Understanding the intuition and math behind Stable Diffusion

Annada Prasad Behera

National Institute of Science Education and Research, Bhubaneswar

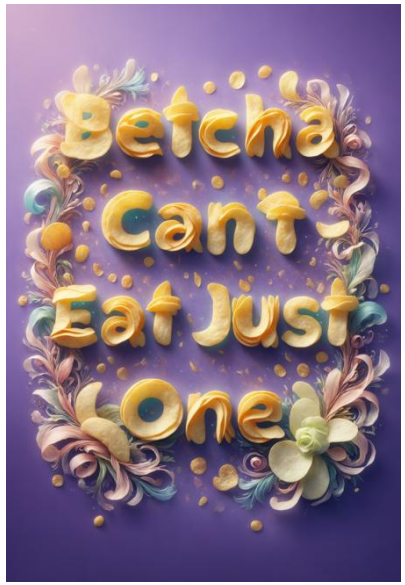
July 29th, 2024

Examples of images generated using SD

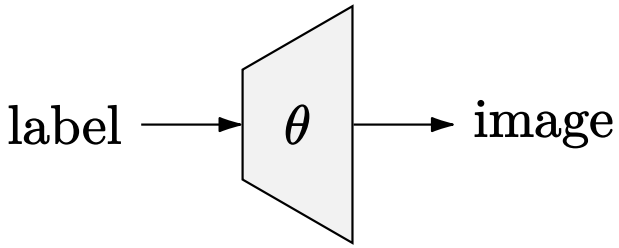




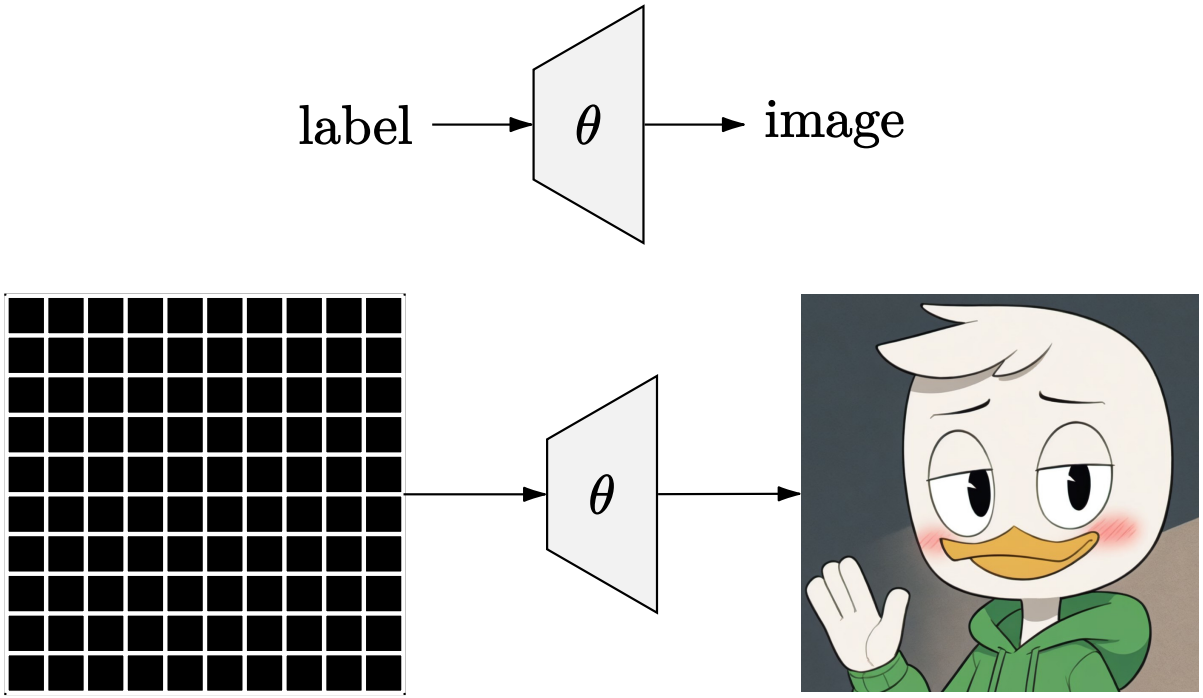


A naive idea

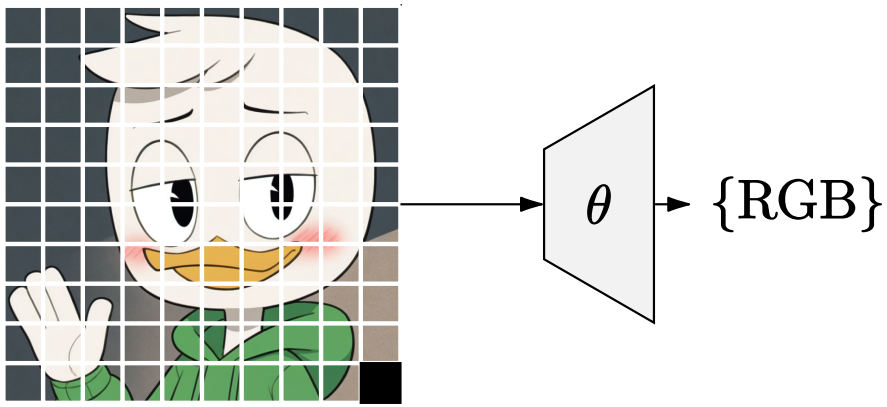


A naive idea

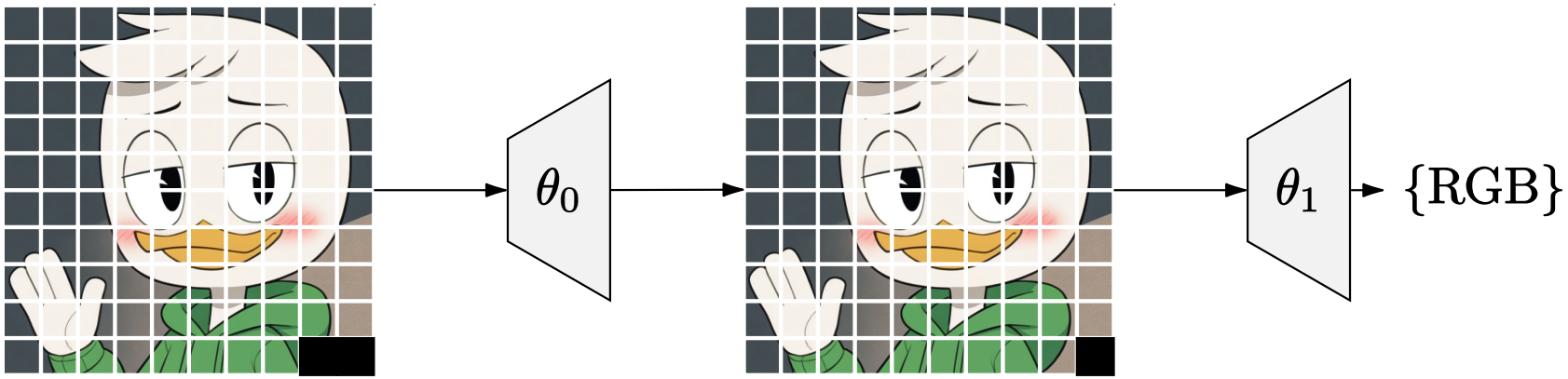


Observation I: For multiple predictions for the same input, the predictors will learn to output the average of the labels.

Predicting a pixel – Autoregressor



- Observation II: There is no blurring effect. Since, the average value of many colors is still a color.
- Observation II: Why not chain the two neural networks?



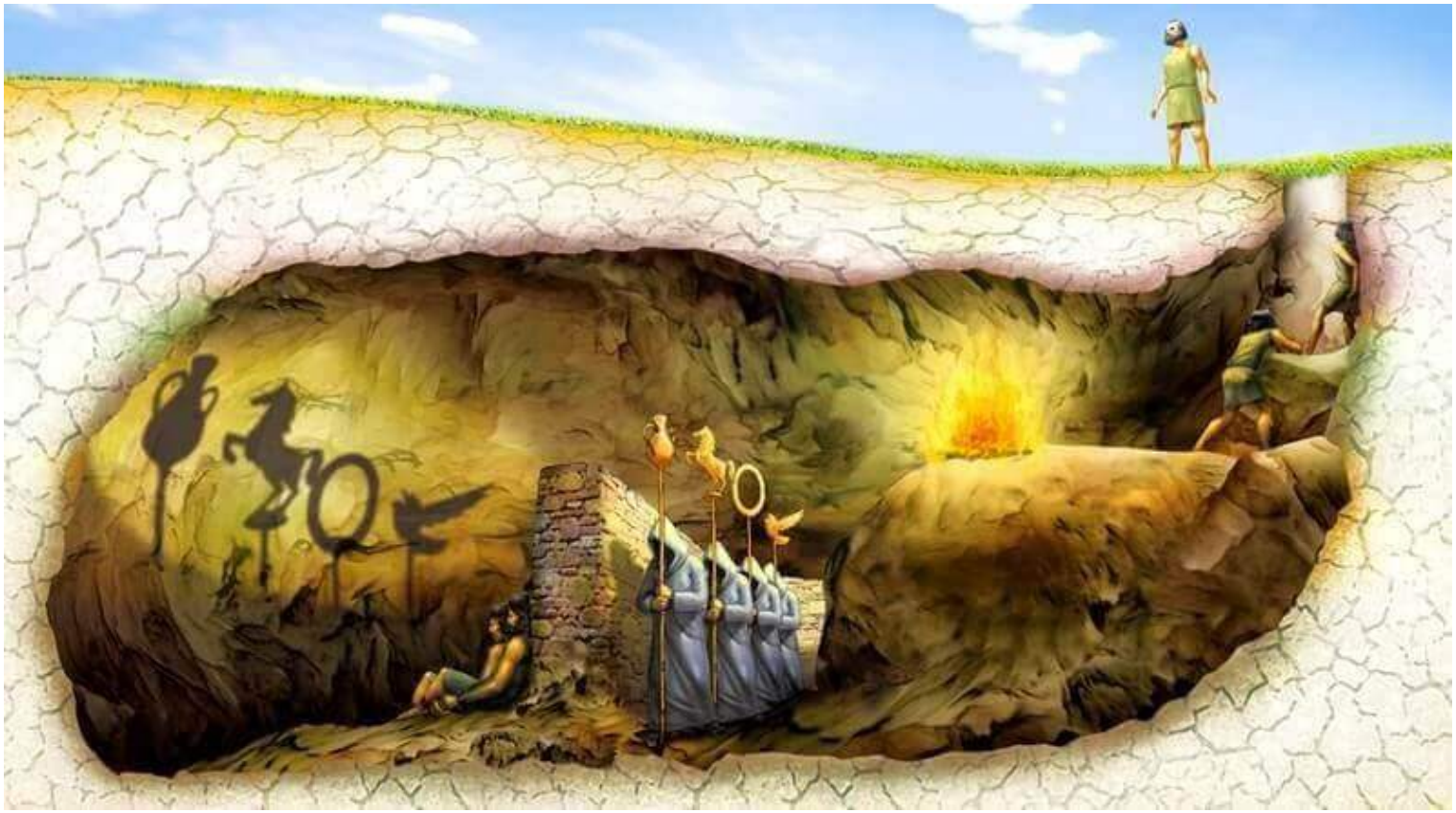
Classical Cartography

Pomponius Mela, circa 43 AD.

Classical Cartography (with better clocks)

Marinus of Tyre, circa 114 AD

Latent variable: Plato's Allegory of the Cave



Probability theory

For given data x and it's latent variable z, the joint probability distribution $p(x, z)$ is given as,

$$p(x, z) = p(x) \cdot p(z|x) \tag{1}$$

The marginalized latent z provides the full probability of seeing the data,

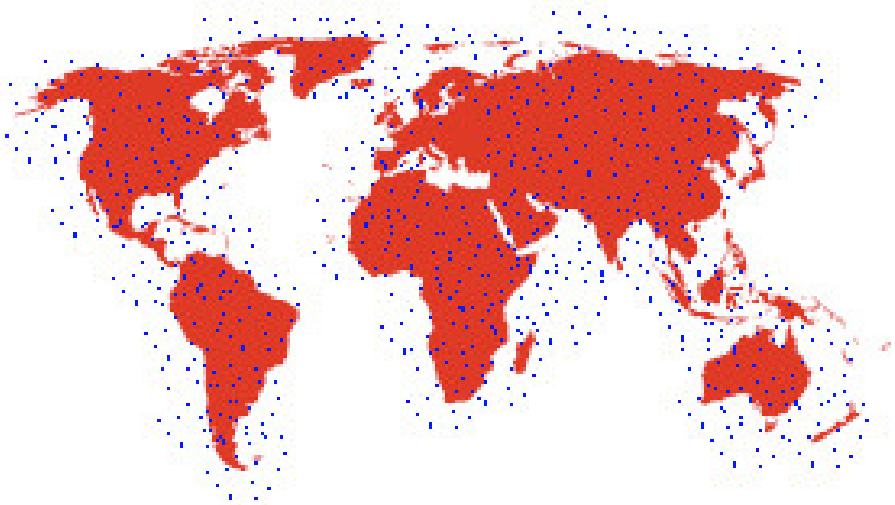
$$p(x) = \int p(x, z)\ dz \tag{2}$$

And from Bayes' rule,

$$p(z|x) = \frac{p(x|z) \cdot p(z)}{p(x)} \tag{3}$$

where, $p(z|x)$ is the *posterior*, $p(x|z)$ is the *likelihood*, $p(z)$ is the *prior* and $p(x)$ is the *evidence.*

Probability theory



x	the data we have collected
$p(x)$	is the true distribution we want to estimate
$q_\phi(x)$	the our current best idea of $p(x)$
$p(x, z)$	is the true data distribution
$q_\phi(z\|x)$	our current best idea of latent with the data at hand

ELBO: Evidence lower bound

The log of the true distribution is given by,

$$\log p(x) = \log p(x) \int q_\phi(z|x)\ dz \tag{4}$$

$$= \int q_\phi(z|x) \log p(x)\ dz = \mathbb{E}_{q_\phi(z|x)}[\log p(x)] \tag{5}$$

$$= \mathbb{E}_{q_\phi(z|x)}\Big[\log \frac{p(x,z)}{p(z|x)}\Big] = \mathbb{E}_{q_\phi(z|x)}\Big[\log \frac{p(x,z)\cdot q_\phi(z|x)}{p(z|x)\cdot q_\phi(z|x)}\Big] \tag{6}$$

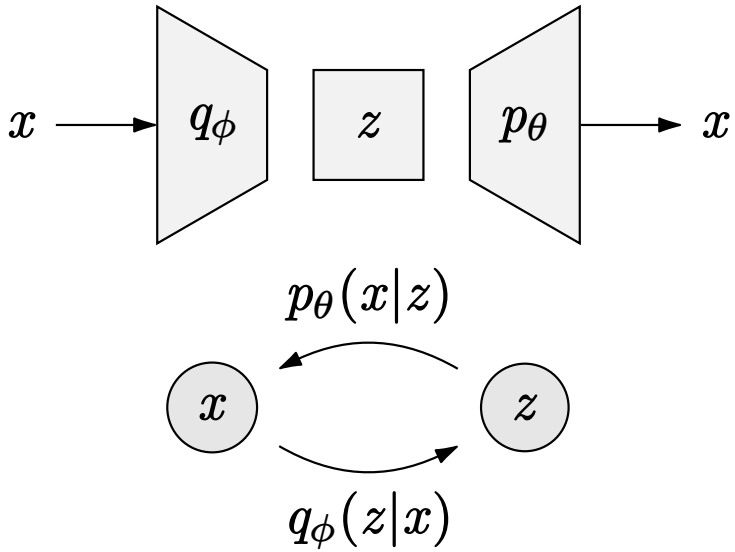
$$= \mathbb{E}_{q_\phi(z|x)}\Big[\log \frac{p(x,z)}{q_\phi(z|x)}\Big] + \mathbb{E}_{q_\phi(z|x)}\Big[\log \frac{q_\phi(z|x)}{p(z|x)}\Big] \tag{7}$$

$$= \mathbb{E}_{q_\phi(z|x)}\Big[\log \frac{p(x,z)}{q_\phi(z|x)}\Big] + D_{\mathrm{KL}}[q_\phi(z|x)\ ||\ p(z|x)] \tag{8}$$

$$\geq \mathbb{E}_{q_\phi(z|x)}\Big[\log \frac{p(x,z)}{q_\phi(z|x)}\Big] \qquad \text{the "}\textit{evidence lower bound}\text{"} \tag{9}$$

Goal: Minimize the KL-divergence, i.e, maximize ELBO.

Variational Autoencoders



$$\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] = \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)\cdot p(z)}{q_\phi(z|x)}\right] \tag{10}$$

$$= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] + \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(z)}{q_\phi(z|x)}\right] \tag{11}$$

$$= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q_\phi(z|x)\,||\,p(z))}_{\text{prior matching term}} \tag{12}$$

VAE: Optimization

The ELBO is optimized jointly over parameters ϕ and θ, with the following priors:

$$q_\phi(z|x) = N(z; \mu_\phi(x), \sigma^2_\phi(x)I) \tag{13}$$

$$p(z) = N(z; 0, I) \tag{14}$$

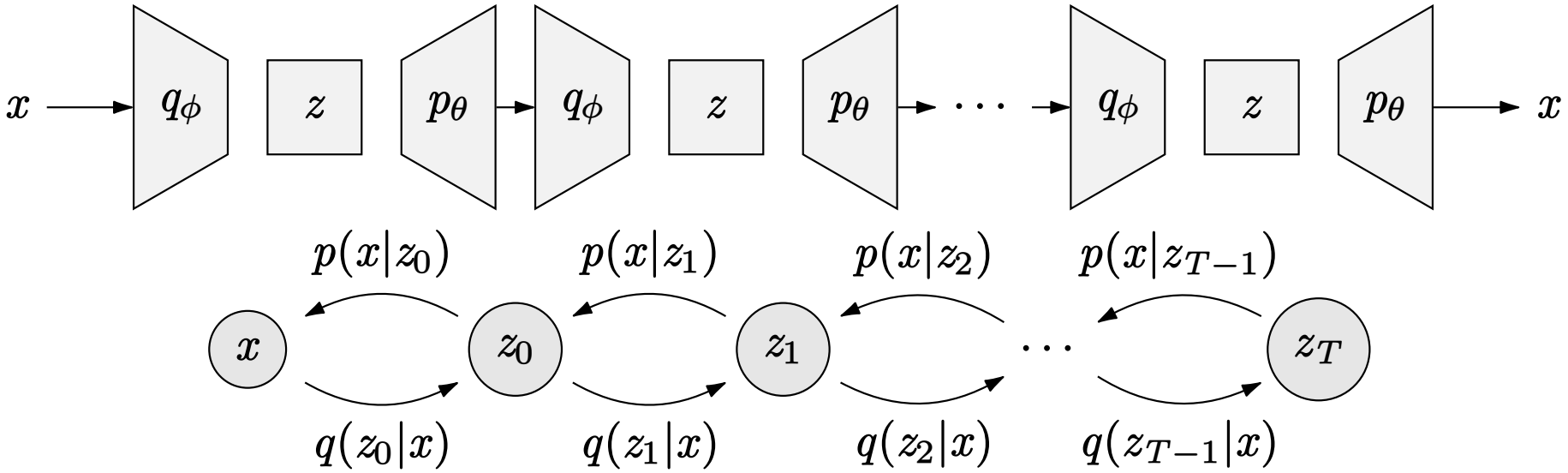
with the objective,

$$\arg\max_{\phi,\theta} \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - D_{\mathrm{KL}}\, q_\phi(z|x) \,||\, p(z) \tag{15}$$

$$\approx \arg\max_{\phi,\theta} \sum_{l=1}^{L} \log p_\theta(x|z^{(l)}) - D_{\mathrm{KL}}\, q_\phi(z|x) \,||\, p(z) \tag{16}$$

where $\left\{z^{(l)}\right\}_{l=1}^{L}$ is the latents sampled from $q_\phi(z|x)$ for every x, with reparameterization $x = \mu + \sigma\epsilon$ and $\epsilon \sim N(0, 1)$.

Markovian Hierarchical Variational Autoencoder



Joint distribution and posterior,

$$p(x, z_{1:T}) = p(z_T)p_\theta(x|z_1)\prod_{t=2}^{T} p_\theta(z_{t-1}|z_t) \tag{17}$$

$$q_\phi(z_{1:T}|x) = q_\phi(z_1|x)\prod_{t=2}^{T} q_\phi(z_t|z_{t-1}) \tag{18}$$

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

Variational Diffusion Model

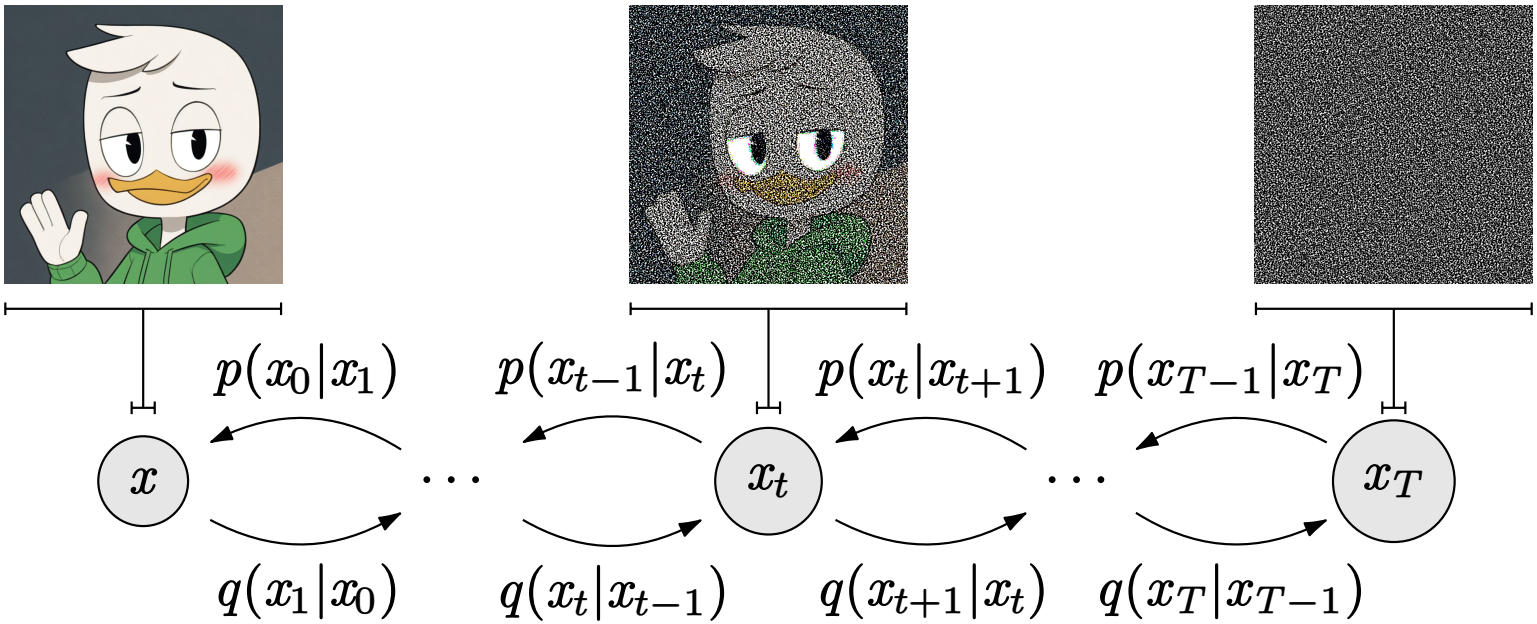
A Variational Diffusion Model is simply a MHVAE with three restrictions,

I. The latent dimension is exactly equal to data dimension.

II. The latent encoder is pre-defined to be linear Gaussian model, i.e, encoder is it is ***not*** learned.

III. The latent encoders' parameters vary over time in such a way that the distribution at the final timestep is standard Gaussian.

Restrictions



Restriction I lets me abuse the notation and write $z = x$ and hence the posterior is,

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x|x_{t-1}) \tag{19}$$

Restriction II lets me encoder as, $q(x_t|x_{t-1} = N(x_t; \sqrt{\alpha_t}\, x_{t-1}, (1-\alpha_t)I)$ And the decoder as (Restriction III),

$$p(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x) \text{ and } p(x_T) = N(x_T; 0, I) \tag{20}$$

VDM ELBO

$$\log p(x) = \underbrace{\mathbb{E}_{q(x_1|x_0)}\left[\log p_\theta(x_0|x_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(x_T|x_0) \,||\, p(x_T))}_{\text{prior matching term}} \tag{21}$$

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

- The *reconstruction term* is the same as the one from vanilla VAE.
- The *prior matching term* from vanilla VAE is KL-div of final noise distribution and standard Gaussian. It is **zero**. (Restriction III)
- The *denoising matching term* is the **most most** important term. This learns how to denoise a noisy input, i.e, minimized when two denoising step match as closely as possible.