Maths Behind Denoising Diffusion Probabilistic Models (DDPM's) II

Ayush Raina

Indian Institute of Science

October 6, 2024

ELBO - Evidance Lower Bound

1. Evidence - Log Likelihood of observed data - $log(p(x))$ where $x \in \mathbb{R}^d$
2. Evidance Lower Bound (ELBO) - Lower bound on the log likelihood of observed data - $log(p(x)) \geq ELBO$

$$\boxed{ELBO = \mathbb{E}_{q(z|x)}[log(\frac{p(x,z)}{q(z|x)})]}$$

So let us see why ELBO is a lower bound on the log likelihood of observed data.

Proof 1

We start with $\log(p(x))$, where z is the latent variable:

$$\begin{aligned}
\log(p(x)) &= \log\left(\sum_z p(x,z)\right) && \text{(starting point)} \\
&= \log\left(\sum_z \frac{p(x,z)}{q(z \mid x)} q(z \mid x)\right) && \text{(multiply and divide by } q(z \mid x)\text{)} \\
&= \log\left(\mathbb{E}_{q(z|x)}\left[\frac{p(x,z)}{q(z \mid x)}\right]\right) \\
&\geq \mathbb{E}_{q(z|x)}\left[\log\left(\frac{p(x,z)}{q(z \mid x)}\right)\right] && \text{(Jensen's inequality)}
\end{aligned}$$

We can use Jensen inequality because log is a concave function.

Proof 2

$$\log(p(x)) = \log(p(x)) \left(\sum_z q(z|x) \right) \qquad \text{(starting point)}$$

$$= \sum_z q(z|x) \log(p(x))$$

$$= \mathbb{E}_{q(z|x)} \log\left(p(x)\right)$$

$$= \mathbb{E}_{q(z|x)} \log \left(\frac{p(x)p(z|x)}{p(z|x)} \right)$$

$$= \mathbb{E}_{q(z|x)} \left[\log \left(\frac{p(x,z)}{p(z|x)} \right) \right]$$

Proof 2 Continued

$$
\begin{aligned}
&= \mathbb{E}_{q(z|x)}\left[\log\left(\frac{p(x,z)}{q(z|x)}\frac{q(z|x)}{p(z|x)}\right)\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log\left(\frac{p(x,z)}{q(z|x)}\right) + \log\left(\frac{q(z|x)}{p(z|x)}\right)\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log\left(\frac{p(x,z)}{q(z|x)}\right)\right] + \mathbb{E}_{q(z|x)}\left[\log\left(\frac{q(z|x)}{p(z|x)}\right)\right] \\
&= ELBO + KL(q(z|x)||p(z|x))
\end{aligned}
$$

Since $KL(q(z|x)||p(z|x)) \geq 0$, we have $\log(p(x)) \geq ELBO$.

Now we know that

$$\boxed{\log(p(x)) = \mathbb{E}_{q(z|x)}\left[\log\left(\frac{p(x,z)}{q(z|x)}\right)\right] + KL(q(z|x)||p(z|x))}$$

and due to the KL divergence term on the right hand side, we can say that ELBO is a lower bound on the log likelihood of observed data.

Variational Autoencoder (VAE)

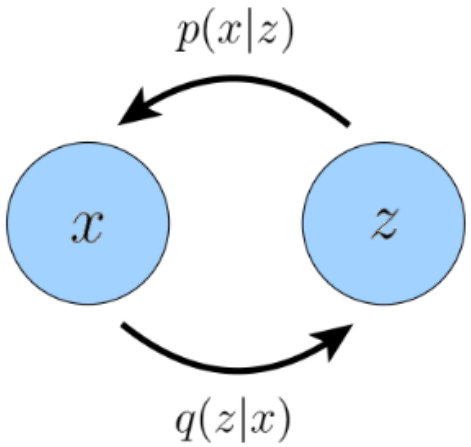

Figure 1: Variational Autoencoder

In default formulation in VAE paper, directly maximize ELBO using a variational approach.
We optimize for best possible $q_\phi(z|x)$ among family of posterior distributions which are parameterized by ϕ.

Family of Posterior Distributions

The family of posterior distributions is generally choosen as multivariate gaussian with diagonal covariance matrix.

$$q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), diag(\sigma_\phi(x)))$$

The prior is choosen as standard normal distribution

$$p(z) = \mathcal{N}(z; 0, \mathbb{I}_{d\times d})$$

Encoder and Decoder Network

When we maximize the ELBO, we are doing the following things to be specific:

1. Adjusting parameters ϕ in such a way that true latent distribution $p(z)$ is as close as possible to encoder outputs $p(z|x)$.
2. Using these latents to regenerate the true data x as close as possible.

Let us see what they mean

ELBO again

$$
\begin{aligned}
ELBO &= \mathbb{E}_{q(z|x)}[log\left(\frac{p(x,z)}{q(z|x)}\right)] \\
&= \mathbb{E}_{q(z|x)}[log\left(\frac{p(z)p(x|z)}{q(z|x)}\right)] \\
&= \mathbb{E}_{q(z|x)}[log(p(x|z)) + log\left(\frac{p(z)}{q(z|x)}\right)] \\
&= \mathbb{E}_{q(z|x)}[log(p(x|z))] + \mathbb{E}_{q(z|x)}[log\left(\frac{p(z)}{q(z|x)}\right)] \\
&= \mathbb{E}_{q(z|x)}[log(p(x|z))] - \mathbb{E}_{q(z|x)}[log\left(\frac{q(z|x)}{p(z)}\right)] \\
&= \mathbb{E}_{q(z|x)}[log(p(x|z))] - KL(q(z|x)||p(z))
\end{aligned}
$$

Continued

While maximizing ELBO, we did 2 things:

1. Maximize $\mathbb{E}_{q(z|x)}[log(p(x|z))]$ - This is the reconstruction term.
2. Minimize $KL(q(z|x)||p(z))$ - This is the prior matching term.

Look at 2nd term first, this tries to bring $q(z|x)$ close to $p(z)$, which means given x, we are trying to model z as close as possible to true latent distribution.

Now look at 1st term in which we maximize the log likelihood of generating back the true x from z.

Computing ELBO

There are 2 terms to be computed in ELBO:

1. Reconstruction term: $\mathbb{E}_{q(z|x)}[log(p(x|z))]$

This can be estimated using sample averages. We sample $\{z^{(i)}\}_{i=1}^{N} \sim q(z|x)$ then above term can be computed as

$$\mathbb{E}_{q(z|x)}[log(p(x|z))] = \frac{1}{N}\sum_{i=1}^{N} log(p(x|z^{(i)}))$$

Computing ELBO

2. KL Divergence term: $KL(q(z|x)||p(z))$

Since both $q(z|x)$ and $p(z)$ are gaussian, there exists a closed form solution to compute KL divergence between 2 gaussians.

Markovian Hierarchical Variational Autoencoder (MHVAE)

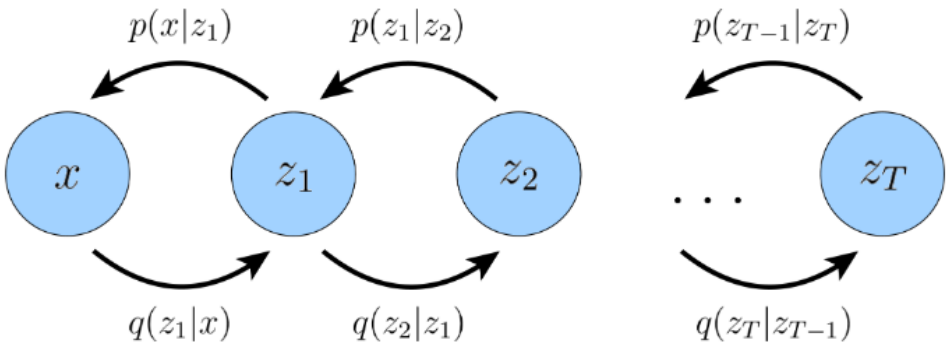

Figure 2: A Markovian Hierarchical Variational Autoencoder with T hierarchical latents. The generative process is modeled as a Markov chain, where each latent $\boldsymbol{z}_t$ is generated only from the previous latent $\boldsymbol{z}_{t+1}$.

Figure 2: Markovian Hierarchical Variational Autoencoder

In MHVAE, we have a hierarchical structure of latent variables.

Markovian Hierarchical Variational Autoencoder (MHVAE)

Now we have a hierarchical structure of latent variables. $z_1, z_2, \ldots, z_T$ are the latent variables at each time step. We can denote all these using $z_{1:T}$.

Joint and Posterior

The Joint Distribution can be written as:

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

The posterior distribution can be written as:

$$\begin{aligned} p(z_{1:T}|x) &= \frac{p(x, z_{1:T})}{p(x)} \\ &= \frac{p(x)p(z_1|x)\prod_{t=2}^{T} p(z_t|z_{t-1})}{p(x)} \\ &= p(z_1|x)\prod_{t=2}^{T} p(z_t|z_{t-1}) \end{aligned}$$

Extending ELBO

$$
\begin{aligned}
\log(p(x)) &= \log\left(\int p(x, z_{1:T}) dz_{1:T}\right) \\
&= \log\left(\int p(x, z_{1:T}) \frac{q(z_{1:T}|x)}{q(z_{1:T}|x)} dz_{1:T}\right) \\
&= \log\left(\mathbb{E}_{q(z_{1:T}|x)}\left[\frac{p(x, z_{1:T})}{q(z_{1:T}|x)}\right]\right) \\
&\geq \mathbb{E}_{q(z_{1:T}|x)}\left[\log\left(\frac{p(x, z_{1:T})}{q(z_{1:T}|x)}\right)\right] \qquad \text{(Jensen's Inequality)}
\end{aligned}
$$

Extending ELBO

$$\begin{aligned}
\log(p(x)) &\geq \mathbb{E}_{q(z_{1:T}|x)}\left[\log\left(\frac{p(x, z_{1:T})}{q(z_{1:T}|x)}\right)\right] \\
&= \mathbb{E}_{q(z_{1:T}|x)}\left[\log\left(\frac{p(z_T)p(x|z_1)\prod_{t=2}^{T} p(z_{t-1}|z_t)}{q(z_{1:T}|x)}\right)\right] \\
&= \mathbb{E}_{q(z_{1:T}|x)}\left[\log\left(\frac{p(z_T)p(x|z_1)\prod_{t=2}^{T} p(z_{t-1}|z_t)}{q(z_1|x)\prod_{t=2}^{T} q(z_t|z_{t-1})}\right)\right]
\end{aligned}$$

We will see how this expression can be broken down into terms that we have seen in VAE.

Variational Diffusion Models (VDM's)

Variational Diffusion Model (VDM) is simply a MHVAE with 3 restrictions:

1. The dimension of latent variables is equal to dimension of data.
2. Distribution of latent variables at each time step is Gaussian Centered at latent variable at previous time step. (This is not learnt, but fixed)
3. Encoding transitions are done in such a way that distribution of latent variable at last time step is standard Gaussian.

Let us see what changes occur in MHVAE when these restrictions are applied to it.

Variational Diffuion Models (VDM's)

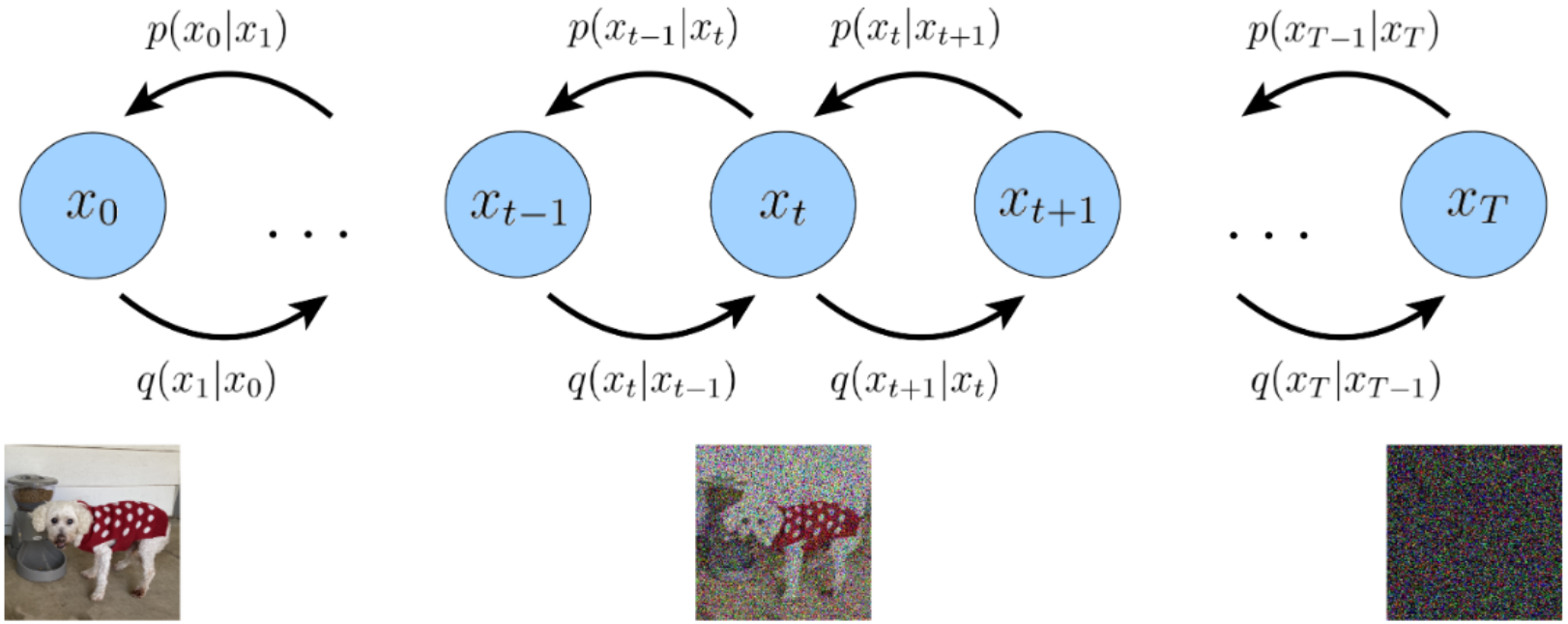

Figure 3: Variational Diffusion Model

Variational Diffusion Models (VDM's)

- x_0 represents true data observations.
- x_t represents intermediate noisy version of data.
- x_T represents pure Gaussian noise.

Restriction 1

According to slight change of notation in previous slide, VDM's posterior is same as MHVAE's posterior which can now be written as:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \qquad \text{(VDM's Posterior)}$$

instead of

$$q(z_{1:T}|x) = p(z_1|x) \prod_{t=2}^{T} q(z_t|z_{t-1}) \qquad \text{(MHVAE's Posterior)}$$

Restriction 2

Unlike MHVAE, distribution of latent variable at each time step is not learnt, but is fixed. Mathematically, the encoder transitions can be written as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbb{I}_{d\times d})$$

Distribution of latent variable x_t at each time step is gaussian centered at latent variable at previous time step x_{t-1}.

Restriction 3

Gaussian Transitions are made in such a way that distribution of final latent variable x_T is standard Gaussian. Mathematically, this can be written as:

$$q(x_T|x_{T-1}) = \mathcal{N}(x_T; 0, \mathbb{I}_{d\times d})$$

Hence Joint Distribution of VDM can be written as:

$$p(x_{0:T}) = p(x_T)\prod_{t=1}^{T} q(x_{t-1}|x_t) \qquad p(x_T) \sim \mathcal{N}(0, \mathbb{I}_{d\times d})$$

Instead of

$$p(x, z_{1:T}) = p(x_T)p(x|z_1)\prod_{t=2}^{T} q(z_{t-1}|z_t) \quad \text{(MHVAE's Joint)}$$

Learning conditionals

Since encoder transitions are fixed for each time steps, we are only interested in learning conditionals $p(x_{t-1}|x_t)$ so that we can simulate new data.

We can achieve this by maximizing ELBO for VDM. To show that maximizing ELBO actually helps in learning conditionals, we will first split ELBO in different components like we did in VAE and got the reconstruction term and prior matching term.

Breaking down ELBO

We know that ELBO for MHVAE can be written as:

$$ELBO = \mathbb{E}_{q(z_{1:T}|x)}\left[\log\left(\frac{p(x, z_{1:T})}{q(z_{1:T}|x)}\right)\right]$$

According to notation's for VDM, we can write ELBO as:

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

Breaking down ELBO

We further saw that ELBO for MHVAE can be written as:

$$ELBO = \quad = \mathbb{E}_{q(z_{1:T}|x)}\left[\log\left(\frac{p(z_T)p(x|z_1)\prod_{t=2}^{T}p(z_{t-1}|z_t)}{q(z_1|x)\prod_{t=2}^{T}q(z_t|z_{t-1})}\right)\right]$$

In our case we can write this as:

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

Let us now further simplify this expression.

Breaking down ELBO

$$
\begin{aligned}
ELBO &= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\left(\frac{p(x_T)\prod_{t=1}^{T} p(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}\right)\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\left(\frac{p(x_T)p(x_0|x_1)\prod_{t=2}^{T} p(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}\right)\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\left(\frac{p(x_T)p(x_0|x_1)\prod_{t=2}^{T} p(x_{t-1}|x_t)}{q(x_T|x_{T-1})\prod_{t=1}^{T-1} q(x_t|x_{t-1})}\right)\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\left(\frac{p(x_T)p(x_0|x_1)\prod_{t=1}^{T-1} p(x_t|x_{t+1})}{q(x_T|x_{T-1})\prod_{t=1}^{T-1} q(x_t|x_{t-1})}\right)\right]
\end{aligned}
$$

Breaking down ELBO

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(\frac{p(x_T)p(x_0|x_1)}{q(x_T|x_{T-1})} \right) + \sum_{t=1}^{T-1} \log \left(\frac{p(x_t|x_{t+1})}{q(x_t|x_{t-1})} \right) \right]$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(\frac{p(x_T)p(x_0|x_1)}{q(x_T|x_{T-1})} \right) \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(\frac{p(x_t|x_{t+1})}{q(x_t|x_{t-1})} \right) \right.$$

First Term can further be simplified into

$$= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(\frac{p(x_T)}{q(x_T|x_{T-1})} \right) \right] + \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(p(x_0|x_1) \right) \right]$$

Reconstruction Term

We got 3 terms:

1. $\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\left(p(x_0|x_1)\right)\right]$

Since inside expression is independent of $x_{2:T}$, we can write this as:

$$= \mathbb{E}_{q(x_1|x_0)}\left[\log\left(p(x_0|x_1)\right)\right] \qquad \text{(reconstruction term)}$$

maximizing ELBO results in maximizing this term which is log likelihood of getting back original data sample from latent variable at first time step.

Prior Matching Term

2. $\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\left(\frac{p(x_T)}{q(x_T|x_{T-1})}\right)\right]$

We can simplify this expression as follows:

$$
\begin{aligned}
&= \mathbb{E}_{q(x_{T-1,T}|x_0)}\left[\log\left(\frac{p(x_T)}{q(x_T|x_{T-1})}\right)\right] \\
&= \int q(x_{T-1,T}|x_0)\log\left(\frac{p(x_T)}{q(x_T|x_{T-1})}\right)dx_{T-1}dx_T \\
&= \int q(x_{T-1}|x_0)\left[q(x_T|x_{T-1})\log\left(\frac{p(x_T)}{q(x_T|x_{T-1})}\right)dx_T\right]dx_{T-1} \\
&= -\int q(x_{T-1}|x_0)D_{KL}\left(q(x_T|x_{T-1})||p(x_T)\right)dx_{T-1} \\
&= -\mathbb{E}_{q(x_{T-1}|x_0)}\left[D_{KL}\left(q(x_T|x_{T-1})||p(x_T)\right)\right]
\end{aligned}
$$

Prior Matching Term

$$= -\mathbb{E}_{q(x_{T-1}|x_0)}\left[D_{KL}\left(q(x_T|x_{T-1})||p(x_T)\right)\right]$$

This is prior matching term, maximizing ELBO results in minimizing this term, which forces the distribution of latent variable at last time as close as possible to $p(x_T) \sim \mathcal{N}(0, \mathbb{I}_{d\times d})$.

Consistency Term

3. $\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\left(\frac{p(x_t|x_{t+1})}{q(x_t|x_{t-1})}\right)\right]$

This term can be simplified as:

$$= \mathbb{E}_{q(x_{t-1,t,t+1}|x_0)}\left[\log\left(\frac{p(x_t|x_{t+1})}{q(x_t|x_{t-1})}\right)\right]$$

$$= \int q(x_{t-1,t,t+1}|x_0)\log\left(\frac{p(x_t|x_{t+1})}{q(x_t|x_{t-1})}\right)dx_{t-1}dx_t dx_{t+1}$$

$$= \int q(x_{t+1}|x_0)q(x_{t-1}|x_{t+1})\left[q(x_t|x_{t-1})\log\left(\frac{p(x_t|x_{t+1})}{q(x_t|x_{t-1})}\right)dx_t\right]dx_{t-1,t+1}$$

$$= -\int q(x_{t+1}|x_0)q(x_{t-1}|x_{t+1})D_{KL}\left(q(x_t|x_{t-1})||p(x_t|x_{t+1})\right)dx_{t-1,t+1}$$

$$= -\mathbb{E}_{q(x_{t+1}|x_0)q(x_{t-1}|x_{t+1})}\left[D_{KL}\left(q(x_t|x_{t-1})||p(x_t|x_{t+1})\right)\right]$$

Consistency Term

$$= -\mathbb{E}_{q(x_{t-1,t+1}|x_0)}\left[D_{KL}\left(q(x_t|x_{t-1})||p(x_t|x_{t+1})\right)\right]$$

This is consistency term, maximizing ELBO results in minimizing this term, which ensures that distribution of noised image at time step t i.e $q(x_t|x_{t-1})$ is as close as possible to distribution of denoised image at time step t i.e $p(x_t|x_{t+1})$.

ELBO

Finally we saw that ELBO = reconstruction term + prior matching term + consistency term where:

1. Reconstruction term $\mathbb{E}_{q(x_1|x_0)}\left[\log\left(p(x_0|x_1)\right)\right]$
2. Prior Matching term $-\mathbb{E}_{q(x_{T-1}|x_0)}\left[D_{KL}\left(q(x_T|x_{T-1})||p(x_T)\right)\right]$
3. Consistency term $-\mathbb{E}_{q(x_{t-1,t+1}|x_0)}\left[D_{KL}\left(q(x_t|x_{t-1})||p(x_t|x_{t+1})\right)\right]$

maximizing ELBO results in minimizing prior matching term and consistency term and maximizing reconstruction term.

Thank You!

Thank You!

References

1. Understanding Diffusion Models: A Unified Perspective