(Extra): Mathematical explanation:

Forward Diffusion Process:

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

$$a_t = 1 - \beta_t$$

$$\overline{a_t} = \prod_{s=1}^{t} a_s$$

$\boldsymbol{Reparameterization\,Trick}$:

$$N(\mu, \sigma^2) = \mu + \sigma \cdot \epsilon$$

$$\downarrow$$

$$(x_t|x_{t-1}) = N\left(x_t, \sqrt{1-\beta}x_{t-1}, \beta_t I\right) = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon = \sqrt{\mathrm{a_t}}\mathrm{x_{t-1}} + \sqrt{1-\mathrm{a_t}}\epsilon$$

$$= \sqrt{a_t \cdot a_{t-1}}x_{t-2} + \sqrt{1 - a_t \cdot a_{t-1}}\epsilon = \sqrt{\mathrm{a_t \cdot a_{t-1} \cdot a_{t-2}}}\mathrm{x_{t-3}} + \sqrt{1 - \mathrm{a_t \cdot a_{t-1} \cdot a_{t-2}}}\epsilon$$

$$= \sqrt{a_t \cdot a_{t-1} \cdot \ldots \cdot a_1 \cdot a_0}x_0 + \sqrt{1 - a_t \cdot a_{t-1} \cdot \ldots a_t \cdot a_0}\epsilon$$

$$\sqrt{\overline{a}_t}x_0 + \sqrt{1 - \overline{a}_t}\epsilon$$

$$\downarrow$$

$$q(x_t|x_0) = N\left(x_t, \sqrt{\overline{a}_t}x_0, (1 - \overline{a}_t)I\right)$$

Reverse Diffusion Process:

$$p(x_{t-1}|x_t) = N\left(x_{t-1}, \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right)$$

Evidence Lower BOund (ELBO):

$$-\log\left(p_\theta(x_0)\right) \le -\log\left(p_\theta(x_0)\right) + D_{KL}\left(q(\,x_{1:T} \mid x_0\,) \parallel p_\theta(\,x_{1:T} \mid x_0\,)\right)$$

$$\downarrow$$

$$-\log\left(p_\theta(x_0)\right) \le -\log\left(p_\theta(x_0)\right) + \log\left(\frac{q(\,x_{1:T} \mid x_0\,)}{p_\theta(\,x_{1:T} \mid x_0\,)}\right)$$

Bayes rule on $p_\theta(x_{1:T}|x_0)$:

$$p_\theta(x_{1:T}|x_0) = \frac{p_\theta(x_0|x_{1:T})p_\theta(x_{1:T})}{p_\theta(x_0)} = \frac{p_\theta(x_{0:T})}{p_\theta(x_0)}$$

Hence:

$$-\log\left(p_\theta(x_0)\right) \le -\log\left(p_\theta(x_0)\right) + \log\left(\frac{q(x_{1:T}|x_0)}{\frac{p_\theta(x_{0:T})}{p_\theta(x_0)}}\right)$$

$$= -\log\left(p_\theta(x_0)\right) \le -\log\left(p_\theta(x_0)\right) + \log\left(\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right) + \log\left(p_\theta(x_0)\right)$$

$$= -\log\left(p_\theta(x_0)\right) \le \log\left(\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right)$$

$$\downarrow$$

$$\log\left(\frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right) = \log\left(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}\right) = -\log(p(x_T)) + \log\left(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}\right)$$

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

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

Apply Bayes' rule on $q(x_1|x_{t-1})$:

$$q(x_1|x_{t-1}) = \frac{q(x_{t-1}|x_t)\cdot q(x_t)}{q(x_{t-1})}$$

In order to know where we came from, we write:

$$\frac{q(x_{t-1}|x_t, x_0)\cdot q(x_t|x_0)}{q(x_{t-1}|x_0)}$$

Hence:

$$-\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t, x_0)\cdot q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)\cdot q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \overbrace{\sum_{t=2}^{T} \log\left(\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}\right)}^{=\log\left(\frac{q(x_T|x_0)}{q(x_1|x_0)}\right)} + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{q(x_1|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log(q(x_T|x_0)) - \log(q(x_1|x_0)) + \log(q(x_1|x_0)) - \log(p_\theta(x_0|x_1))$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log(q(x_T|x_0)) - \log(p_\theta(x_0|x_1))$$

$$= \log\left(\frac{q(x_T|x_0)}{p(x_T)}\right) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)}\right) - \log(p_\theta(x_0|x_1))$$

$$= D_{KL}(q(x_T|x_0) \parallel p(x_T)) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$\boldsymbol{D_{KL}(q(x_T|x_0) \parallel p(x_T))}$ can be ignored, no learnable parameters:

$$\sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$= \sum_{t=2}^{T} |\epsilon - \epsilon_\theta(x_t, t)|^2 - \log(p_\theta(x_0|x_1))$$

After some simplifications made by the authors, we can write:

$$L_{simple} = E_{t,x_0,\epsilon}[|\epsilon - \epsilon_\theta(x_t, t)|^2]$$