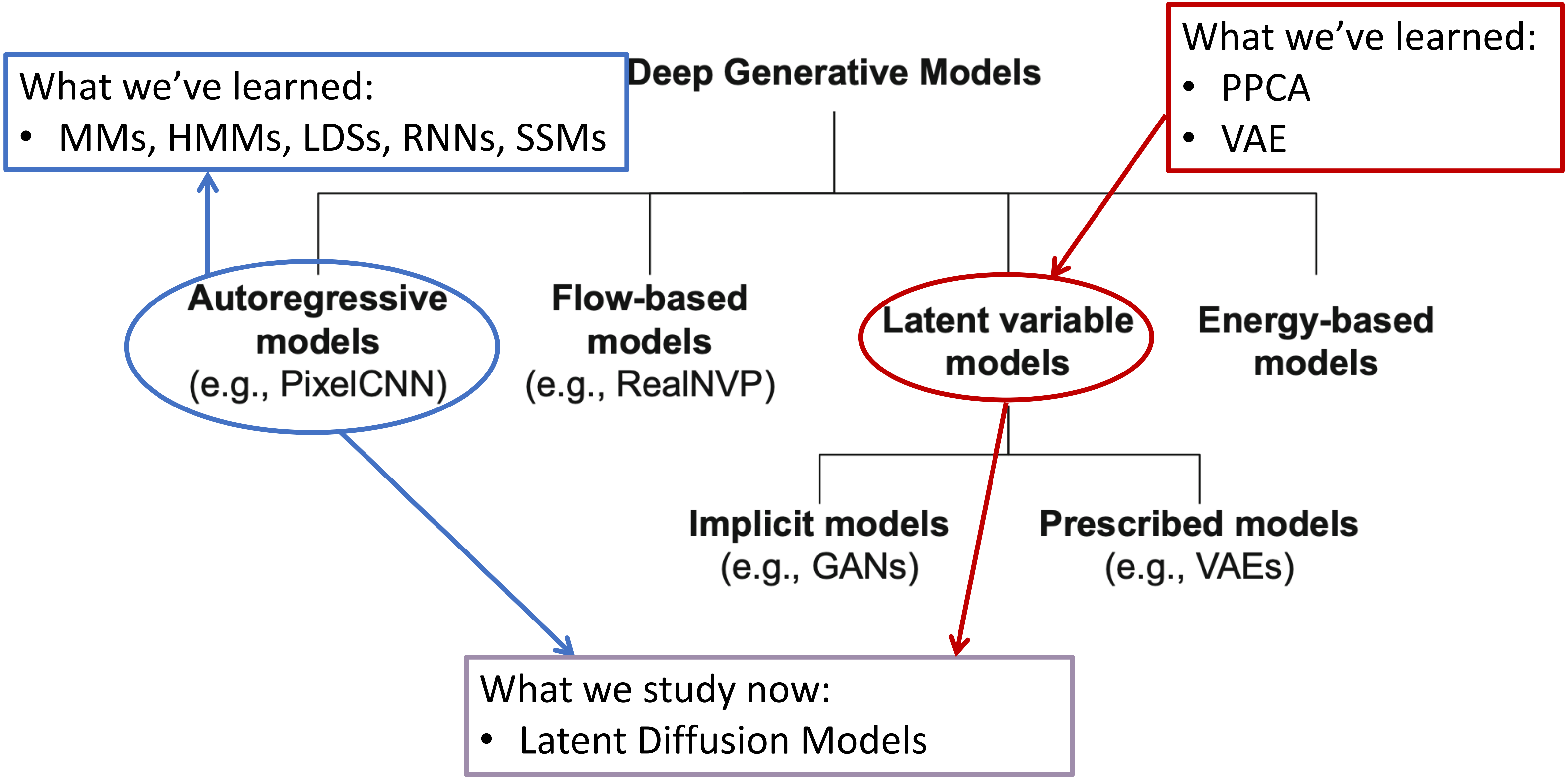
Deep Generative Models: Diffusion Models

Fall Semester 2025

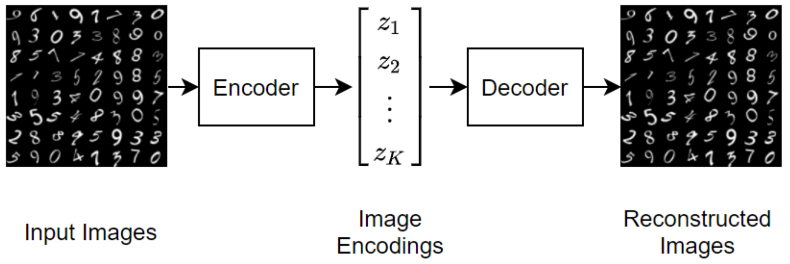
René Vidal
Director of the Center for Innovation in Data Engineering and Science (IDEAS)
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE

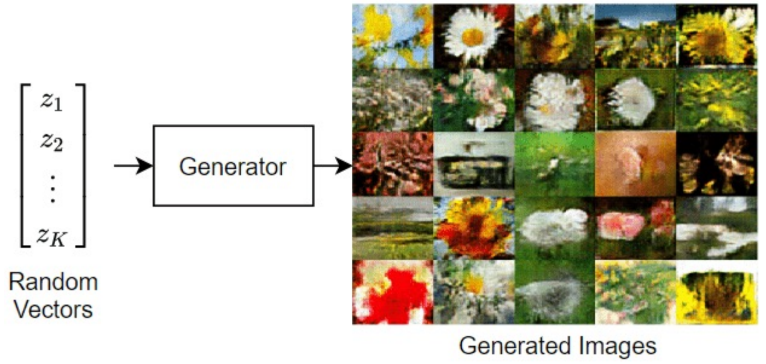
Taxonomy of Generative Models



Diffusion Models

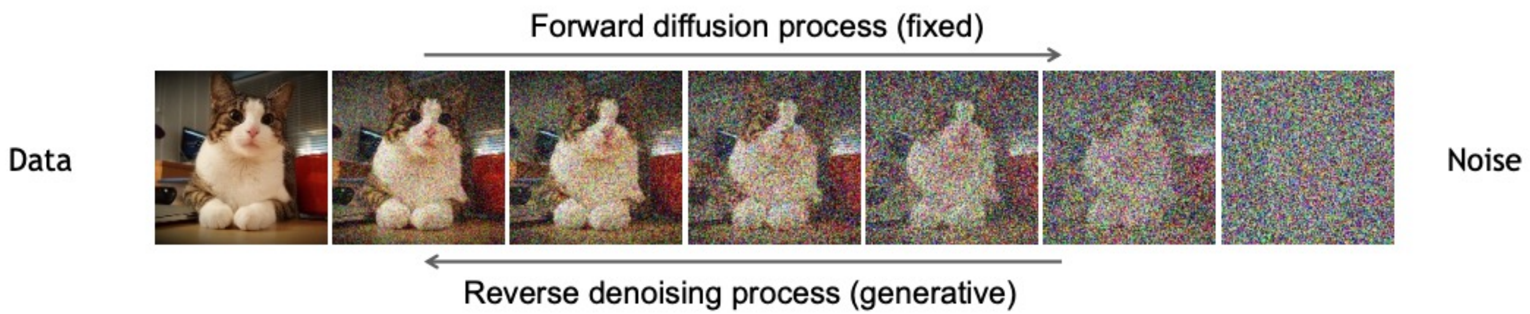
- The journey of generative models has evolved significantly in recent years.
- **Variational Autoencoders (VAEs)** introduce probabilistic modeling for latent representations but struggled with generating high-quality images.
- This led to the rise of **Generative Adversarial Networks (GANs)**, which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

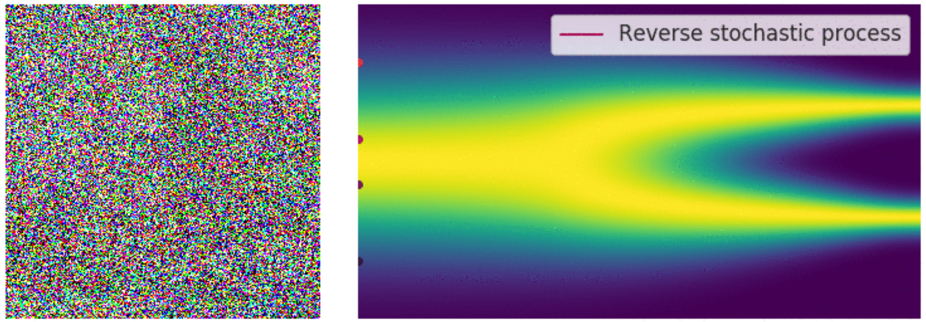




Diffusion Models

- 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

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

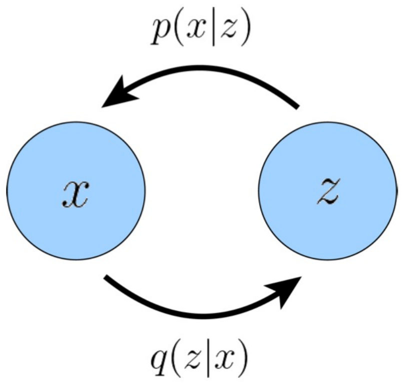
Recall the Evidence Lower Bound (ELBO)

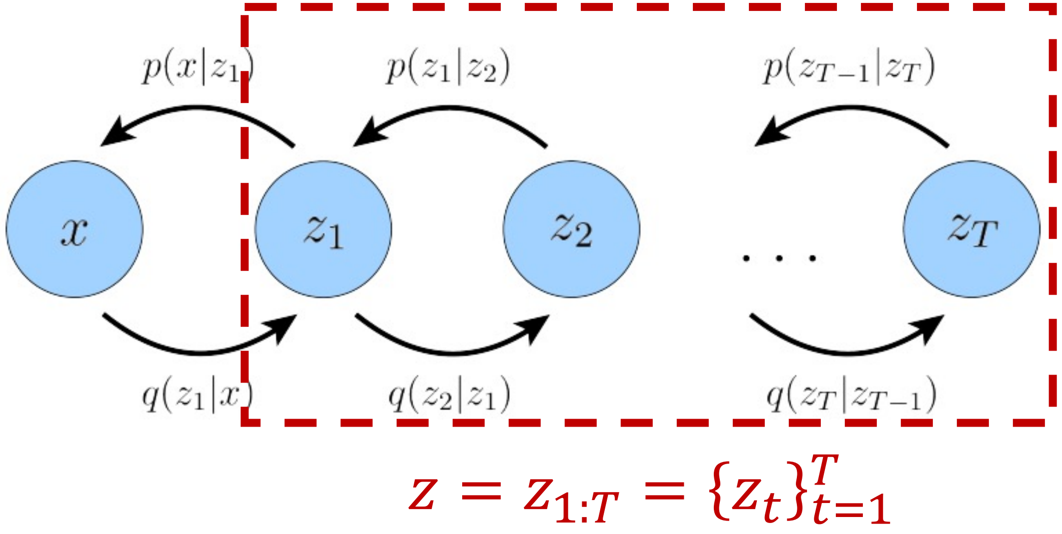
- The ELBO is the sum of a reconstruction term and a prior matching term

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

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}.

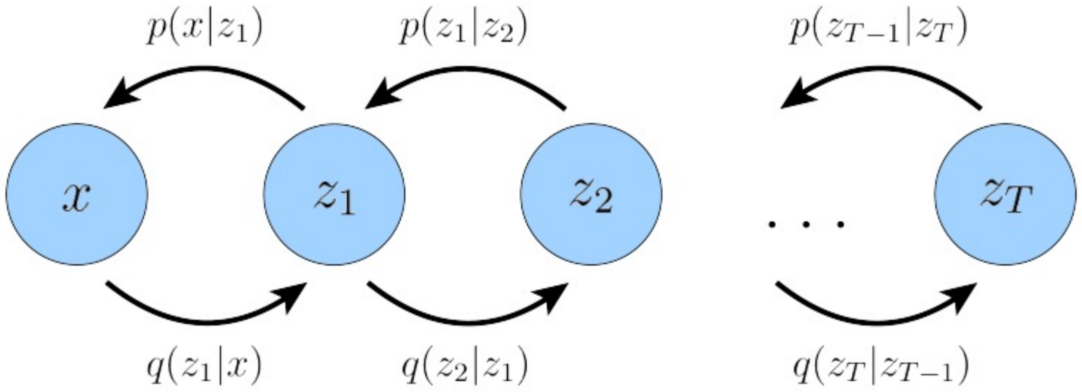




- What is the VAE encoder $q_\phi(z \mid x)$ of a Diffusion Model ?
- What is the VAE decoder $p_\theta(x \mid z)$ of a Diffusion Model ?
- What is the ELBO of a Diffusion Model ?

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 $x \to x_0$ and $z_{1:T} \to x_{1:T}$.
- 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

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

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

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

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

Why can we Simplify Expectations?

- **For the first term**:

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

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

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

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

- **For the second term**:

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

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

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

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

Why can we Simplify Expectations?

- **For the third term**:

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

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

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

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

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

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

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

Interpretation of the ELBO of an MHVAE

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

- $\mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0 \mid x_1)]$ 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.
- $D_{\text{KL}}\left(q_\phi(x_T \mid x_0) \,||\, p_\theta(x_T)\right)$ represents how close the distribution of the final latent distribution is to the standard Gaussian prior.
- $\mathbb{E}_{q_\phi(x_t|x_0)}\left[D_{\text{KL}}\left(q_\phi(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)\right)\right]$ 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 $q_\phi(x_{t-1} \mid x_t, x_0)$.