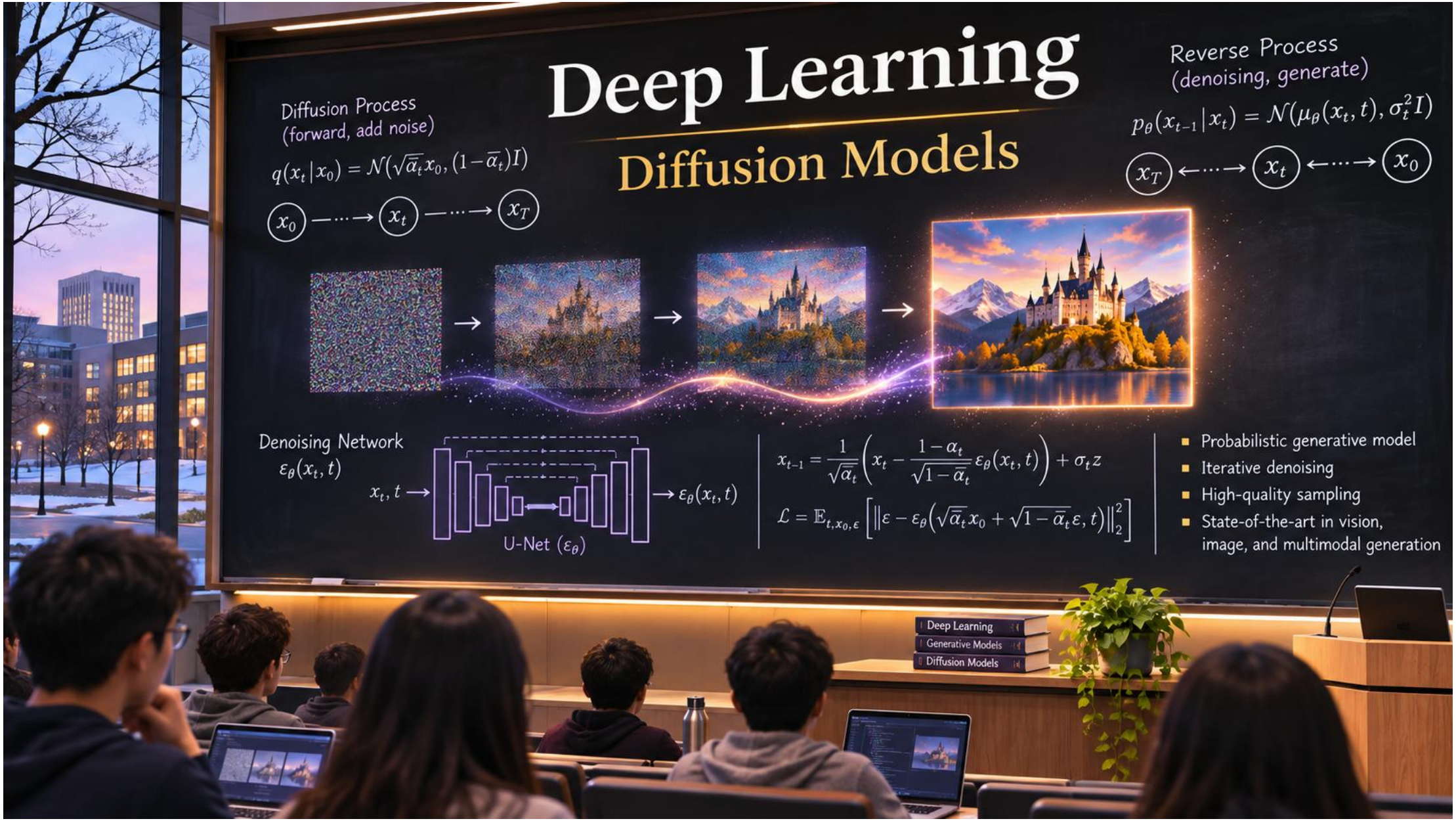
Lecture 15

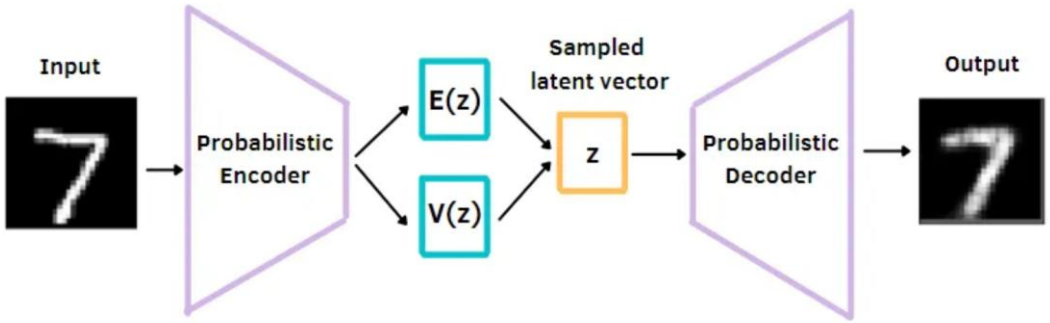
Diffusion Models

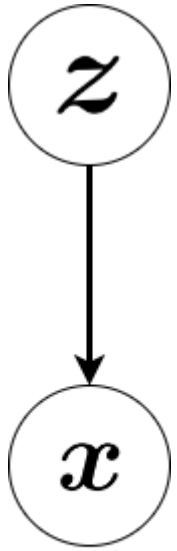
- Diffusion models take their inspiration from the principles of non-equilibrium thermodynamics

- Diffusion models use a Markov chain to gradually introduce noise into data, then learn to reverse this process to recreate the desired data from the noise.


Deep Learning
Diffusion Models
Diffusion Process
(forward, add noise)
$q(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$
$x_0 \longrightarrow x_t \longrightarrow x_T$
Reverse Process
(denoising, generate)
$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t,t), \sigma_t^2 I)$
$x_T \longleftarrow x_t \longleftarrow x_0$
Denoising Network
$\varepsilon_\theta(x_t, t)$
$x_t, t \rightarrow$
$\rightarrow \varepsilon_\theta(x_t, t)$
U-Net (ε_θ)
$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t,t)\right) + \sigma_t z$
$\mathcal{L} = \mathbb{E}_{t,x_0,\varepsilon}\left[\left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t\right)\right\|_2^2\right]$
Probabilistic generative model
Iterative denoising
High-quality sampling
State-of-the-art in vision, image, and multimodal generation
Deep Learning
Generative Models
Diffusion Models

Variational Autoencoder





VAE

Image: Eugenia Anello

Diffusion Models

$$x_0 \longrightarrow x_1 \longrightarrow \cdots \longrightarrow x_T$$

Data



Noise

Image: Arash Vahdat

Diffusion Models

$$x_0 \longrightarrow x_1 \longrightarrow \cdots \longrightarrow x_T$$

Data  Noise

$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$

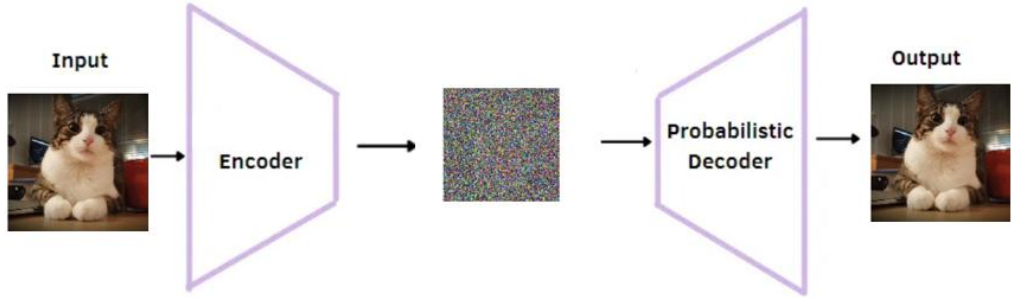
Image: Arash Vahdat

Diffusion Models vs VAE

$$x_0 \longrightarrow x_1 \longrightarrow \cdots \longrightarrow x_T$$

Data  Noise

$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$



Diffusion Models vs VAE

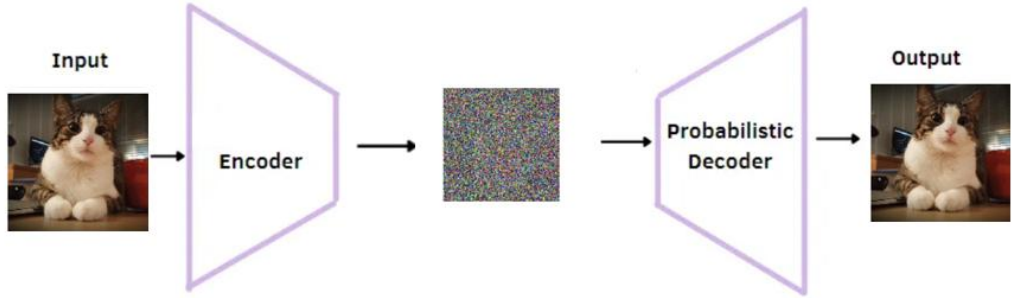
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x_t}; \sqrt{1-\beta_t}\mathbf{x_{t-1}}, \beta_t\mathbf{I})$$

Data



Noise

$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$



Diffusion Models vs VAE

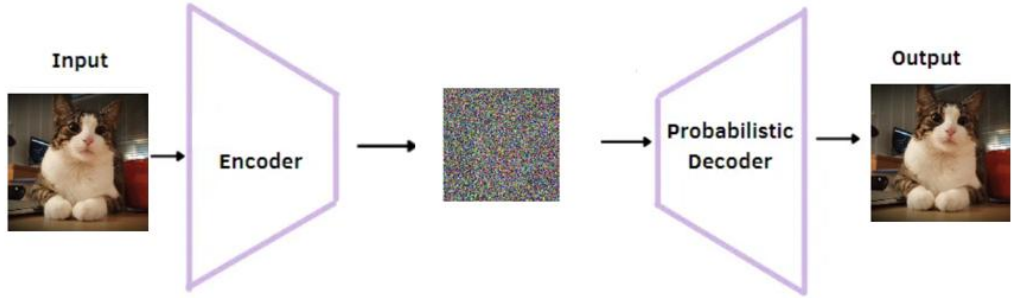
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x_t}; \sqrt{1-\beta_t}\mathbf{x_{t-1}}, \beta_t\mathbf{I})$$

Data

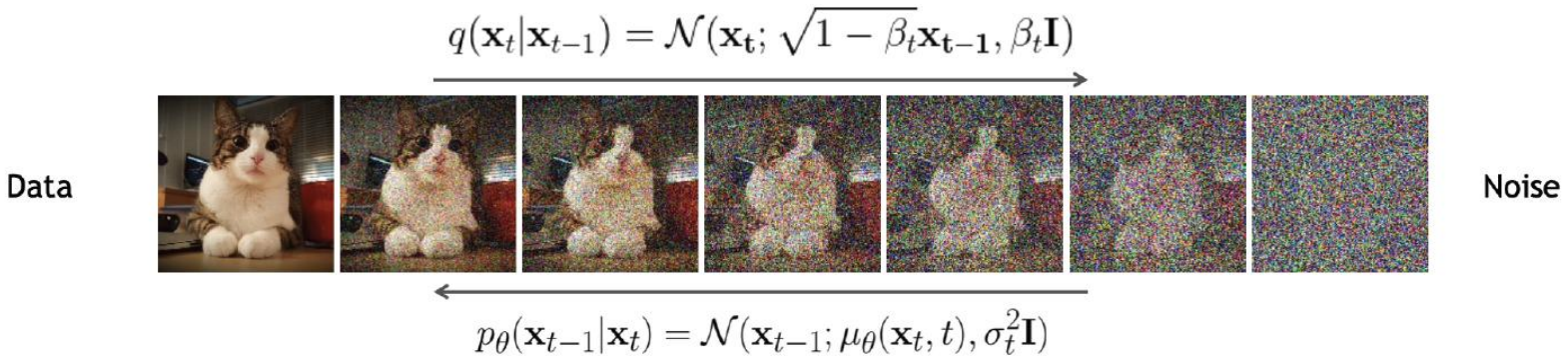


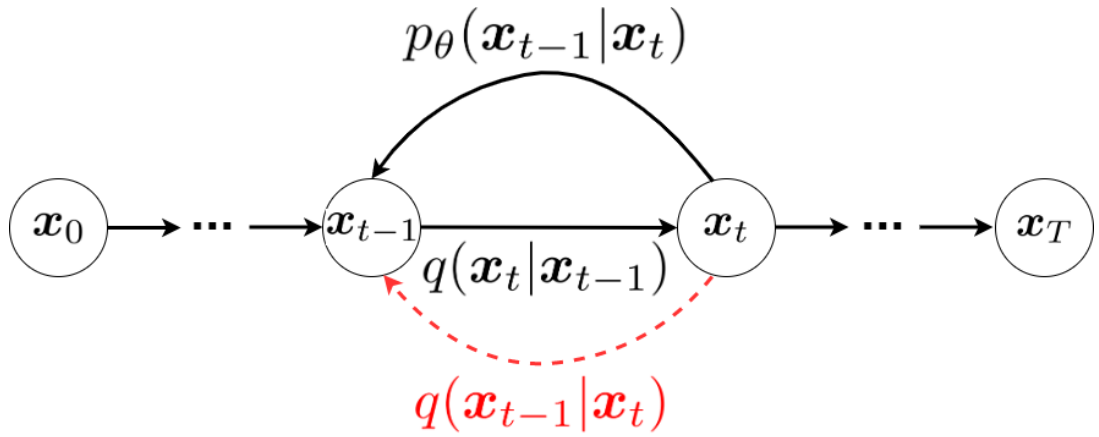
Noise

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$$



Denoising Diffusion Probabilistic Models (DDPMs)





The Forwards Encoder Process in DDPMs

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

Data



Noise

The forward process is defined by the equation:

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

Where:

- β_t is a noise level parameter, part of a predetermined noise schedule.

$$\beta_1 < \beta_2 < \cdots < \beta_T$$

$$\beta_t \in (0, 1)$$

Joint Distribution over Latent States in DDPMs

The joint distribution is expressed as:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

Where:

- $x_{1:T}$ represents all latent states from time 1 to T.
- x_0 is the initial state.

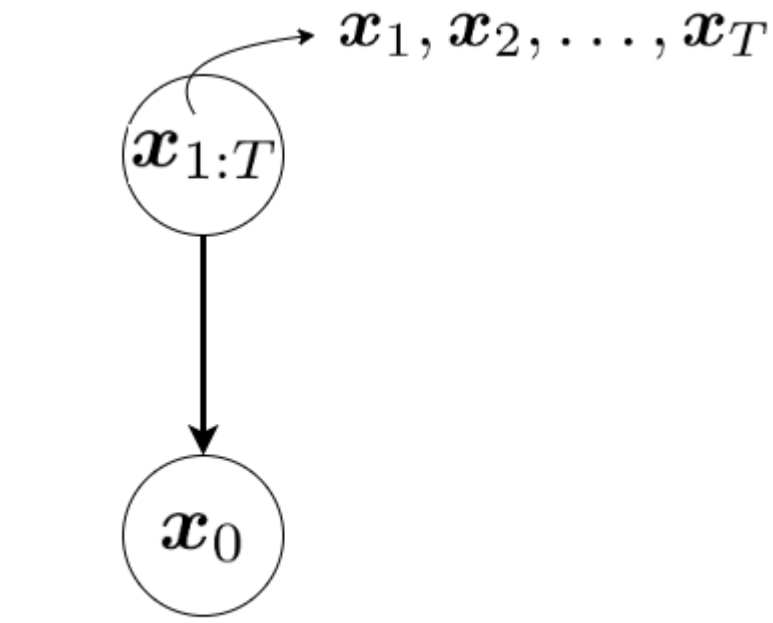
Joint Distribution over Latent States in DDPMs

The joint distribution is expressed as:

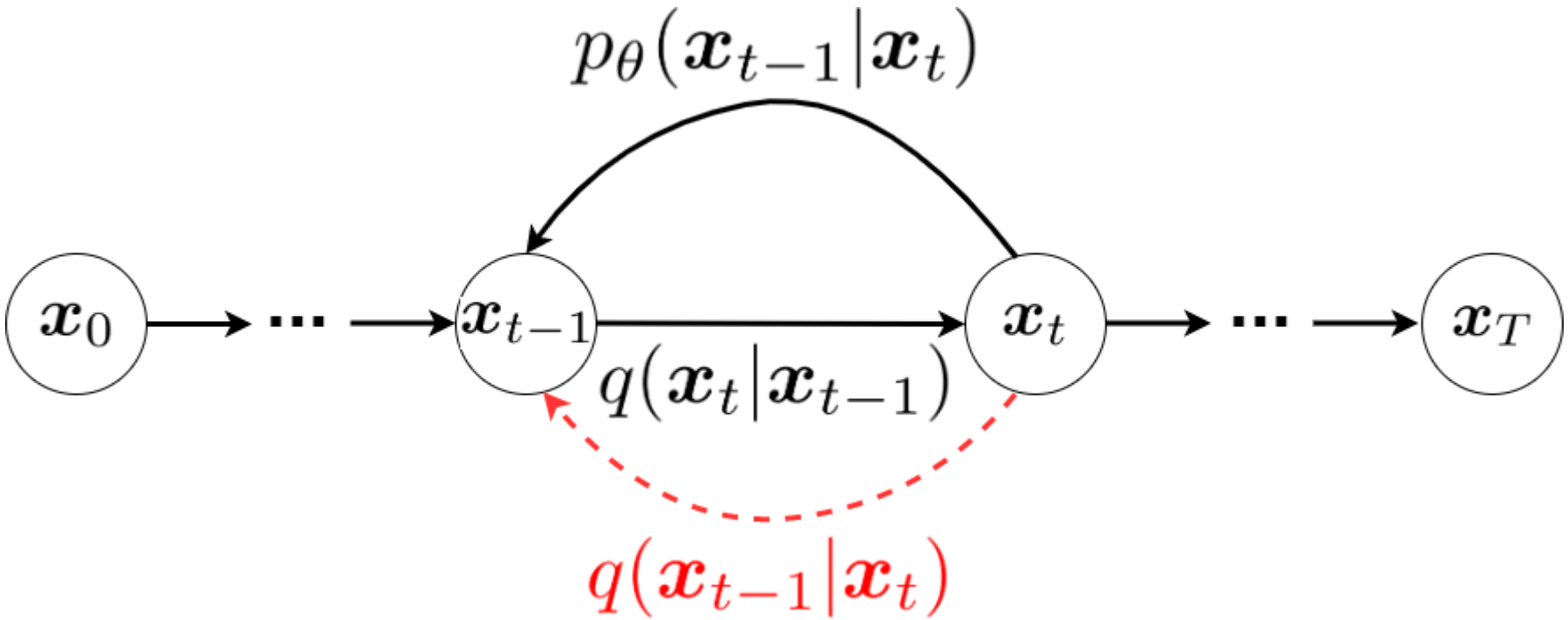
$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

Where:

- $x_{1:T}$ represents all latent states from time 1 to T.
- x_0 is the initial state.



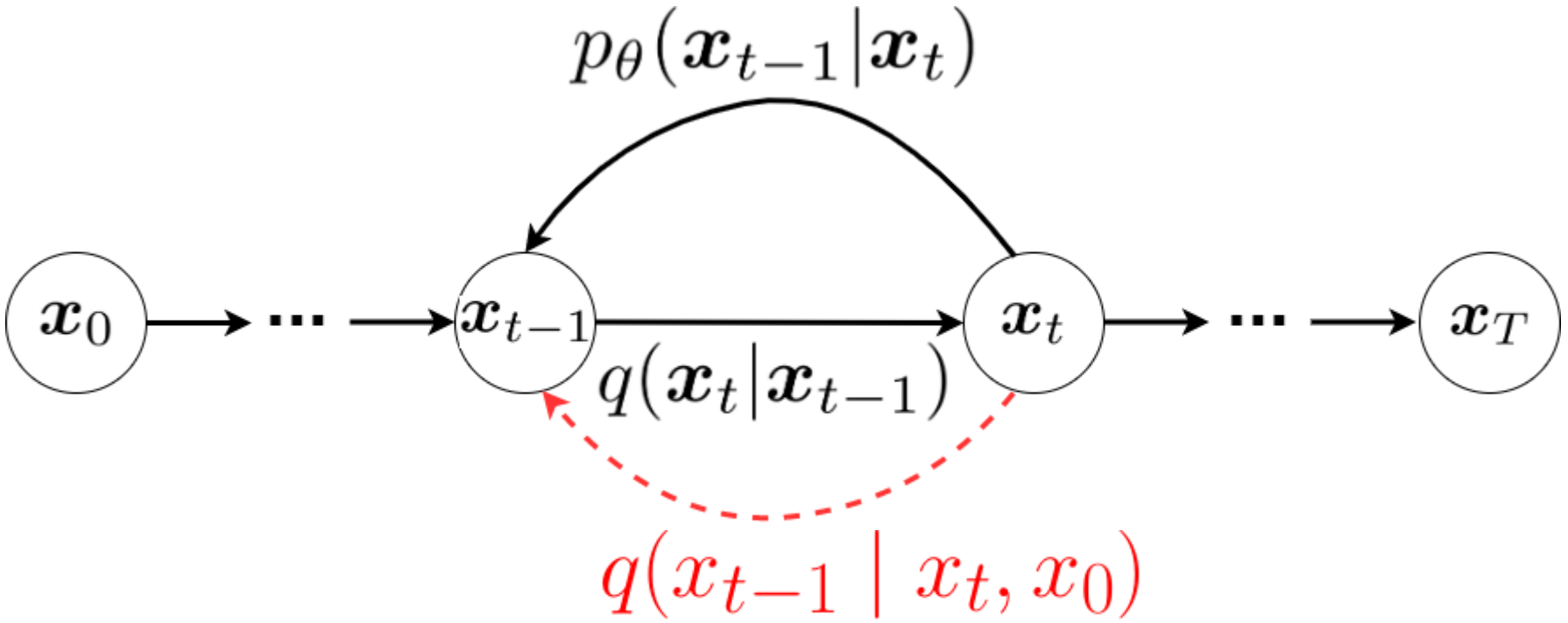
Diffusion model


$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$
$\boldsymbol{x}_0$
$\boldsymbol{x}_{t-1}$
$\boldsymbol{x}_t$
$\boldsymbol{x}_T$
$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$
$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

A Tractable Reverse Target

$q(x_{t-1} \mid x_t)$ is generally intractable. With the clean x_0 known,

$q(x_{t-1} \mid x_t, x_0)$ is a tractable Gaussian target for $p_\theta(x_{t-1} \mid x_t)$, for $t \geq 2$.



The network uses the noisy state and timestep. The clean observation supplies the target.

The Training Goal: Matching the Means

For each intermediate step $t \geq 2$, train the reverse model by minimizing

$$\min_{\theta} \mathbb{E}_{\substack{x_0 \sim q_{\text{data}} \\ x_t \sim q(x_t|x_0)}} D_{\text{KL}}(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)) \,.$$

For two Gaussians with the same fixed isotropic covariance,

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \sigma_t^2 I),$$
$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I),$$

$$D_{\text{KL}}(q \| p_\theta) = \frac{1}{2\sigma_t^2} \| \tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t) \|^2.$$

We therefore match the means. First compute the target $\tilde{\mu}_t$, then learn μ_θ.

Computing the Target Mean: Gaussian Product

$$\alpha_t = 1 - \beta_t, \qquad \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i, \qquad \bar{\alpha}_t = \alpha_t \bar{\alpha}_{t-1}.$$

Bayes' rule gives the tractable posterior for $t \geq 2$:

$$q(x_{t-1} \mid x_t, x_0) \propto q(x_t \mid x_{t-1}) q(x_{t-1} \mid x_0).$$

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, \beta_t I),$$
$$q(x_{t-1} \mid x_0) = \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}} x_0, (1 - \bar{\alpha}_{t-1}) I).$$

Substitute these densities and collect the terms involving x_{t-1}:

$$\log q(x_{t-1} \mid x_t, x_0) = -\frac{1}{2} \left[\frac{\|x_t - \sqrt{\alpha_t} x_{t-1}\|^2}{\beta_t} + \frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0\|^2}{1 - \bar{\alpha}_{t-1}} \right] + C$$
$$= -\frac{1}{2} \left[a_t \|x_{t-1}\|^2 - 2 b_t^\top x_{t-1} \right] + C,$$

$$a_t = \frac{\alpha_t}{\beta_t} + \frac{1}{1 - \bar{\alpha}_{t-1}}, \qquad b_t = \frac{\sqrt{\alpha_t}}{\beta_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} x_0.$$

Here C collects terms independent of x_{t-1}.

Computing the Target Mean: Completing the Square

$$a_t = \frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}, \qquad b_t = \frac{\sqrt{\alpha_t}}{\beta_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}} x_0.$$

Complete the square in the posterior exponent:

$$a_t \|x_{t-1}\|^2 - 2 b_t^\top x_{t-1} = a_t \left\| x_{t-1} - \frac{b_t}{a_t} \right\|^2 - \frac{\|b_t\|^2}{a_t}.$$

The Gaussian mean and variance are therefore

$$\tilde{\mu}_t(x_t, x_0) = \frac{b_t}{a_t} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) x_t + \sqrt{\bar{\alpha}_{t-1}} \beta_t x_0}{1-\bar{\alpha}_t},$$

$$\tilde{\beta}_t = \frac{1}{a_t} = \frac{\beta_t (1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}. \qquad \text{For equal covariances, choose } \sigma_t^2 = \tilde{\beta}_t.$$

This is the target mean in terms of x_t and x_0. Next, express the same target using the forward noise.

The Target Mean: Final Form

With $\epsilon \sim \mathcal{N}(0, I)$, express the posterior mean using the forward noise:

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

Rearrange the direct-sampling formula and substitute into this mean:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\,\epsilon \quad \Longrightarrow \quad x_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\,\epsilon}{\sqrt{\bar{\alpha}_t}}.$$

$$\begin{aligned}
\tilde{\mu}_t &= \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\frac{x_t - \sqrt{1 - \bar{\alpha}_t}\,\epsilon}{\sqrt{\bar{\alpha}_t}} \\
&= \frac{\alpha_t(1 - \bar{\alpha}_{t-1}) + \beta_t}{\sqrt{\alpha_t}(1 - \bar{\alpha}_t)}x_t - \frac{\beta_t \epsilon}{\sqrt{\alpha_t}\sqrt{1 - \bar{\alpha}_t}} \\
\tilde{\mu}_t(x_t, x_0) &= \boxed{\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon\right)}.
\end{aligned}$$

We used $\bar{\alpha}_t = \alpha_t \bar{\alpha}_{t-1}$ and $\alpha_t + \beta_t = 1$.

Learning the Mean by Predicting Noise

Parameterize the learned mean by replacing ϵ with the network output $\epsilon_\theta(x_t, t)$:

$$\tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t \epsilon}{\sqrt{1-\bar{\alpha}_t}} \right),$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t \epsilon_\theta(x_t, t)}{\sqrt{1-\bar{\alpha}_t}} \right).$$

Substitute these means into the Gaussian KL loss:

$$D_{\mathrm{KL}}(q \| p_\theta) = \frac{\|\tilde{\mu}_t - \mu_\theta\|^2}{2\sigma_t^2}$$

$$= \underbrace{\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)}}_{w_t} \|\epsilon - \epsilon_\theta(x_t, t)\|^2.$$

The input is (x_t, t), and the target is the noise ϵ used to construct x_t.

The Simplified Training Objective

The intermediate KL terms give a weighted noise-prediction loss:

$$\sum_{t=2}^{T} \mathbb{E}_{x_0,\epsilon}\left[w_t \|\epsilon - \epsilon_\theta(x_t, t)\|^2\right], \qquad w_t = \frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)}.$$

The simplified objective uses uniform timesteps and unit weights, including $t = 1$:

$$L_{\text{simple}} = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right].$$

Why drop the weights? Ho et al. reported better sample quality with this simplified objective.

This changes the relative importance of the timesteps. It is a deliberate reweighting, not an algebraic equality. The ELBO section will account for the endpoint terms.

Generating the Network Input Directly

The network needs x_t. We can sample it directly from the clean training example x_0.

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I), \qquad \bar{\alpha}_t = \prod_{i=1}^{t} (1 - \beta_i).$$

Choose t, draw one Gaussian noise vector, and compute

$$\epsilon \sim \mathcal{N}(0, I), \qquad \boxed{x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon.}$$

There is no need to generate $x_1, \ldots, x_{t-1}$. The posterior mean $\tilde{\mu}_t$ is not used to generate this input.

Use this x_t as the network input and the sampled ϵ as its training target.

Training a DDPM

Repeat until convergence:

1. Sample a clean training observation, a timestep, and independent noise.

$$x_0 \sim q_{\text{data}}, \qquad t \sim \text{Uniform}\{1, \dots, T\}, \qquad \epsilon \sim \mathcal{N}(0, I).$$

2. Construct the noisy input in one step.

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon.$$

3. Predict the noise and take a gradient descent step.

$$\theta \leftarrow \theta - \eta \nabla_\theta \|\epsilon - \epsilon_\theta(x_t, t)\|^2.$$

The network input is (x_t, t), and the training target is ϵ.

Sampling from a DDPM

Start with Gaussian noise:

$$x_T \sim \mathcal{N}(0, I).$$

For $t = T, T-1, \ldots, 1$:

$$z \sim \mathcal{N}(0, I) \text{ if } t > 1, \qquad z = 0 \text{ if } t = 1.$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right),$$

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z.$$

Return x_0. At generation time, no clean observation is supplied.

The Generative Model and Its Likelihood

The reverse transitions define a joint distribution:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t), \qquad p(x_T) = \mathcal{N}(x_T; 0, I).$$

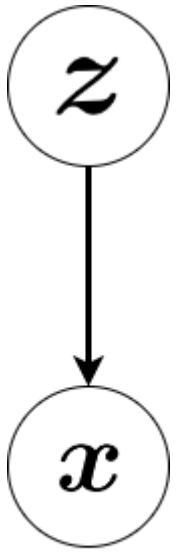
The likelihood of an observed training example integrates out the latent trajectory:

$$p_\theta(x_0) = \int p_\theta(x_{0:T})\, dx_1 \cdots dx_T.$$

We now derive a lower bound on the log likelihood and recover the reverse-step matching objective.

DDPM vs VAE

- **In VAE**
  - Observed variable x
  - Hidden Variable z



VAE

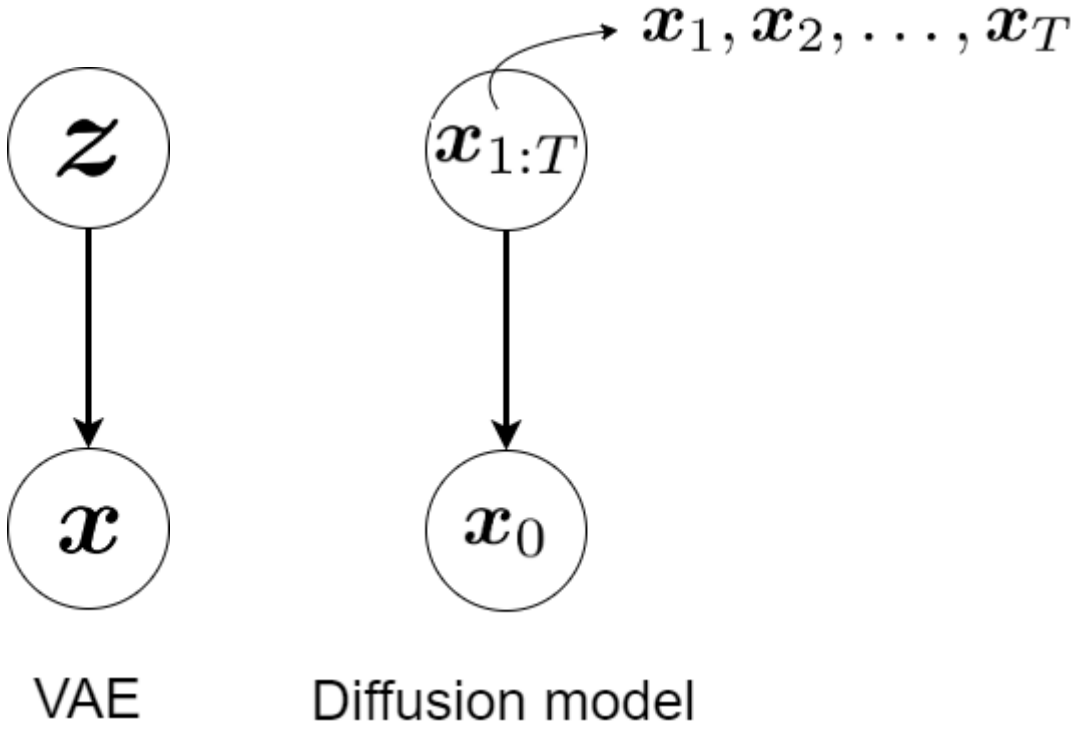
DDPM vs VAE

- In VAE
  - Observed variable x
  - Hidden Variable z



$$x_0 \longleftarrow x_1 \longleftarrow \cdots \longleftarrow x_T$$

- **In DDPM**
  - Observed variable x_0
  - Hidden variable $x_1, x_2, \ldots x_T$



Objective Functions in VAEs vs. DDPMs

VAEs:

Maximize the likelihood of the data:

$$p_\theta(x) = \int p_\theta(x, z)\, dz$$

Objective Functions in VAEs vs. DDPMs

VAEs:

Maximize the likelihood of the data:

$$p_\theta(x) = \int p_\theta(x, z)\, dz$$

Diffusion Models:

$$p_\theta(x_0) = \int p_\theta(x_0, x_{1:T})\, dx_{1:T}$$

Variational Lower Bounds in VAEs vs. DDPMs

VAEs:

$$\log p_\theta(x) \geq \mathbb{E}_{q(z|x)}[\log p_\theta(x \mid z)] - D_{\mathrm{KL}}(q(z \mid x) \,\|\, p_\theta(z)) \,.$$

The KL term in the lower bound compares the encoder with the latent prior.

Variational Lower Bounds in VAEs vs. DDPMs

VAEs:

$$\log p_\theta(x) \geq \mathbb{E}_{q(z|x)}[\log p_\theta(x \mid z)] - D_{\mathrm{KL}}(q(z \mid x) \,\|\, p_\theta(z)) \,.$$

Diffusion models:

$$\log p_\theta(x_0) \geq \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0 \mid x_{1:T})] - D_{\mathrm{KL}}(q(x_{1:T} \mid x_0) \,\|\, p_\theta(x_{1:T})) \,.$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} \right]$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) - \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} \right]$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) - \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) + \log \frac{p_\theta(x_{1:T})}{q(x_{1:T}|x_0)}\right]$$

Variational Lower Bound

$$\mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}))$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \int q(x_{1:T}|x_0) \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})} dx_{1:T}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0|x_{1:T})] - \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) - \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T})}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p_\theta(x_0|x_{1:T}) + \log \frac{p_\theta(x_{1:T})}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_0|x_{1:T})p_\theta(x_{1:T})}{q(x_{1:T}|x_0)}\right] = \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

$$\mathbb{E}_q \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

Throughout the derivation, the shorthand means

$$\mathbb{E}_q[\,\cdot\,] = \mathbb{E}_{q(x_{1:T}|x_0)}[\,\cdot\,], \qquad x_0 \text{ fixed.}$$

$$\mathbb{E}_q \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

$$= \mathbb{E}_q \left[\log \frac{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)} \right]$$

Substitute the reverse-model factorization:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t).$$

$$\mathbb{E}_q\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_q\left[\log \frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q[\log q(x_{1:T}|x_0)]$$

$$\mathbb{E}_q\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_q\left[\log \frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q[\log q(x_{1:T}|x_0)]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q\left[\log \prod_{t=1}^{T} q(x_t|x_{t-1})\right]$$

Substitute the forward-process factorization:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}).$$

$$\mathbb{E}_q\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_q\left[\log \frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q[\log q(x_{1:T}|x_0)]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q\left[\log \prod_{t=1}^{T} q(x_t|x_{t-1})\right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \sum_{t=1}^{T} \mathbb{E}_q[\log q(x_t|x_{t-1})]$$

$$\mathbb{E}_q\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_q\left[\log \frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}\right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q[\log q(x_{1:T}|x_0)]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \mathbb{E}_q\left[\log \prod_{t=1}^{T} q(x_t|x_{t-1})\right]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q[\log p_\theta(x_{t-1}|x_t)] - \sum_{t=1}^{T} \mathbb{E}_q[\log q(x_t|x_{t-1})]$$

$$= \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q\left[\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right]$$

Interpreting the Expanded ELBO

$$\mathcal{L}(x_0) = \mathbb{E}_q[\log p(x_T)] + \sum_{t=1}^{T} \mathbb{E}_q \left[\log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})} \right].$$

Here $\mathbb{E}_q = \mathbb{E}_{q(x_{1:T}|x_0)}$, with x_0 fixed.

The forward process and $p(x_T)$ are fixed. The learned parameters appear in the reverse transitions.

The displayed ratios are not yet KL divergences: numerator and denominator describe different conditional distributions.

Bayes' rule will introduce $q(x_{t-1} \mid x_t, x_0)$, the tractable target used earlier.

Rewriting the Forward Transition

$$\mathcal{L}(x_0) = \mathbb{E}_q \left[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})} \right] .$$

Here $\mathbb{E}_q = \mathbb{E}_{q(x_{1:T}|x_0)}$.

To obtain a KL divergence, express the denominator using the reverse posterior.

The Forward Markov Property

$$\mathcal{L}(x_0) = \mathbb{E}_q \left[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})} \right].$$

Given x_{t-1}, the forward transition does not depend on earlier states:

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

Adding x_0 leaves this forward density unchanged. It lets us apply Bayes' rule with x_0 held fixed.

Bayes' Rule with the Clean Observation

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

For $t \geq 2$, apply Bayes' rule with x_0 fixed:

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

Thus the transition ratio becomes

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

Keep the $t = 1$ reconstruction term separate.

Substituting into the ELBO

$$\mathcal{L}(x_0) = \mathbb{E}_q \left[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})} \right].$$

Substitute the following ratio for each intermediate term:

$$\frac{p_\theta(x_{t-1} \mid x_t)}{q(x_t \mid x_{t-1})} = \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_{t-1} \mid x_t, x_0)} \frac{q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)}, \quad t \geq 2.$$

$$\mathcal{L}(x_0) = \mathbb{E}_q \Bigg[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_{t-1} \mid x_t, x_0)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)} + \log \frac{p_\theta(x_0 \mid x_1)}{q(x_1 \mid x_0)} \Bigg].$$

The Telescoping Sum

$$\sum_{t=2}^{T} \log \frac{q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)} = \sum_{t=2}^{T} \left[\log q(x_{t-1} \mid x_0) - \log q(x_t \mid x_0) \right]$$
$$= \log q(x_1 \mid x_0) - \log q(x_T \mid x_0).$$

Substitute this result into the expanded ELBO:

$$\mathcal{L}(x_0) = \mathbb{E}_q \Bigg[\log p(x_T) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_{t-1} \mid x_t, x_0)}$$
$$+ \sum_{t=2}^{T} \log \frac{q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)} + \log \frac{p_\theta(x_0 \mid x_1)}{q(x_1 \mid x_0)} \Bigg].$$

The $\log q(x_1 \mid x_0)$ terms cancel. The remaining terminal ratio uses $q(x_T \mid x_0)$.

The Negative ELBO

After the cancellation, negate the lower bound:

$$-\mathcal{L}(x_0) = -\mathbb{E}_q\left[\log \frac{p(x_T)}{q(x_T \mid x_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1} \mid x_t)}{q(x_{t-1} \mid x_t, x_0)} + \log p_\theta(x_0 \mid x_1)\right].$$

Integrate out unused states, then recognize each expected log ratio as a KL:

$$-\mathcal{L}(x_0) = \underbrace{D_{\mathrm{KL}}(q(x_T \mid x_0) \,\|\, p(x_T))}_{L_T(x_0)} + \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(x_t|x_0)} D_{\mathrm{KL}}(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t))}_{L_{t-1}(x_0)} \underbrace{-\mathbb{E}_{q(x_1|x_0)} \log p_\theta(x_0 \mid x_1)}_{L_0(x_0)}.$$

The intermediate KL terms give our earlier training losses. The terminal term is fixed, and L_0 measures reconstruction.

The Three Terms in the Negative ELBO

$$-\mathcal{L}(x_0) = L_T(x_0) + \sum_{t=2}^{T} L_{t-1}(x_0) + L_0(x_0).$$

First term: fixed forward process and fixed prior, so no dependence on θ.

$$L_T(x_0) = D_{\mathrm{KL}}(q(x_T \mid x_0) \,\|\, p(x_T)) \,.$$

Second term: the intermediate denoising losses we optimized earlier.

$$L_{t-1}(x_0) = \mathbb{E}_{q(x_t|x_0)} D_{\mathrm{KL}}(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)) \,.$$

Third term: learn the final step from x_1 to the clean observation x_0.

$$L_0(x_0) = -\mathbb{E}_{q(x_1|x_0)} \log p_\theta(x_0 \mid x_1).$$

Both the second and third terms train the model. The third term is reconstruction.

The Final Step Is Also Mean Matching

For continuous data, choose a Gaussian decoder with fixed $\sigma_1^2 > 0$:

$$p_\theta(x_0 \mid x_1) = \mathcal{N}\big(x_0; \mu_\theta(x_1, 1), \sigma_1^2 I\big) .$$

Its negative log density is

$$-\log p_\theta(x_0 \mid x_1) = \frac{1}{2\sigma_1^2}\|x_0 - \mu_\theta(x_1, 1)\|^2 + c, \qquad c = \frac{d}{2}\log(2\pi\sigma_1^2).$$

Substitute this expression into the reconstruction term:

$$\begin{aligned} L_0(x_0) &= -\mathbb{E}_{q(x_1|x_0)} \log p_\theta(x_0 \mid x_1) \\ &= \frac{1}{2\sigma_1^2}\mathbb{E}_{q(x_1|x_0)}\|x_0 - \mu_\theta(x_1, 1)\|^2 + c. \end{aligned}$$

The target is now the observed x_0 itself. We train the same network at $t = 1$.

Reconstruction Becomes Noise Prediction

At $t = 1$, use $\bar{\alpha}_1 = \alpha_1$, $1 - \alpha_1 = \beta_1$, and $\epsilon \sim \mathcal{N}(0, I)$:

$$x_1 = \sqrt{\alpha_1} x_0 + \sqrt{\beta_1}\epsilon \quad \Longrightarrow \quad x_0 = \frac{x_1 - \sqrt{\beta_1}\epsilon}{\sqrt{\alpha_1}}.$$

The learned mean uses the same form, replacing noise by its prediction:

$$\mu_\theta(x_1, 1) = \frac{x_1 - \sqrt{\beta_1}\epsilon_\theta(x_1, 1)}{\sqrt{\alpha_1}}, \qquad x_0 - \mu_\theta(x_1, 1) = \sqrt{\frac{\beta_1}{\alpha_1}}\big(\epsilon_\theta(x_1, 1) - \epsilon\big).$$

Substitute this difference into the reconstruction loss from the previous slide:

$$\begin{aligned} L_0(x_0) &= \frac{1}{2\sigma_1^2}\mathbb{E}\|x_0 - \mu_\theta(x_1, 1)\|^2 + c \\ &= \boxed{\frac{\beta_1}{2\sigma_1^2\alpha_1}\mathbb{E}_\epsilon\|\epsilon - \epsilon_\theta(x_1, 1)\|^2 + c}. \end{aligned}$$

Sampling $t = 1, \ldots, T$ in the simplified noise loss includes this final step. Dropping its weight, like the other weights, gives a reweighted objective.