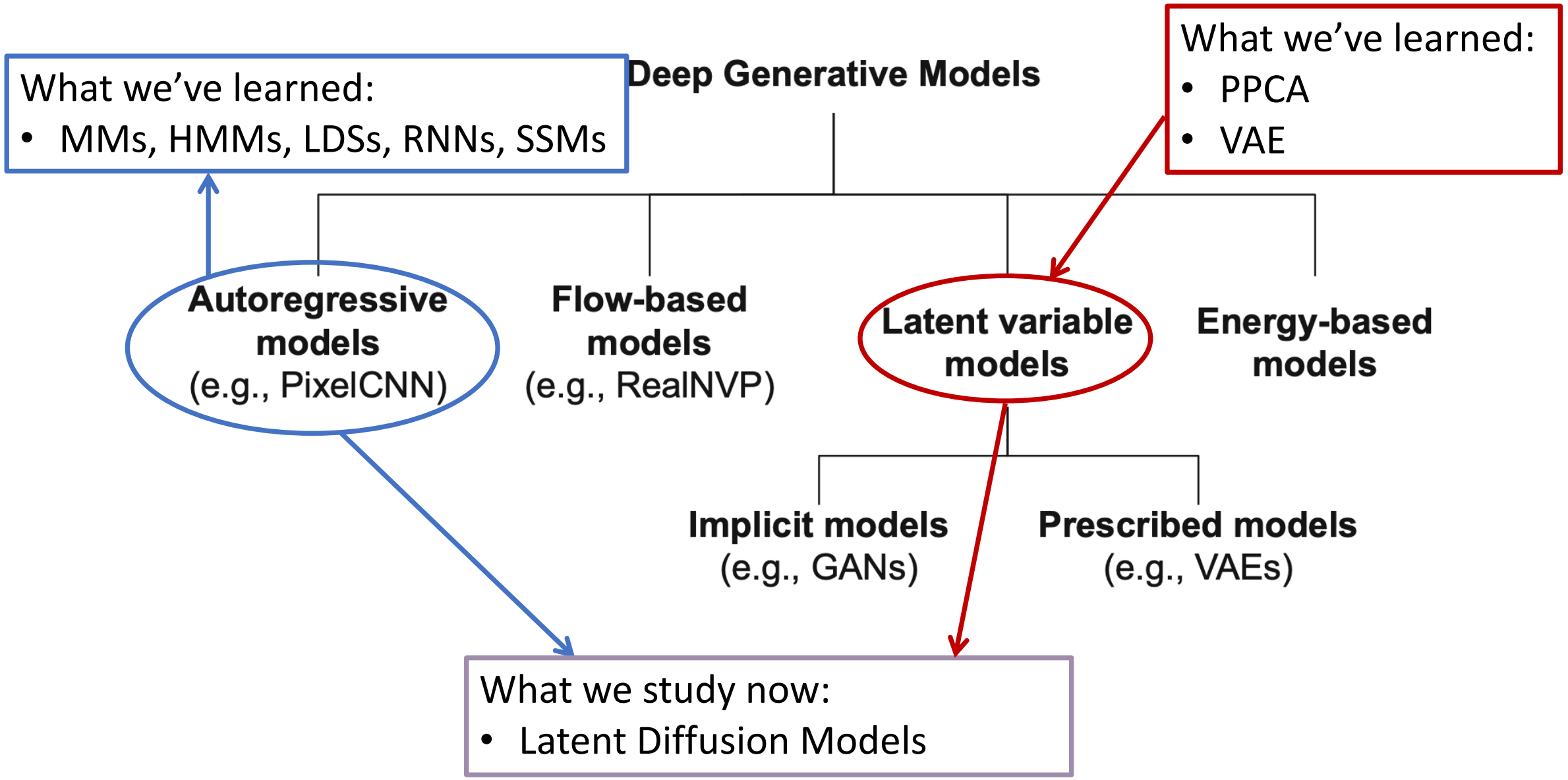
Deep Generative Models: Diffusion Models

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

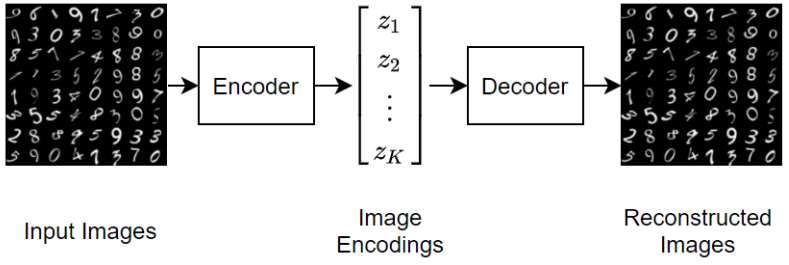
Penn
University of Pennsylvania

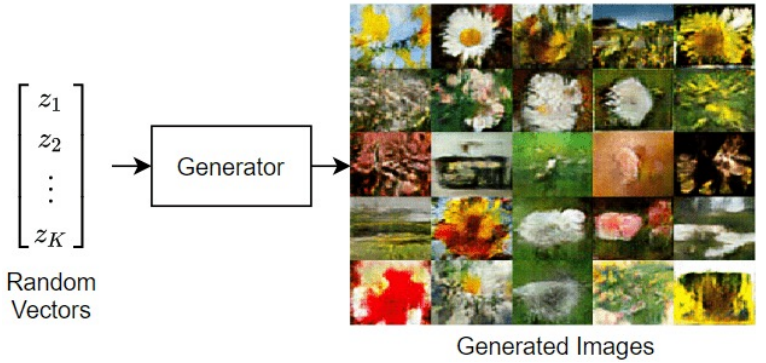
Taxonomy of Generative Models



Diffusion Models

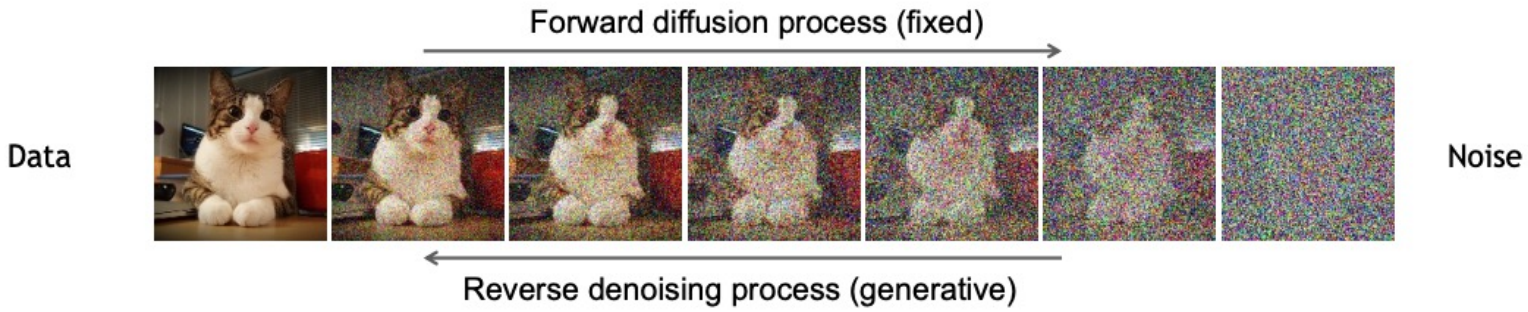
- The journey of generative models has evolved significantly in recent years.
- **Variational Autoencoders (VAEs)** introduce probabilistic modeling for latent representations but struggled with generating high-quality images.
- This led to the rise of **Generative Adversarial Networks (GANs)**, which 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.

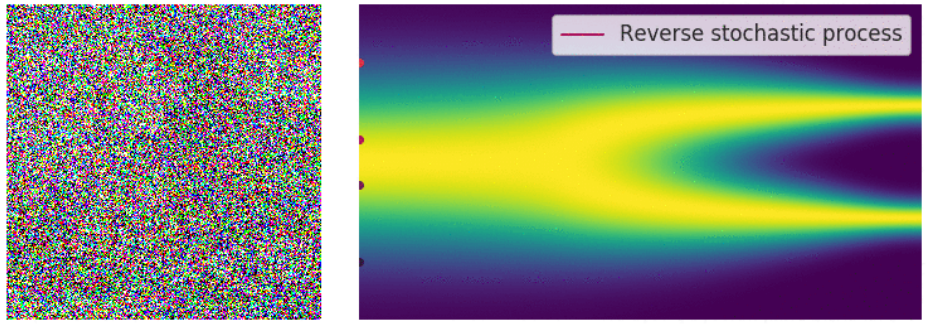




Diffusion Models

- 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 (forward) diffusion process.
- The VAE decoder maps noise to data by learning a transformation that aims to reverse the forward diffusion process.





Outline

- **Markov Hierarchical Variational Auto Encoders (MHVAEs)**
  - Autoregressive Encoder and Autoregressive Decoder of an MHVAE
  - Derivation of the ELBO of an MHVAE

- Diffusion Models as MHVAEs with a Linear Gaussian Autoregressive Latent Space
  - Forward Diffusion Process
  - Reverse Diffusion Process
  - ELBO for Diffusion Models as a particular case of the ELBO for MHVAEs
  - Implementation Details: UNet architecture, Training and Sampling Strategies

- 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

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

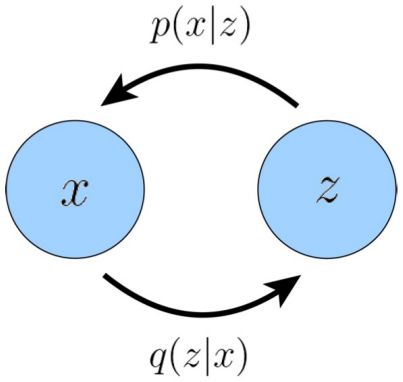
Recall the Evidence Lower Bound (ELBO)

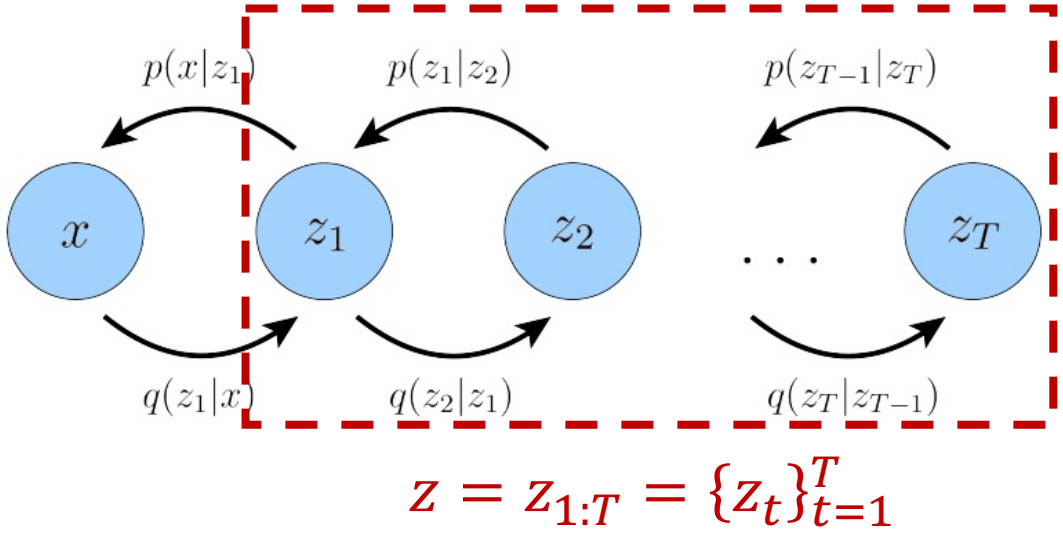
- The ELBO is the sum of a reconstruction term and a prior matching term

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

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}.



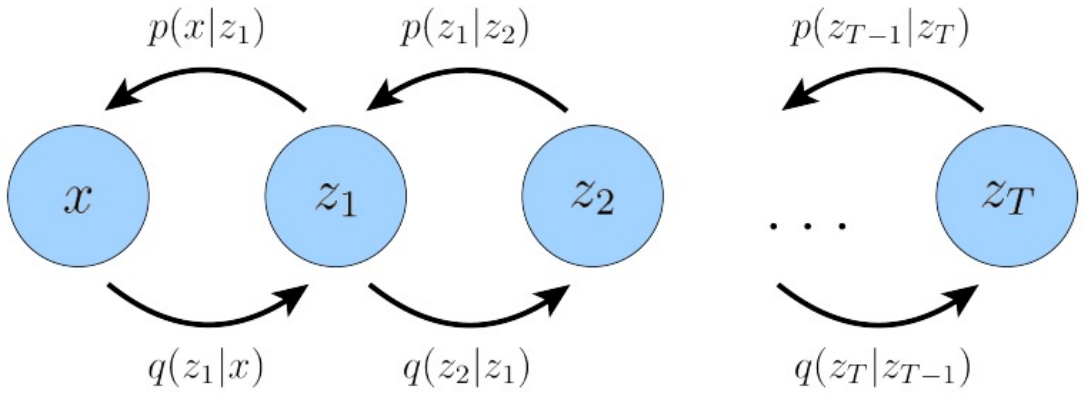


$z = z_{1:T} = \{z_t\}_{t=1}^{T}$

- What is the VAE encoder $q_\phi(z \mid x)$ of a Diffusion Model ?
- What is the VAE decoder $p_\theta(x \mid z)$ 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 an MHVAE

- Let us make the change of variables $x \rightarrow x_0$ and $z_{1:T} \rightarrow x_{1:T}$.
- The ELBO is hard to evaluate because it requires sampling from $q_\phi(x_{1:T} \mid x_0)$.

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

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

$$\underbrace{\mathbb{E}_{q_\phi(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(x_T \mid x_0) \,||\, p_\theta(x_T)\right)}_{\text{prior matching term}}$$

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

Decomposition of the ELBO for an MHVAE

- **Proof (1/2):** Reversing $q_\phi(x_t \mid x_{t-1})$

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

- Substituting $q_\phi(x_t \mid x_{t-1})$ and using telescopic product to cancel factors

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

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

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

Decomposition of the ELBO for an MHVAE

- **Proof (2/2):** expanding into three terms and simplifying expectations

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

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

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

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

Why can we Simplify Expectations?

- For the first term:

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

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

$$\int q_\phi(x_1, x_{2:T} \mid x_0)\, dx_{2:T} = q(x_1 \mid x_0)$$

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

- For the second term:

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

$$= \int \log\left(\frac{p_\theta(x_T)}{q_\phi(x_T \mid x_0)}\right) q_\phi(x_{1:T-1}, x_T \mid x_0)\, dx_{1:T-1} dx_T$$

$$\int q_\phi(x_{1:T-1}, x_T \mid x_0)\, dx_{1:T-1} = q(x_T \mid x_0)$$

$$= \int \log\left(\frac{p_\theta(x_T)}{q_\phi(x_T \mid x_0)}\right) q_\phi(x_T \mid x_0)\, dx_T = \mathbb{E}_{q_\phi(x_T|x_0)}\left[\log \frac{p_\theta(x_T)}{q_\phi(x_T \mid x_0)}\right]$$

Why can we Simplify Expectations?

- For the third term:

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

$$=\int\log\left(\frac{p_\theta(x_{t-1}\mid x_t)}{q_\phi(x_{t-1}\mid x_t,x_0)}\right)q_\phi(x_{1:t-2},x_{t-1:t},x_{t+1:T}\mid x_0)dx_{1:t-2}dx_{t+1:T}dx_{t-1}dx_t$$

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

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

$$=-\int D_{\mathrm{KL}}\left(q_\phi(x_{t-1}\mid x_t,x_0)\,||\,p_\theta(x_{t-1}\mid x_t)\right)q_\phi(x_t\mid x_0)dx_t$$

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

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

Interpretation of the ELBO of an MHVAE

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

- $\mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)]$ 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 see, the diffusion model learns the denoising step $p_\theta(x_{t-1} \mid x_t)$ as an approximation to the tractable, ground-truth denoising step $q_\phi(x_{t-1} \mid x_t, x_0)$.

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

Outline

- Markov Hierarchical Variational Auto Encoders (MHVAEs)
  - Autoregressive Encoder and Autoregressive Decoder of an MHVAE
  - Derivation of the ELBO of an MHVAE

- **Diffusion Models as MHVAEs with a Linear Gaussian Autoregressive Latent Space**
  - Forward Diffusion Process
  - Reverse Diffusion Process
  - ELBO for Diffusion Models as a particular case of the ELBO for MHVAEs
  - Implementation Details: UNet architecture, Training and Sampling Strategies

- Application of Diffusion Models
  - Stable Diffusion: Text-Conditioned Diffusion Model
  - ControlNet: Multimodal Control for Consistent Synthesis

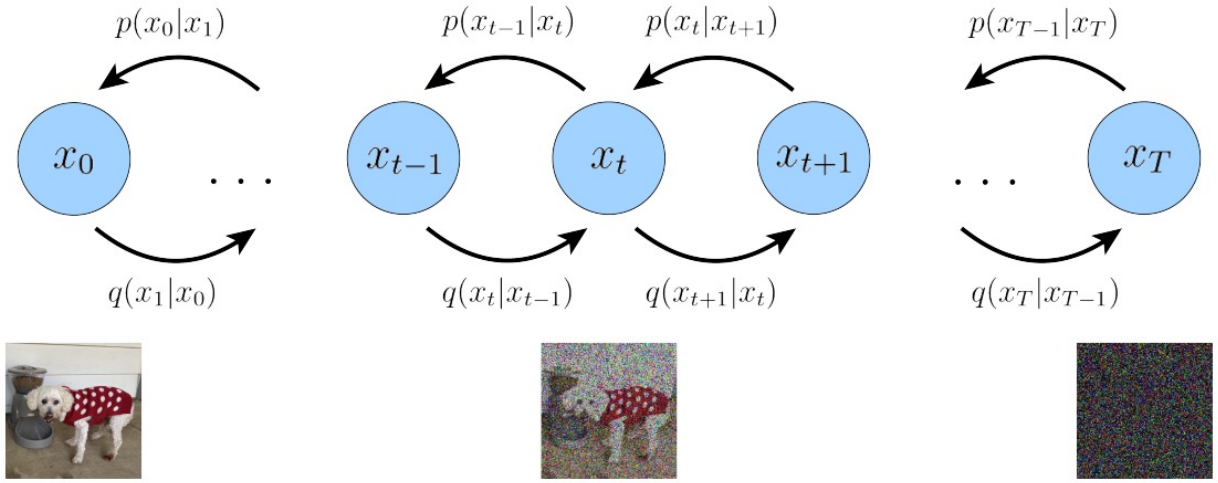
Diffusion Model as MHVAEs with Gaussian Latents

- **A Diffusion Model is an MHVAE** where the latent variables $x_{1:T}$ have the same dimension as the data x_0, and 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

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

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

- 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

- Consider the formulation of a single noising step:

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

- Since $x_1 \mid x_0$ is Gaussian, and $x_t \mid x_{t-1}$ is Gaussian, $x_t \mid x_0$ is also Gaussian.

- We can recursively derive the closed form update for $\mathbb{E}[\, x_t \mid x_0\,]$ as follows

$$\begin{aligned}
\mathbb{E}[\, x_t \mid x_0\,] &= \mathbb{E}\big[\, \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t \mid x_0\,\big] \\
&= \sqrt{\alpha_t}\ \mathbb{E}[\, x_{t-1} \mid x_0\,] + \sqrt{1-\alpha_t}\mathbb{E}[\epsilon_t] \\
&= \sqrt{\alpha_t}\mathbb{E}[\, x_{t-1} \mid x_0\,] \\
&= \sqrt{\alpha_t}\sqrt{\alpha_{t-1}}\ \mathbb{E}[\, x_{t-2} \mid x_0\,] \\
&= \sqrt{\alpha_t}\sqrt{\alpha_{t-1}} \cdots \sqrt{\alpha_1}\, x_0 \\
&= \sqrt{\bar{\alpha}_t} x_0
\end{aligned}$$

$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$$

The Forward Process of Diffusion Model

- We can recursively derive the closed form update for $\mathrm{Var}[\, x_t \mid x_0 \,]$ as follows

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

- That is:

$$\begin{aligned}\mathrm{Var}(\, x_t \mid x_0 \,) &= \alpha_t \left[\alpha_{t-1}\mathrm{Var}(\, x_{t-2} \mid x_0 \,) + (1-\alpha_{t-1})\, I\right] + (1-\alpha_t)\, I \\ &= \alpha_t\alpha_{t-1}\mathrm{Var}(\, x_{t-2} \mid x_0 \,) + (1-\alpha_t\alpha_{t-1})I \\ &= \cdots \\ &= \alpha_t\alpha_{t-1} \ldots \alpha_1\, \mathrm{Var}(\, x_0 \mid x_0 \,) + \left(1 - \prod_{i=1}^{t} \alpha_i\right) I \\ &= \left(1 - \textstyle\prod_{i=1}^{t} \alpha_i\right) I = (1-\overline{\alpha_t})I\end{aligned}$$

$\mathrm{Var}(\, x_0 \mid x_0 \,) = \mathbf{0}$

$$\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$$

The Forward Process of Diffusion Model

- Therefore, given x_0, we can sample x_t directly without having to generate all x_t's. This is because x_t is a linear Gaussian transformation of x_0 with scheduled randomness (controlled by $\overline{\alpha_t}$) drawn from a standard normal distribution:

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

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

- Moreover, we can also generate x_0 from x_t as

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

- This suggests we can reverse the noising process. However, exact reversal requires knowing the exact $\overline{\epsilon_t}$. The reverse diffusion process will be designed to predict it. In addition, the formula will be used for the reparameterization trick.

The Backward Diffusion Process

- We have designed a forward diffusion process $q_\phi(x_t \mid x_{t-1})$ that
  - At each step adds Gaussian noise to the input until it becomes pure noise
  - Allows us to sample $x_t \mid x_0$ without having to compute x_t recursively
  - Allows us to sample $x_0 \mid x_t$ without having to compute x_t recursively

- We now need to design a reverse diffusion process $p_\theta(x_{t-1} \mid x_t)$ that makes the calculation of the ELBO easy. We do this by
  - Understanding the specific structure of both the ELBO and the $q_\phi(x_t \mid x_{t-1})$
  - Making $p_\theta(x_{t-1} \mid x_t)$ match that structure

ELBO for Diffusion Model: Score Matching Term

- To compute the third term, we need

$$q(x_{t-1} \mid x_t, x_0) = \frac{q(x_t \mid x_{t-1}, x_0)\, q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)}$$

$$= \frac{\mathcal{N}\left(x_{t-1}; \frac{x_t}{\sqrt{\alpha_t}}, \frac{1-\alpha_t}{\alpha_t} I\right) \mathcal{N}\left(x_{t-1}; \sqrt{\overline{\alpha_{t-1}}} x_0, (1-\overline{\alpha_{t-1}}) I\right)}{\mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}} x_0, (1-\overline{\alpha_t}) I\right)}$$

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

$q(x_t \mid x_{t-1}, x_0) = q(x_t \mid x_{t-1})$
$= \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t) I)$
$\propto \exp(-\frac{1}{2(1-\alpha_t)} || x_t - \sqrt{\alpha_t} x_{t-1} ||^2)$
$\propto \exp(-\frac{1}{2(1-\alpha_t)} \alpha_t || x_{t-1} - \frac{x_t}{\sqrt{\alpha_t}} ||^2$
$\propto N\left(x_{t-1}; \frac{x_t}{\sqrt{\alpha_t}}, \frac{1-\alpha_t}{\alpha_t} I\right)$

$\mathcal{N}(x; \mu_1, \Sigma_1)\, \mathcal{N}(x; \mu_2, \Sigma_2) \propto \mathcal{N}(x; \bar{\mu}, \bar{\Sigma})$
$\bar{\mu} = \bar{\Sigma}(\Sigma_1^{-1} \mu_1 + \Sigma_2^{-1} \mu_2), \bar{\Sigma} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$

$\mu_1 = \frac{x_t}{\sqrt{\alpha_t}}, \Sigma_1 = \frac{1-\alpha_t}{\alpha_t} I, \mu_2 = \sqrt{\overline{\alpha_{t-1}}} x_0, \Sigma_2 = (1-\overline{\alpha_{t-1}}) I$

- Applying the product rule, we get $q(x_{t-1} \mid x_t, x_0) \propto \mathcal{N}(x_{t-1}; \mu_q, \Sigma_q)$, where

$$\Sigma_q(t) \coloneqq \mathrm{Cov}(x_{t-1} \mid x_t, x_0) = \left(\frac{\alpha_t}{1-\alpha_t} I + \frac{1}{1-\overline{\alpha_{t-1}}} I\right)^{-1} = \frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}} I$$

$$\mu_q(x_t, x_0) \coloneqq \mathbb{E}(x_{t-1} \mid x_t, x_0) = \Sigma_q \left(\frac{\alpha_t}{1-\alpha_t} I \frac{x_t}{\sqrt{\alpha_t}} + \frac{1}{1-\overline{\alpha_{t-1}}} I \sqrt{\overline{\alpha_{t-1}}} x_0\right)$$

$$= \frac{(1-\alpha_t)(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}} I \left(\frac{\sqrt{\alpha_t}}{1-\alpha_t} x_t + \frac{\sqrt{\overline{\alpha_{t-1}}}}{1-\overline{\alpha_{t-1}}} x_0\right) = \frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}}) x_t + \overline{\alpha_{t-1}}(1-\alpha_t) x_0}{1-\overline{\alpha_t}}$$

ELBO for Diffusion Model: Matching the Mean

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

- Recall KL divergence for Gaussians

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

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

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

- Choose variance of p to match exactly variance of q

$$D_{\mathrm{KL}}\left(q(x_{t-1}\mid x_t,x_0)\,||\,p_\theta(x_{t-1}\mid x_t)\right)$$
$$= D_{\mathrm{KL}}\left(\mathcal{N}\left(x_{t-1};\mu_q,\Sigma_q(t)\right)||\,\mathcal{N}\left(x_{t-1};\mu_\theta,\Sigma_q(t)\right)\right)$$
$$= \frac{1}{2\sigma_q^2(t)}\left[||\,\mu_\theta - \mu_q\,||_2^2\right]$$
$$= \frac{1}{2\sigma_q^2(t)}\frac{\overline{\alpha_{t-1}}(1-\alpha_t)^2}{(1-\overline{\alpha_t})^2}\left[||\widehat{x_\theta}(x_t,t) - x_0||_2^2\right]$$

Reparameterization as an Alternative Form for ELBO

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

$$\mu_q(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}$$

$$= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t}{\sqrt{\bar{\alpha}_t}}}{1-\bar{\alpha}_t}$$

$$= \frac{1-\bar{\alpha}_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_t$$

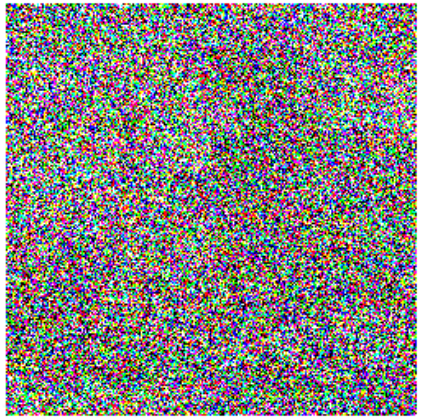
$$= \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_t$$

- This inspires us to approximate the denoising transition mean as choosing the mean of p to match q: $\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_t$

Progressive Denoising or Direct Reconstruction?

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

$$
\begin{aligned}
&\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\big(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t)\big) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\Big(\mathcal{N}\big(x_{t-1}; \mu_q, \Sigma_q(t)\big) \parallel \mathcal{N}\big(x_{t-1}; \mu_\theta, \Sigma_q(t)\big)\Big) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2\sigma_q^2(t)} \left[\left\| \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}} \hat{\epsilon}_\theta(x_t, t) - \frac{1}{\sqrt{\alpha_t}} x_t + \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}} \epsilon_0 \right\|_2^2\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t} \left[\|\epsilon_0 - \hat{\epsilon}_\theta(x_t, t)\|_2^2\right]
\end{aligned}
$$



- Predicting x_0 from a highly noisy x_t in one step is complex, as the signal is dominated from significant noise for large t.
- Predicting the noise at each step and refining x_t towards x_0 makes the learning task more manageable (e.g., converges better or requires smaller network capacity).

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

Outline

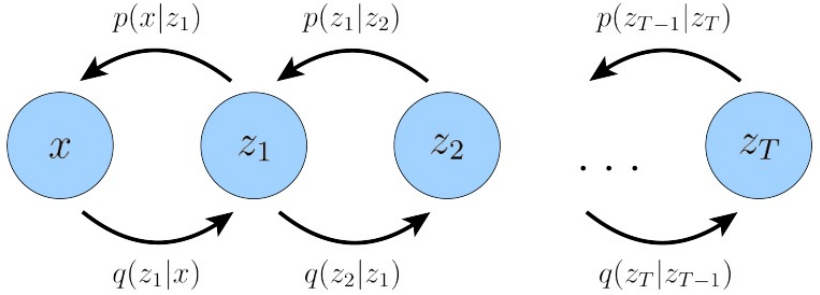
- Markov Hierarchical Variational Auto Encoders (MHVAEs)
  - Autoregressive Encoder and Autoregressive Decoder of an MHVAE
  - Derivation of the ELBO of an MHVAE

- Diffusion Models as MHVAEs with a Linear Gaussian Autoregressive Latent Space
  - Forward Diffusion Process
  - Reverse Diffusion Process
  - ELBO for Diffusion Models as a particular case of the ELBO for MHVAEs
  - **Implementation Details: UNet Architecture, Training and Sampling Strategies**

- Application of Diffusion Models
  - Stable Diffusion: Text-Conditioned Diffusion Model
  - ControlNet: Multimodal Control for Consistent Synthesis

Implementation (DDPM)

- The Denoising Diffusion Probabilistic Model (DDPM) fixes the noise variances α_t of the forward process and learns only the backward (denoising) process [Ho et al., 2020].



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

- For the backward process, we use reparameterization of ELBO

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

$$= \underbrace{-\textstyle\sum_{t=1}^{T} \mathbb{E}_{q_\phi(x_t|x_0)}\left[D_{\mathrm{KL}}\left(q_\phi(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\right)\right]}_{\text{reconstruction term} + \text{score matching term}} = -\textstyle\sum_{t=1}^{T} \mathbb{E}_{q_\phi(x_t|x_0)}\left[\frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}\left[\|\epsilon_0 - \hat{\epsilon}_\theta(x_t, t)\|_2^2\right]\right]$$

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

- For each x, we need to compute T terms in the sum, which is expensive for typical values of T (T= 1000).

SGD over time-steps

- DDPM minimizes the ELBO efficiently by performing SGD over the set of timesteps $[T] = \{1, 2, \ldots, T\}$.
- Using linearity of expectation and letting $t \sim \text{Unif}([T])$

$$\log p\,(x) \geq -D_{\text{KL}}\left(q_\phi(\,x_T \mid x_0\,) \,||\, p_\theta(x_T)\right) - \sum_{t=1}^{T} \mathbb{E}_{q_\phi(\,x_t|x_0\,)}\left[\frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}\|\epsilon_0 - \hat{\epsilon}_\theta(x_t,t)\|_2^2\right]$$

$$= -D_{\text{KL}}\left(q_\phi(\,x_T \mid x_0\,) \,||\, p_\theta(x_T)\right) - \mathbb{E}_{q_\phi(\,x_t|x_0\,)}\left[\sum_{t=1}^{T}\frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}\|\epsilon_0 - \hat{\epsilon}_\theta(x_t,t)\|_2^2\right]$$

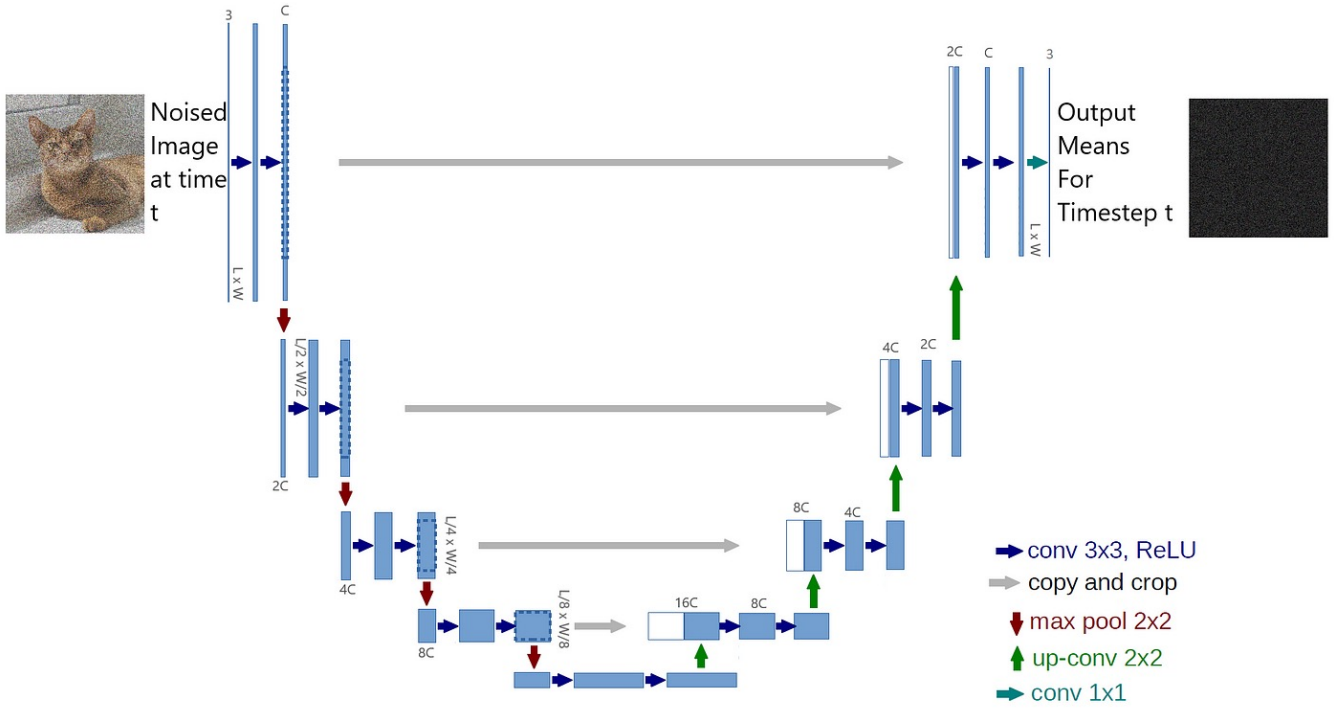
$$= -D_{\text{KL}}\left(q_\phi(\,x_T \mid x_0\,) \,||\, p_\theta(x_T)\right) - \mathbb{E}_{q_\phi(\,x_t|x_0\,),\,t}\left[\frac{T}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}\|\epsilon_0 - \hat{\epsilon}_\theta(x_t,t)\|_2^2\right]$$

- Thus, the sampling procedure computes only 1 term instead of T terms.
- The term $-D_{\text{KL}}\left(q_\phi(\,x_T \mid x_0\,) \,||\, p_\theta(x_T)\right)$ is constant during training, since $q_\phi(\,x_T \mid x_0\,)$ is not learnable.
- Thus, the simplified (unweighted) learning objective used in DDPM is

$$\mathbb{E}_{q_\phi(\,x_t|x_0\,),\,t}\left[\|\epsilon_0 - \hat{\epsilon}_\theta(x_t,t)\|_2^2\right]$$

Architecture of DDPM

- DDPM uses a U-Net with residual connection and self-attention layers to represent the noise $\overline{\epsilon_\theta}(x_t, t)$.
- Since the parameter $\overline{\epsilon_\theta}(x_t, t)$ depends both on x_t and on time t, a time embedding needs to be provided as input to each unit of the U-Net.

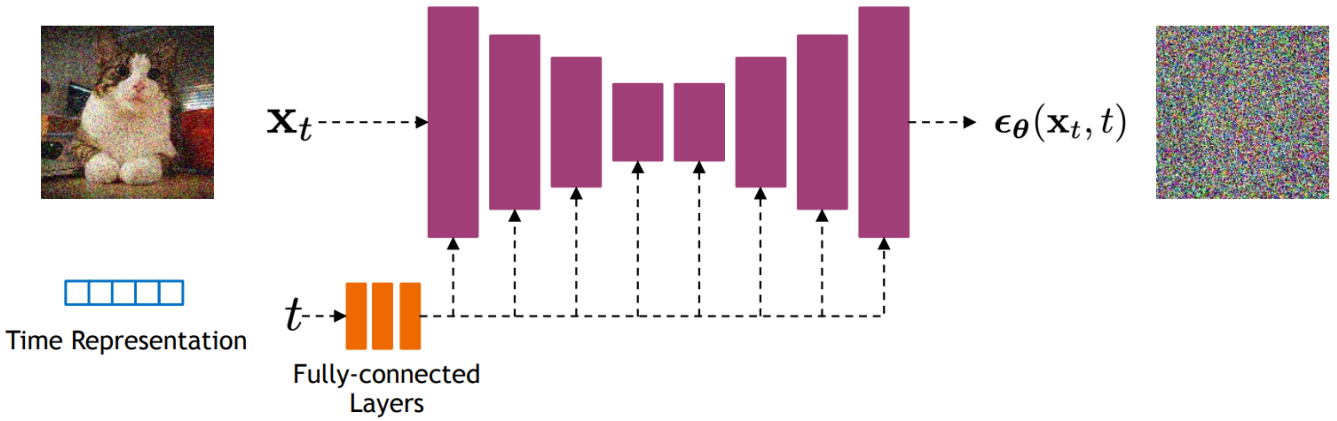


Time Encoding

- The time representation is implemented using sinusoidal positional embeddings.
- Given a time step $t \in [T]$ and an embedding dimension d, the sinusoidal positional embedding $\mathrm{SPE}(t)$ is given for $i = 0, 1, \ldots, \frac{d}{2} - 1$ as

$$\mathrm{SPE}(t)_{2i} = \sin\left(\frac{t}{10^{\frac{6i}{d}}}\right), \mathrm{SPE}(t)_{2i+1} = \cos\left(\frac{t}{10^{\frac{6i}{d}}}\right)$$

- This produces an fixed-length vector $\mathrm{SPE}(t)$ that encodes the time smoothly, since small changes in t cause predictable oscillations in the time embedding.
- The time embedding is given as input to all units of the U-Net.



Training and Sampling in DDPM

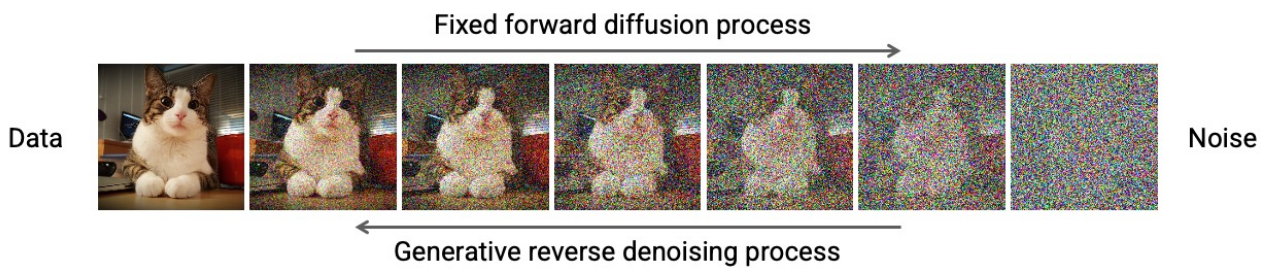
- DDPM implements the training procedure by performing SGD on the set of training images over timesteps.
- The sampling procedure executes iteratively the denoising process from a Gaussian initialization x_T.

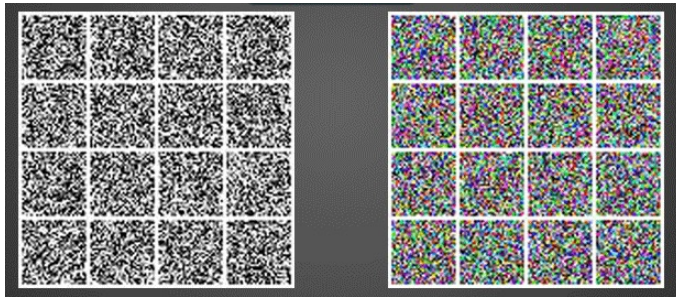
Algorithm 1 Training

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
$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
x_t
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: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
$\mu_\theta(x_t, t)$ $\sigma_t = \sqrt{\beta_t}$
5: **end for**
6: **return** $\mathbf{x}_0$





DDPM Noise Scheduler

- **Noise Scheduler for Forward Process**: in DDPM, the scheduler refers to how the noise variance ($\beta_t = 1 - \alpha_t$) changes across the diffusion timesteps $t = 1, \dots, T$.
  - **Linear Schedule**: β_t increases linearly from initial value (10^{-4}) to a maximum value (0.02)

$$\beta_t = \beta_{\text{start}} + t\frac{\beta_{\text{end}} - \beta_{\text{start}}}{T-1}, t = 0, \dots, T-1$$

  - **Cosine Schedule**: Uses a cosine function to define β_t, which better preserves signal early in the process and decays more gently near the end. It helps maintain more information in the intermediate steps, improve sample quality, require fewer steps for comparable results.

$$f(t) = \cos\left(\frac{t/T+s}{1+s} \cdot \frac{\pi}{2}\right), t = 0, \dots, T-1, \qquad \overline{\alpha_t} = \frac{f(t)}{f(0)}, \qquad \beta_t = 1 - \frac{\overline{\alpha_t}}{\overline{\alpha_{t-1}}}, \qquad s = 0.008.$$

- **Training Data**: For each training image x_0 and timestep t, a noisy image x_t is generated as: $x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\,\epsilon, \quad \epsilon \sim \mathcal{N}(\epsilon; 0, I)$.

Training Curves & Test Metrics

- The Inception Score (IS) and Fréchet Inception Distance (FID) are two common metrics for evaluating the quality of the generated images.
- IS measures how realistic and diverse the generated samples are.
- The FID measures how close the distribution of the generated images is to the real data distribution.
- Desiderata: High IS and low FID for realistic and diverse images.

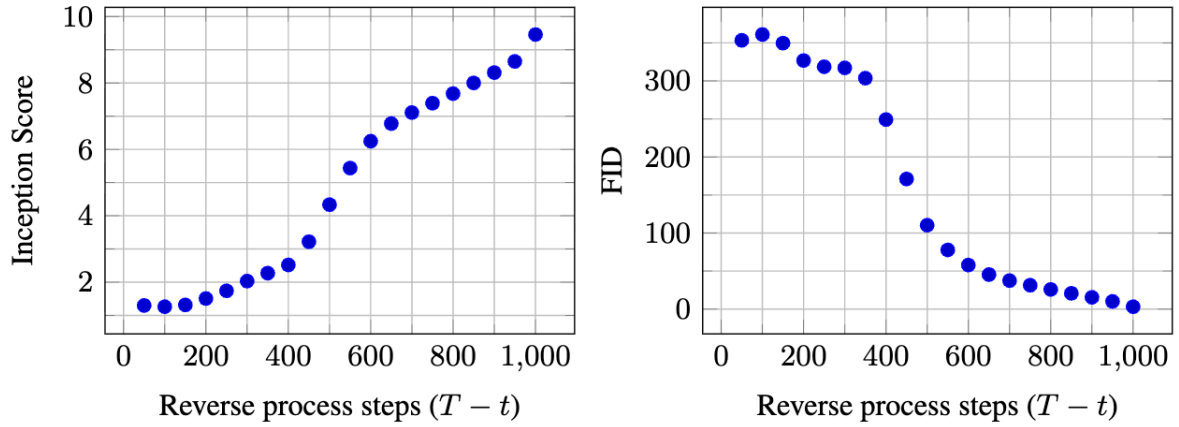


Figure 10: Unconditional CIFAR10 progressive sampling quality over time

Table 1: CIFAR10 results. NLL measured in bits/dim.

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [11]	8.30	37.9	
JEM [17]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	**10.06**	**2.67**	
Unconditional			
Diffusion (original) [53]			$\leq$ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			**2.80**
PixelIQN [43]	5.29	49.46	
EBM [11]	6.78	38.2	
NCSNv2 [56]		31.75	
NCSN [55]	8.87$\pm$0.12	25.32	
SNGAN [39]	8.22$\pm$0.05	21.7	
SNGAN-DDLS [4]	9.09$\pm$0.10	15.42	
StyleGAN2 + ADA (v1) [29]	**9.74** $\pm$ 0.05	3.26	
Ours (L, fixed isotropic $\mathbf{\Sigma}$)	7.67$\pm$0.13	13.51	$\leq$ 3.70 (3.69)
Ours (L_{simple})	9.46$\pm$0.11	**3.17**	$\leq$ 3.75 (3.72)

Generated Samples of DDPM