Carnegie Mellon

School of Computer Science

Deep Reinforcement Learning and Control

Diffusion models for imitation learning

Fall 2025, CMU 10-703

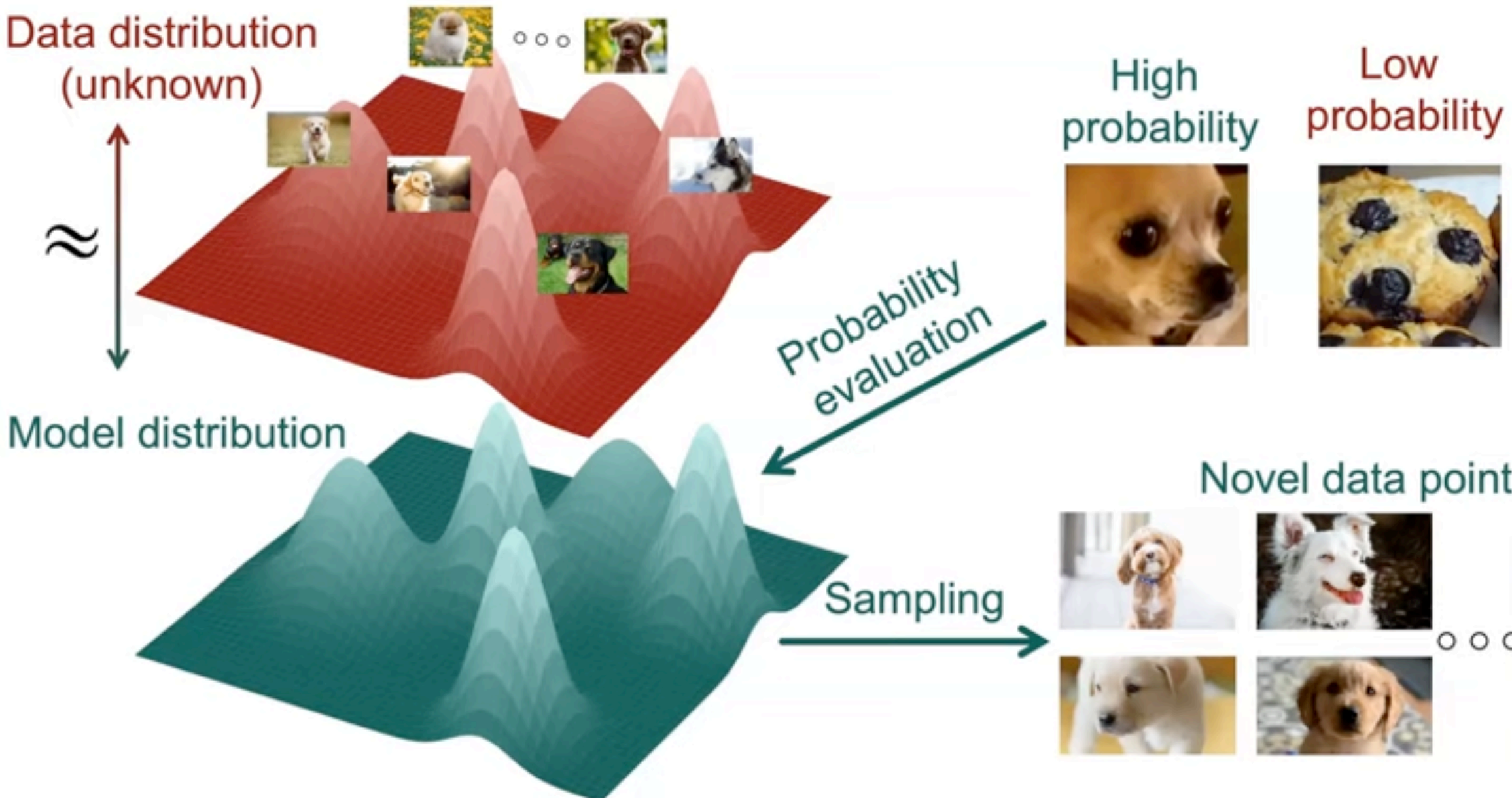
Instructors: Katerina Fragkiadaki and Aviral Kumar

Used Materials

- **Disclaimer:** The material and slides for this lecture are borrowed from the CVPR 2022 tutorial of Kreis, Gao and Vahdat on diffusion models the blogpost of Calvin Luo on Understanding Diffusion Models: A Unified Perspective, and the excellent course of diffusion models and applications CS492 of Minhyuk Sung

Deep Generative Learning

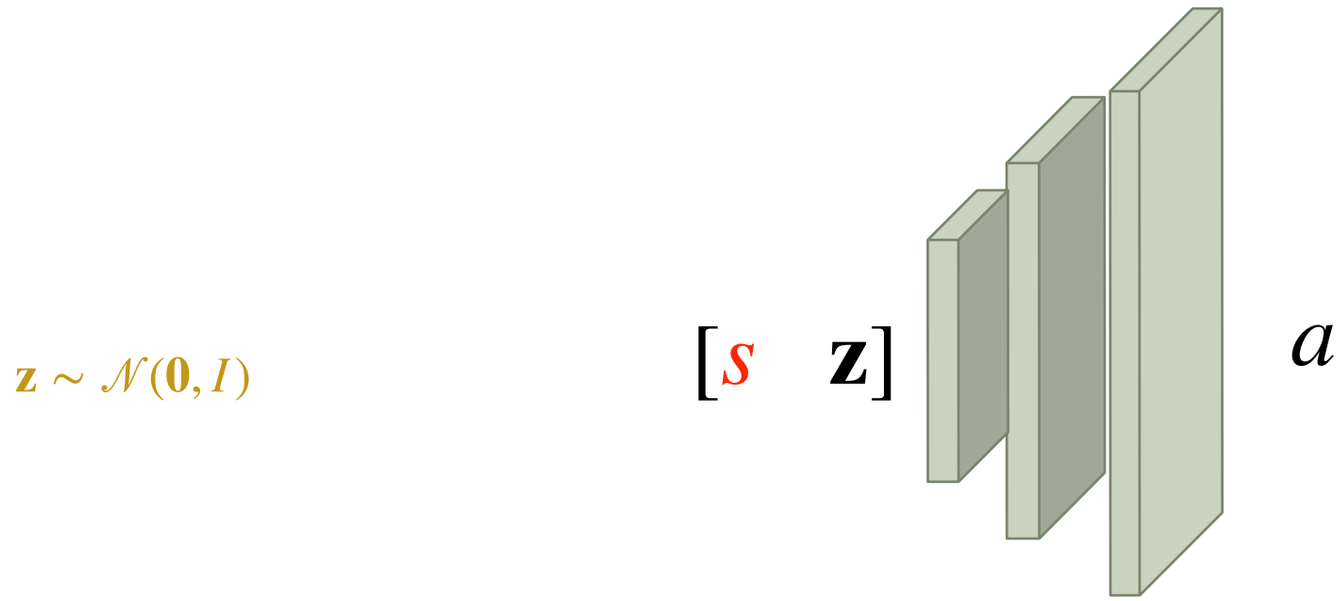
Learning to generate data



From Yang Song

Learning stochastic generative models

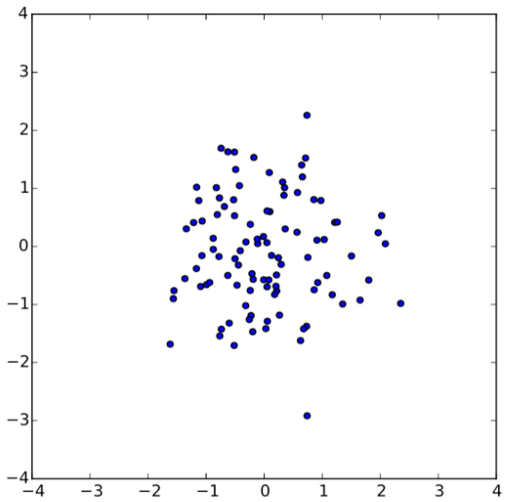
- As we vary the input noisy samples z, we generate an action a.



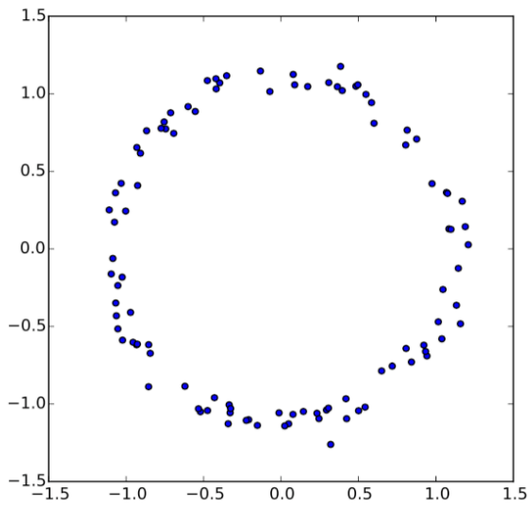
State s: additional conditioning information

Learning stochastic generative models

- Our generative model will transforms the input Gaussian distributions into the desired action distribution.
- Why simple gaussian noise suffices to create complex outputs?
- The neural net will transform it to a complex distribution!



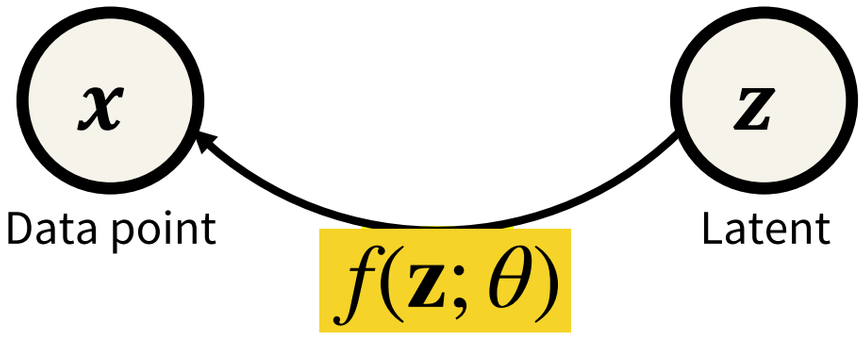
$$z \sim \mathcal{N}(\mathbf{0}, I)$$



$$f(z) = \frac{z}{10} + \frac{z}{\|z\|}$$

Tutotial on variational Autoencoders, Doersch

Variational Autoencoders



$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

We will represent the mapping from a latent distribution $p(\mathbf{z})$ to the data distribution $p(\mathbf{x})$ with the conditional distribution $p(\mathbf{x} \mid \mathbf{z})$:

$$p(\mathbf{x} \mid \mathbf{z}; \theta) = \mathcal{N}(\mathbf{x} \mid f(\mathbf{z}; \theta), \sigma^2 \cdot I)$$

Let's maximize the marginal probaility of the data:

$$\max_{\theta} . \quad p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p(\mathbf{x} \mid \mathbf{z}; \theta) p(\mathbf{z}) d\mathbf{z}$$

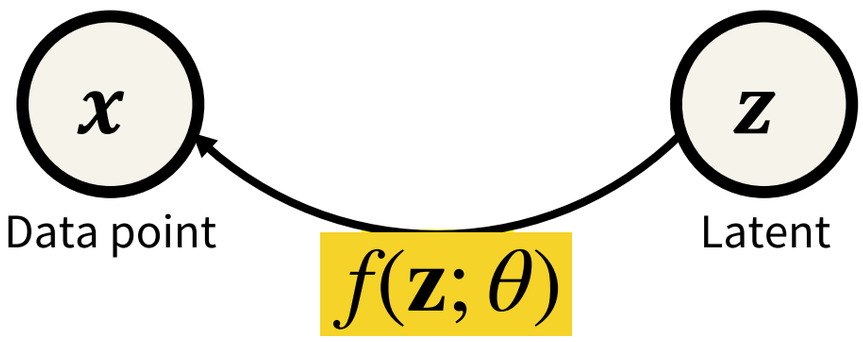
What if we approximate it with sampling:

$$\min_{\theta} . \quad \sum_j -\log p(\mathbf{x}_j) = -\sum_j \sum_{\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, I)} \log p(\mathbf{x}_j \mid \mathbf{z}_i; \theta) = -\sum_j \sum_{i=0, \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, I)}^{K} \frac{1}{K} [\text{const.} - \frac{\|f(\mathbf{z}_i; \theta) - \mathbf{x}_j\|^2}{2\sigma^2}]$$

It is just a bad approximation…

Bayes Rule

$$p(\mathbf{x}\,|\,\mathbf{z};\theta) = \mathcal{N}(\mathbf{x}\,|\,f(\mathbf{z};\theta), \sigma^2 \cdot I)$$



$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{z}\,|\,\mathbf{x}) = p(\mathbf{x}\,|\,\mathbf{z})p(\mathbf{z})$$

$$p(\mathbf{x}) = \frac{p(\mathbf{x}\,|\,\mathbf{z})p(\mathbf{z})}{p(\mathbf{z}\,|\,\mathbf{x})}$$

unknown

Deriving the Evidence Lower Bound (ELBO)

$$\log p(\boldsymbol{x}) = \log p(\boldsymbol{x}) \int q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) d\boldsymbol{z}$$

$$= \int q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{z}$$

Let's consider a helper distribution $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ to sample zs from..

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log p(\boldsymbol{x}) \right]$$

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

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

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})} \right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})}{p(\boldsymbol{z} \mid \boldsymbol{x})} \right]$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})} \right] + \mathcal{D}_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p(\boldsymbol{z} \mid \boldsymbol{x}))$$

$$\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x})} \right]$$

Maximizing the Evidence Lower Bound (ELBO)

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

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

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

- Evidence: the log-likelihood of observed data $p(\mathbf{x})$
- Variational posterior $q_\phi(\mathbf{z}\,|\,\mathbf{x})$ parametrizes a multivariate Gaussian distribution with diagonal covariance: $q_\phi(\mathbf{z}\,|\,\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x})\mathbf{I})$
- Prior $p(\mathbf{z})$ is often a constant multivariate Gaussian: $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$

Maximizing the Evidence Lower Bound (ELBO)

$$\underset{\theta,\phi}{\operatorname{argmax}}\, \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})\right)$$

Approximates using a Monte Carlo estimate:

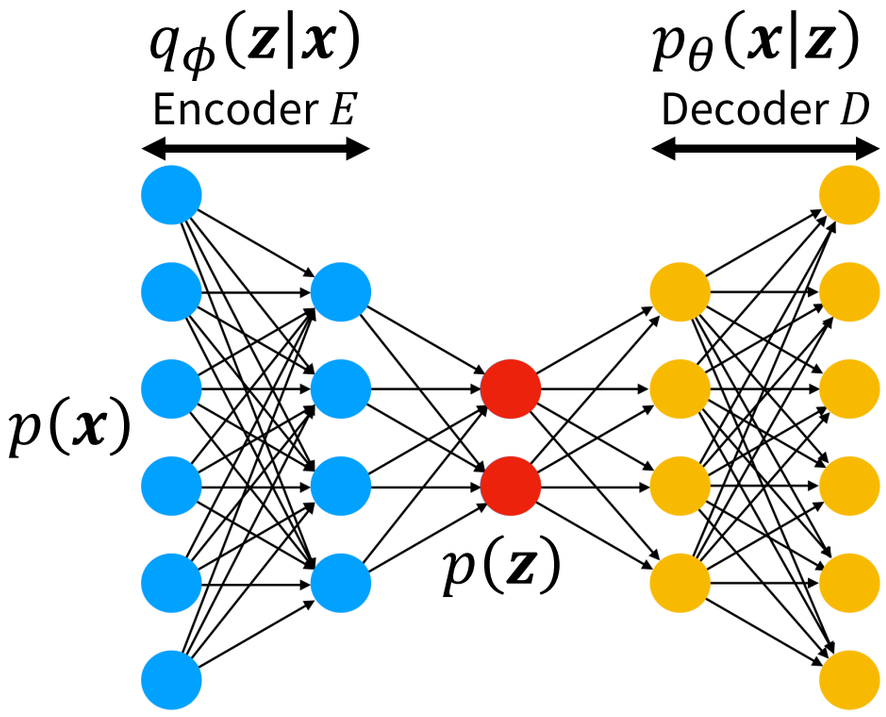
$$\underset{\theta,\phi}{\operatorname{argmax}}\, \frac{1}{N}\sum_{i=1}^{N} \log p_\theta\left(\boldsymbol{x}|\boldsymbol{z}^{(i)}\right) - D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})\right)$$

where $\boldsymbol{z}^{(i)} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$ for the given $\boldsymbol{x}$.

Training Variational Autoencoders

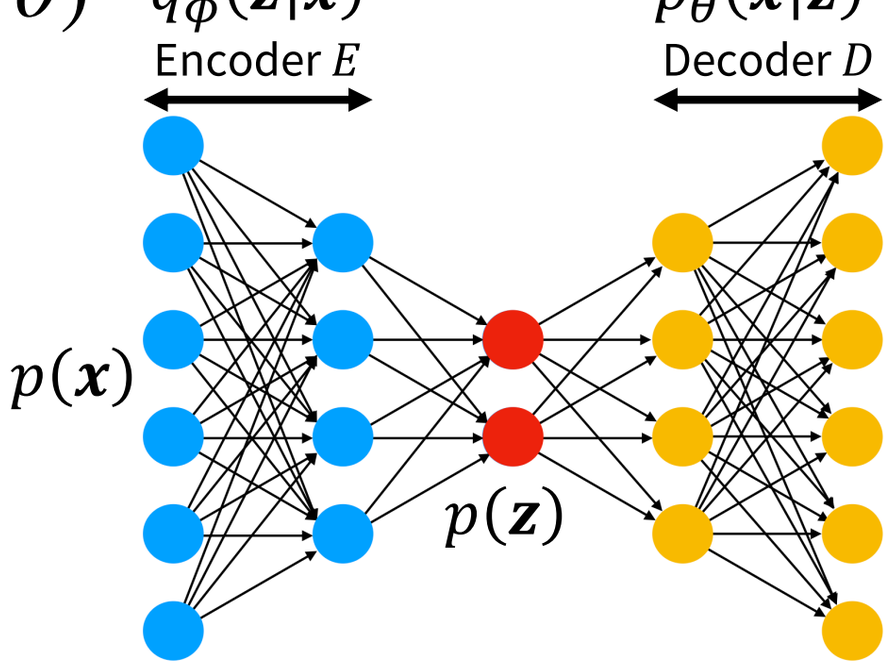
Summary

Data distribution	$p(\boldsymbol{x})$
Encoder	$q_\phi(\boldsymbol{z}\|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_\phi(\boldsymbol{x}), \sigma^2_\phi(\boldsymbol{x})\mathbf{I}\right)$
Prior distribution	$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$
Decoder	$p_\theta(\boldsymbol{x}\|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; D_\theta(\boldsymbol{z}), \sigma^2 I)$

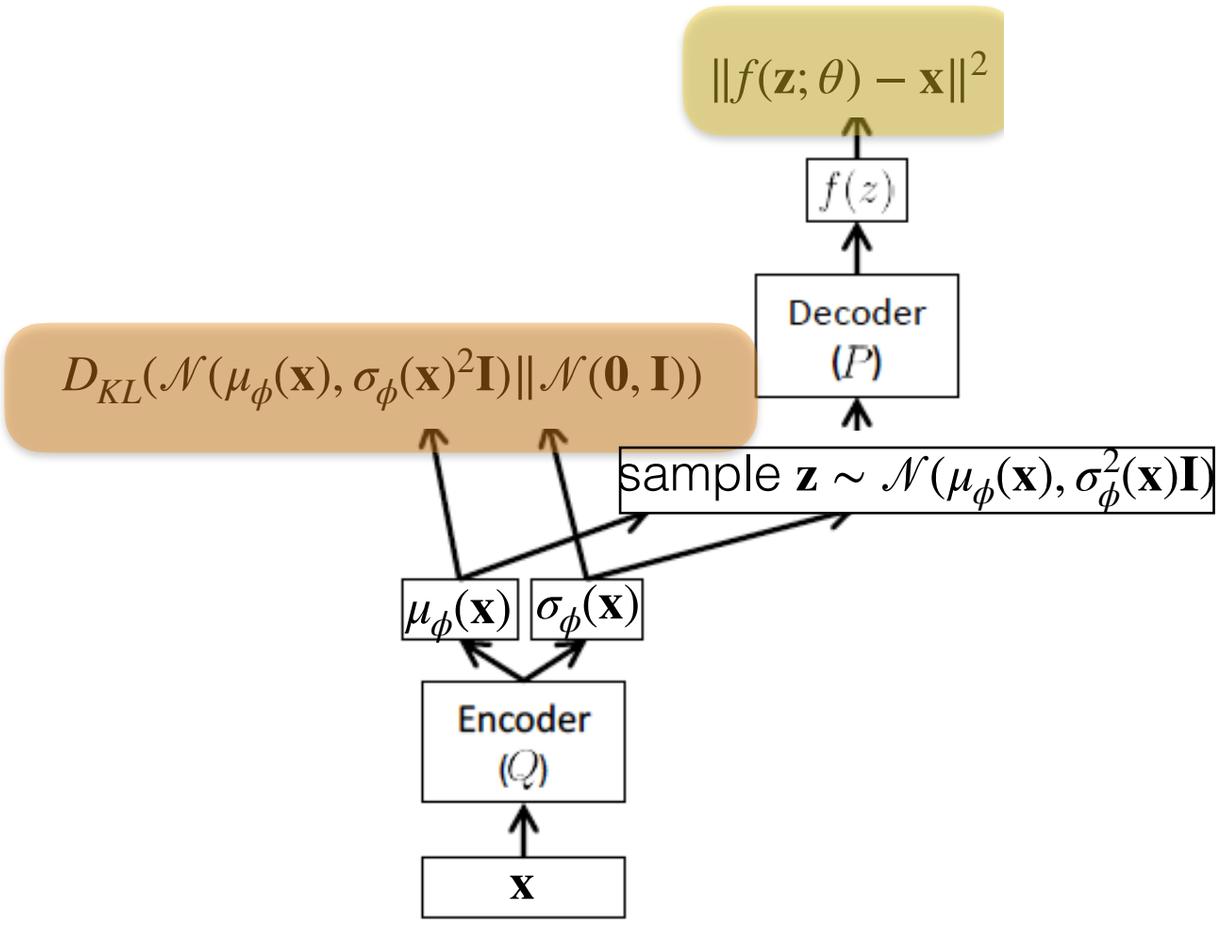


Training Variational Autoencoders

1. Feed a data point $\boldsymbol{x}$ to the encoder to predict $\boldsymbol{\mu}_\phi(\boldsymbol{x})$ and $\sigma^2_\phi(\boldsymbol{x})$.
2. Sample a latent variable $\boldsymbol{z}$ from $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_\phi(\boldsymbol{x}), \sigma^2_\phi(\boldsymbol{x})\mathbf{I}\right)$.
3. Feed $\boldsymbol{z}$ to the decoder to predict $\hat{\boldsymbol{x}} = f(\mathbf{z}; \theta)$
4. Compute the gradient decent through the negative ELBO.



Variational Autoencoder



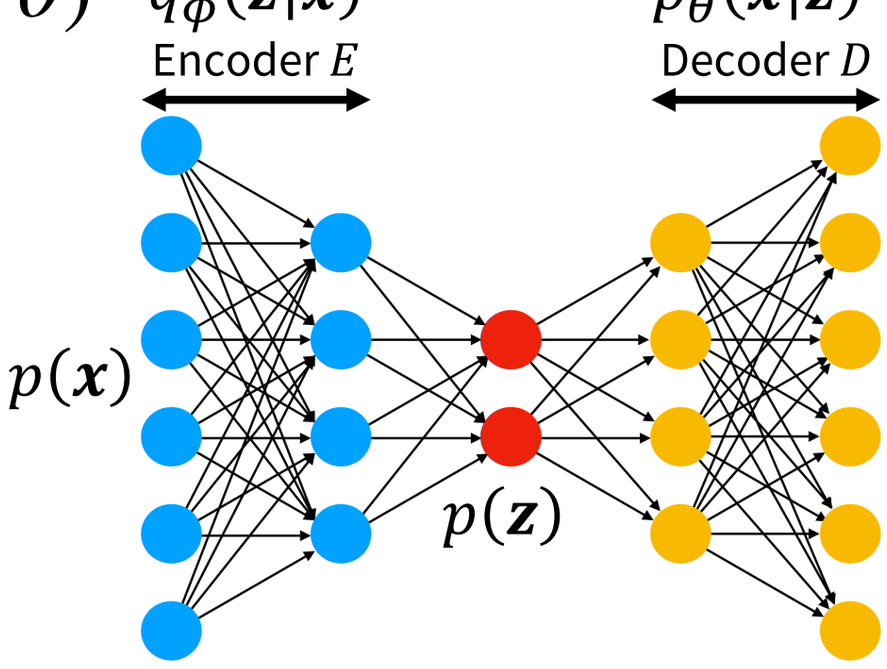
$$\underset{\theta,\phi}{\operatorname{argmax}} \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})\right)$$

$$D_{KL}(\mathcal{N}(\mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x})^2\mathbf{I}) \| \mathcal{N}(\mathbf{0}, \mathbf{I})) = \frac{1}{2}\sum_{\ell=1}^{L}(1 + \log(\sigma_\phi(\mathbf{x})^2) - \mu_\phi(\mathbf{x})^2 - \sigma_\phi(\mathbf{x})^2)$$

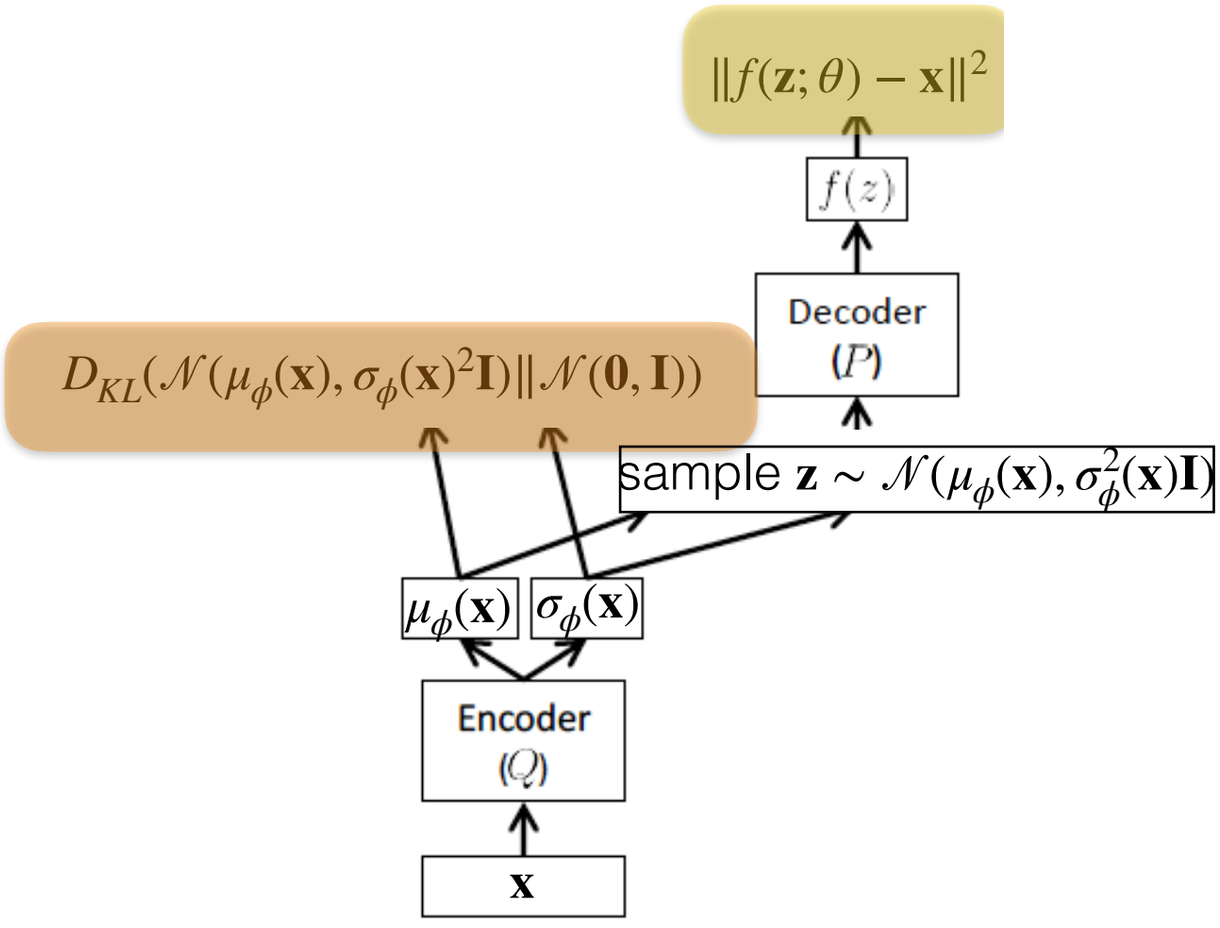
Training Variational Autoencoders

1. Feed a data point $\boldsymbol{x}$ to the encoder to predict $\boldsymbol{\mu}_\phi(\boldsymbol{x})$ and $\sigma^2_\phi(\boldsymbol{x})$.
2. Sample a latent variable $\boldsymbol{z}$ from $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_\phi(\boldsymbol{x}), \sigma^2_\phi(\boldsymbol{x})\mathbf{I}\right)$.
3. Feed $\boldsymbol{z}$ to the decoder to predict $\hat{\boldsymbol{x}} = f(\mathbf{z}; \theta)$
4. Compute the gradient decent through the negative ELBO.

Q. Why is the sampling differentiable?

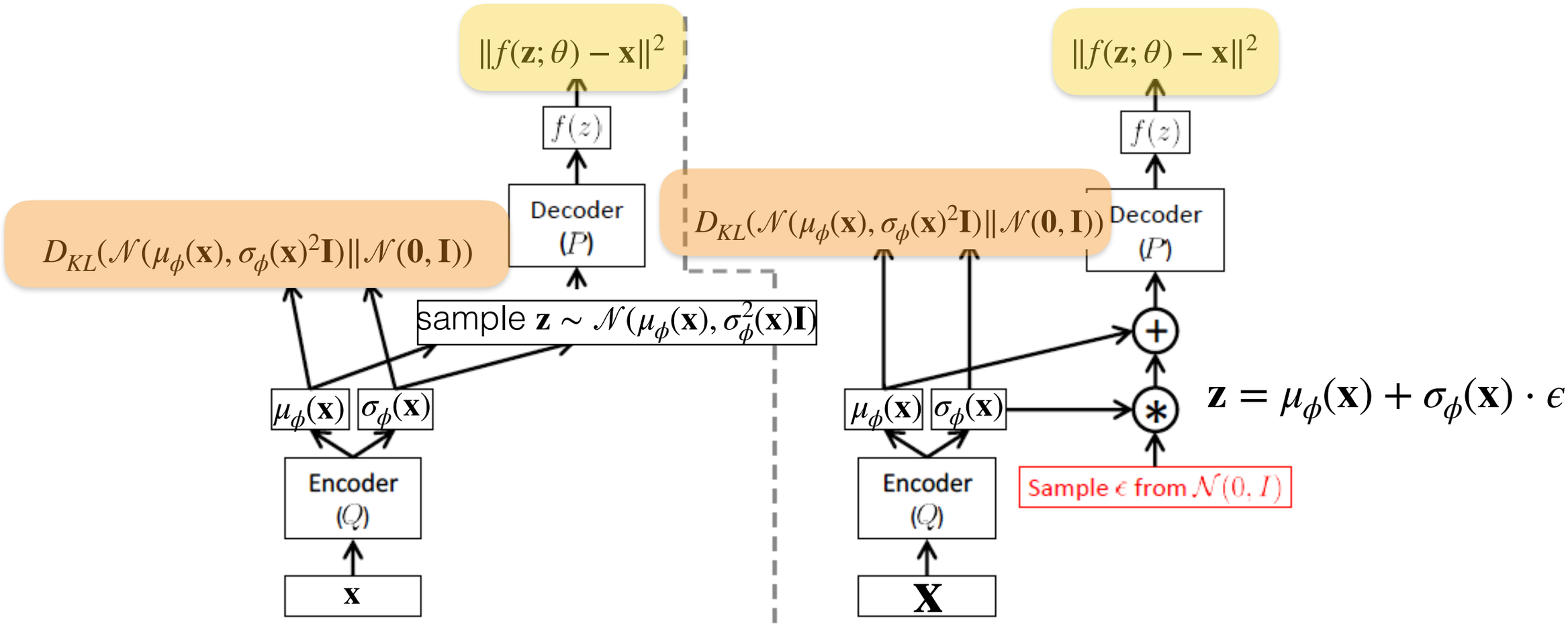


Variational Autoencoder



$$\underset{\theta,\phi}{\operatorname{argmax}} \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})\right)$$

Variational Autoencoder

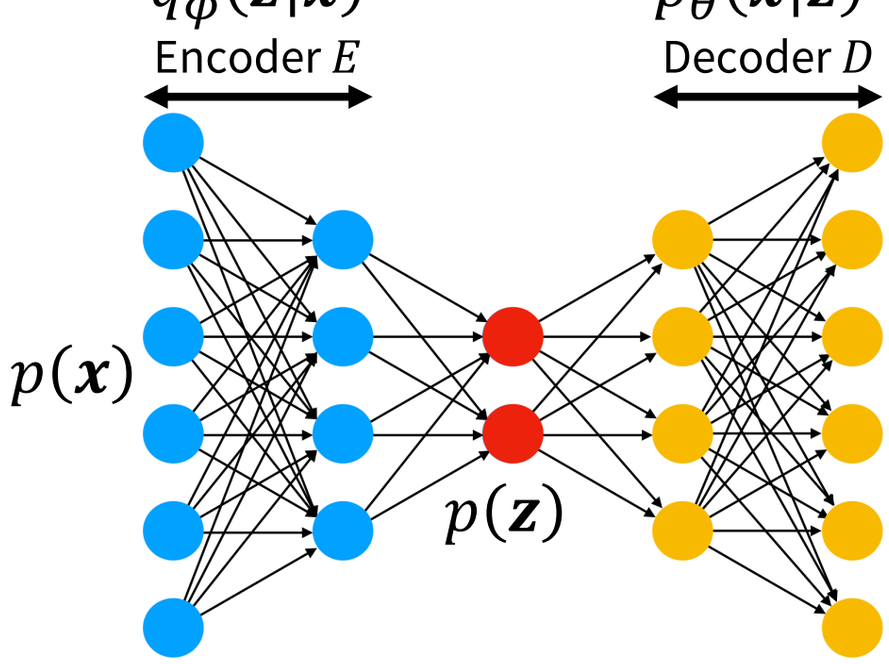


$$\underset{\theta,\phi}{\operatorname{argmax}}\ \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\parallel p(\boldsymbol{z})\right)$$

Variational Autoencoder: Generation

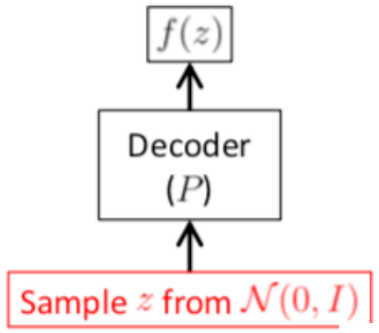
1. Sample a latent variable $\boldsymbol{z}$ from $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$.

2. Feed $\boldsymbol{z}$ to the decoder to predict $\hat{\boldsymbol{x}} = f(\mathbf{z}; \theta)$

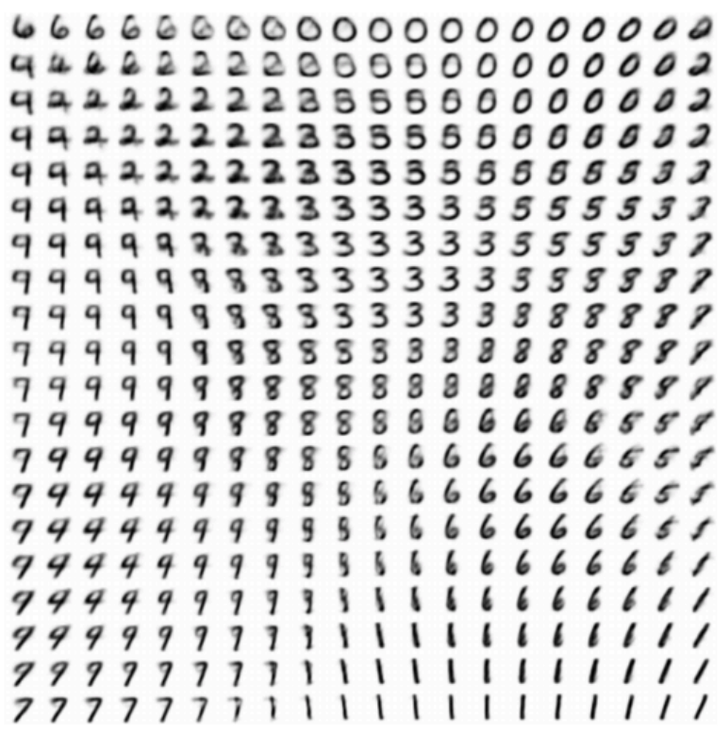


Variational Autoencoder: Generation

At test time



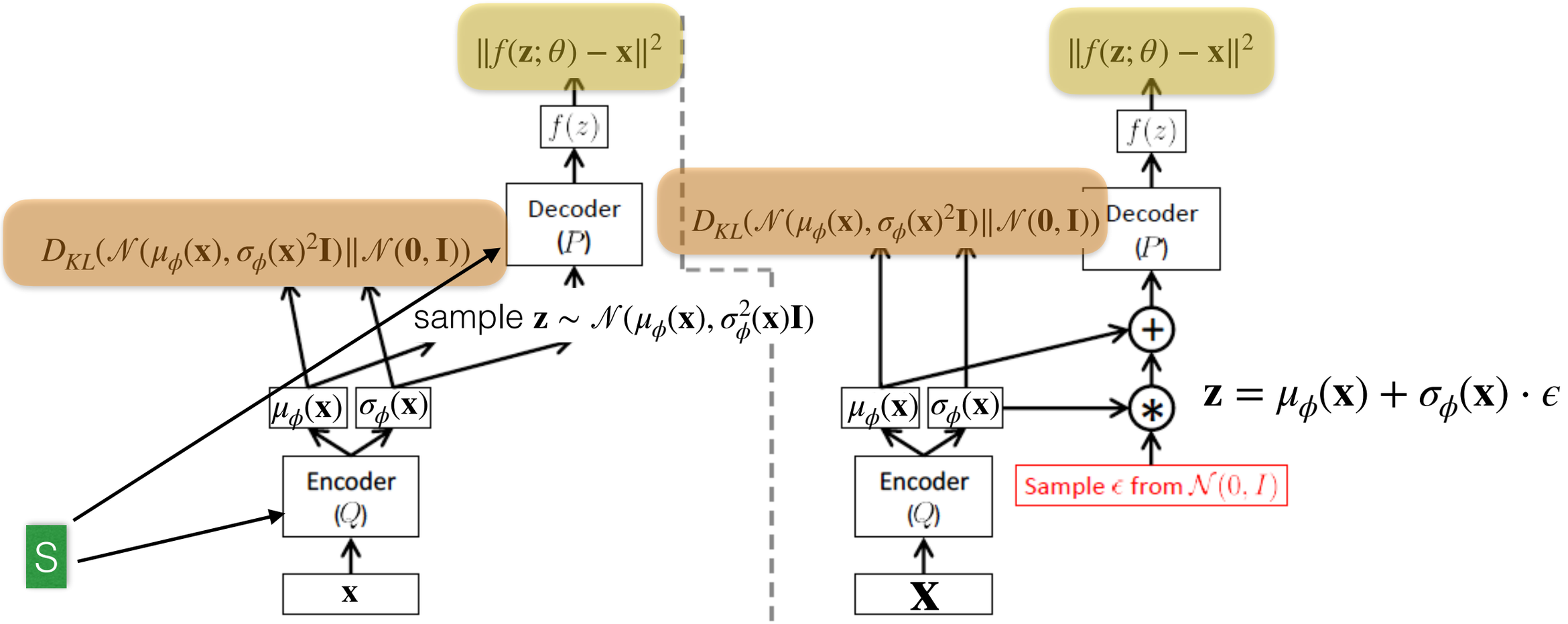
(a) Learned Frey Face manifold



(b) Learned MNIST manifold

Auto-Encoding Variational Bayes, *Kingma and Welling*

Conditional Variational Autoencoder

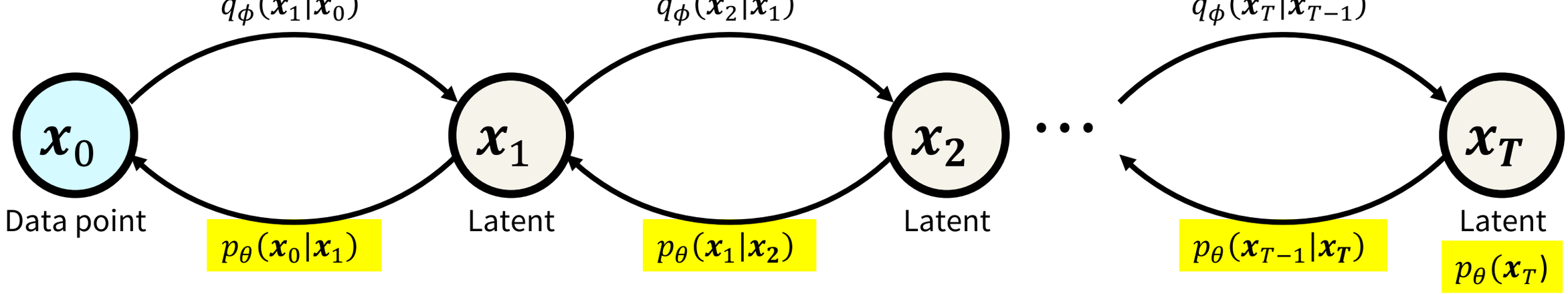


$$\underset{\theta,\phi}{\operatorname{argmax}}\ \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\parallel p(\boldsymbol{z})\right)$$

Markovian Hierarchical VAEs

Joint distribution:

$$p_\theta(\boldsymbol{x}_{0:T}) = p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$$

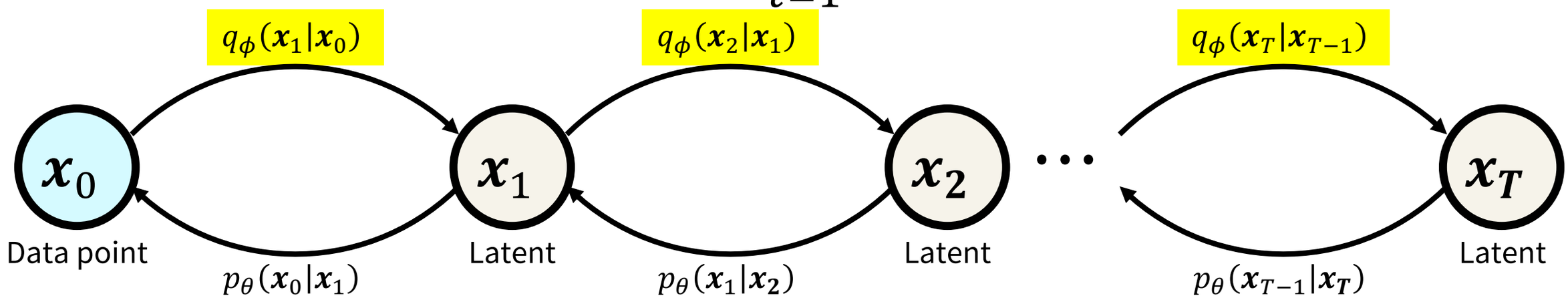


The generative process is modeled as a Markov chain, where decoding each latent only conditions on the previous latent.

Markovian Hierarchical VAEs

Variational posterior:

$$q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q_\phi(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$



The generative process is modeled as a Markov chain, where decoding each latent only conditions on the previous latent.

Deriving the ELBO for Markovian Hierarchical VAEs

$$\log p(\boldsymbol{x}_0) = \log \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T}$$

$$= \log \int p_\theta(\boldsymbol{x}_{0:T}) \frac{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T}$$

$$= \log \mathbb{E}_{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p_\theta(\boldsymbol{x}_{0:T})}{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right]$$

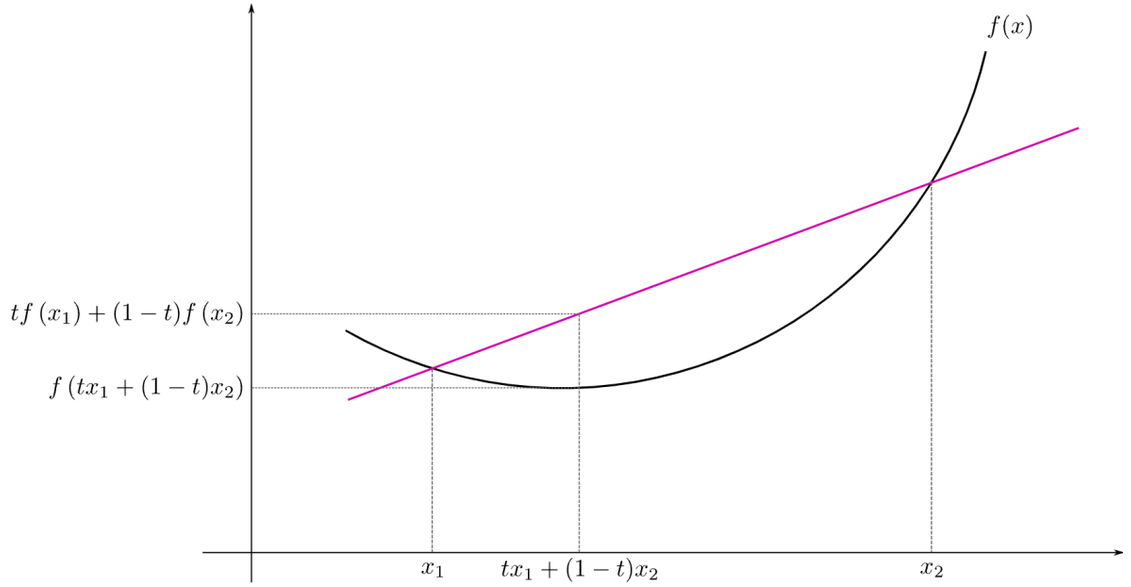
$$\geq \mathbb{E}_{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right]$$

Jensen's Inequality

f is a convex function if

$$f(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_2) \leq tf(\boldsymbol{x}_1) + (1-t)f(\boldsymbol{x}_2)$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2$, and $t \in [0,1]$.



Log is concave..

Deriving the ELBO for Markovian Hierarchical VAEs

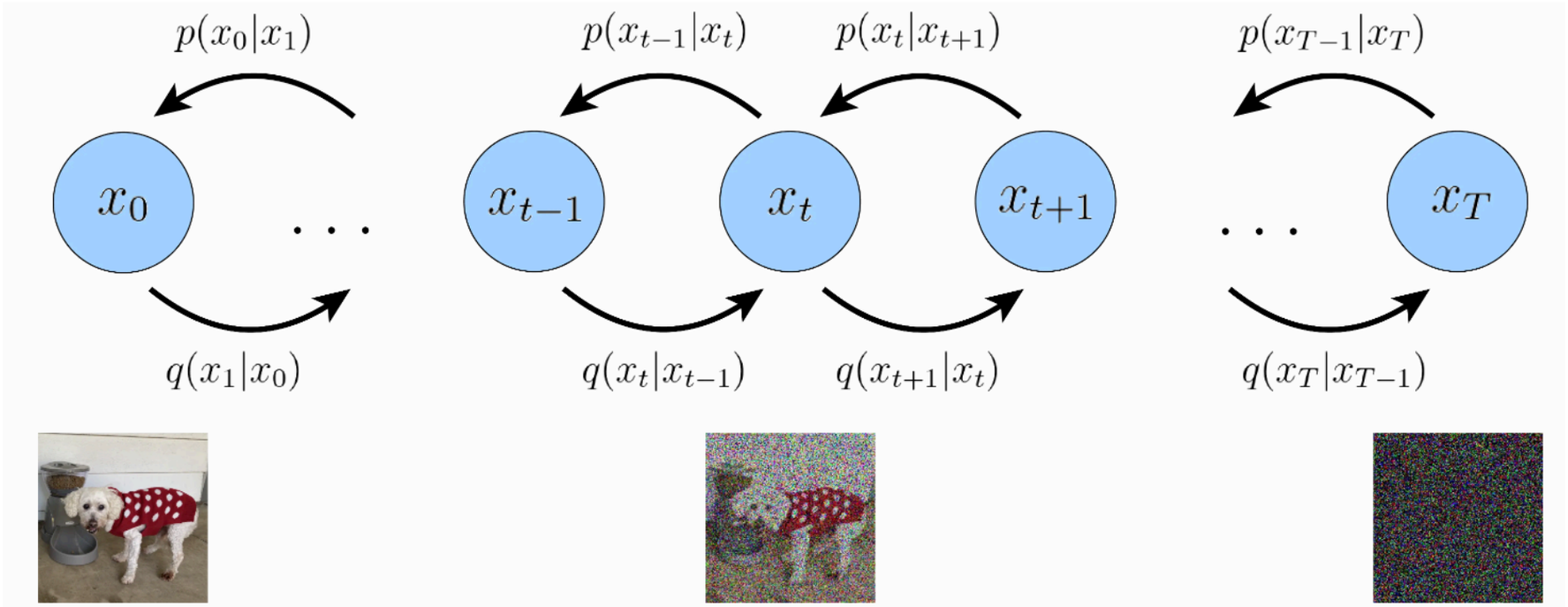
$$\log p(\boldsymbol{x}_0) = \log \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T}$$

$$= \log \int p_\theta(\boldsymbol{x}_{0:T}) \frac{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T}$$

$$= \log \mathbb{E}_{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p_\theta(\boldsymbol{x}_{0:T})}{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right]$$

$$\geq \mathbb{E}_{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right]$$

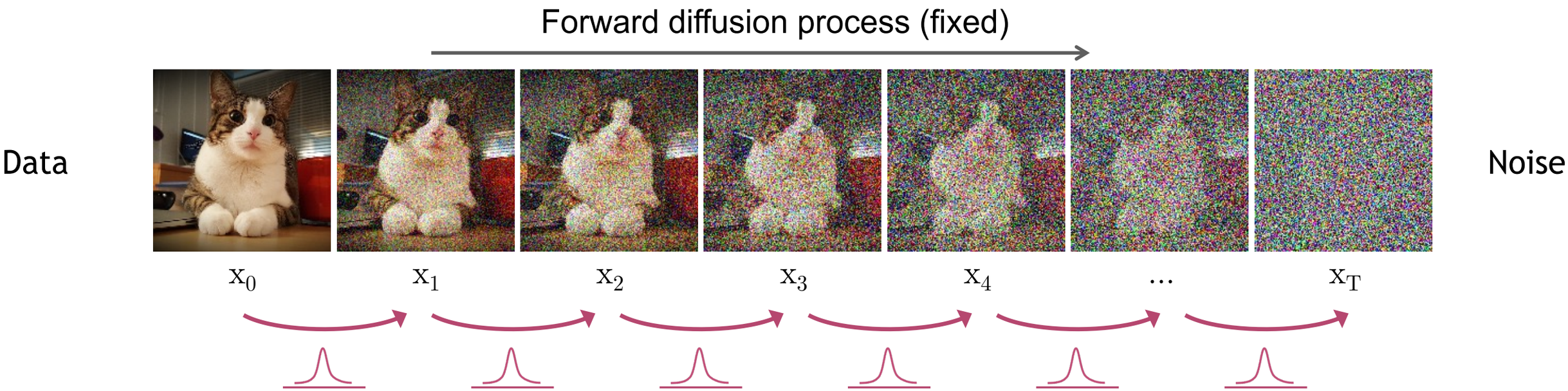
Diffusion Models



One way to understand Diffusion Models is as Markovian Hierarchical VAEs with:

- The **latent dimension is exactly equal to the data dimension**.
- **The encoder at each timestep is not learned**; it is pre-defined as a Gaussian distribution centered around the output of the previous timestep
- The Gaussian parameters of the latent encoders vary over time in such a way that **the distribution of the latent at final timestep T is a standard Gaussian**.

Fixed encoding process

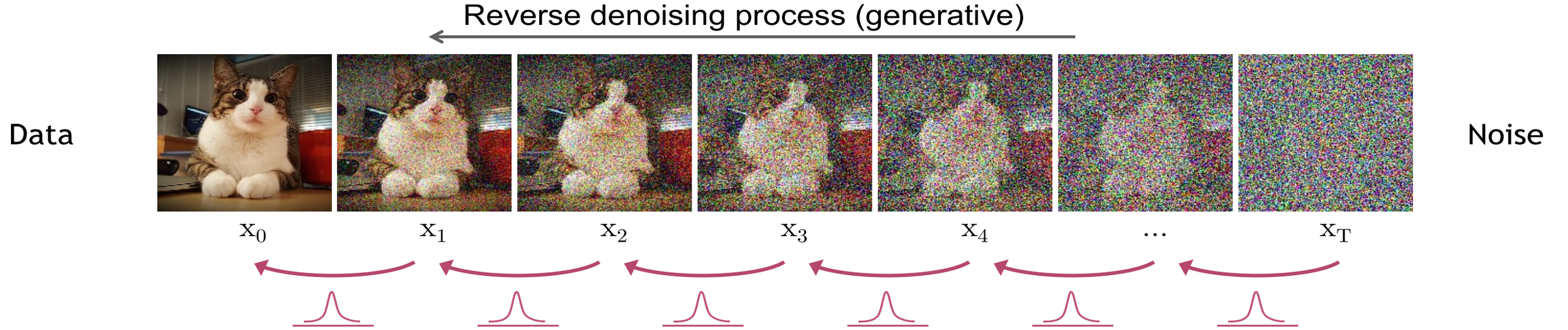


The distribution of each latent variable in the encoder is a Gaussian centered around its previous hierarchical latent:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

Reverse denoising process

Data generation: Sample Gaussian noise from $p(\mathbf{x}_T)$ and iteratively run the denoising transitions $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ for T steps to generate a novel $\mathbf{x}_0$.



$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I})$$

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

Summary so far

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

Data

Noise

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

Observations:

- No learnable parameters ϕ for the encoding process
- We are only interested in learning $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ to generate data

$q(\mathbf{x}_t \mid \mathbf{x}_0)$ is Gaussian

Data

Noise

x_0 x_1 x_2 x_3 x_4 … x_T

$q(\mathbf{x}_t \mid \mathbf{x}_0)$

$q(\mathbf{x}_t \mid \mathbf{x}_0)$ is Gaussian!

Data

Noise

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

For an arbitrary sample $\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)$:

$$\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1}
\end{aligned}$$

Combination of Gaussian Variables

Suppose $\boldsymbol{x}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2 \mathbf{I})$ and $\boldsymbol{x}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2 \mathbf{I})$.

Q. What is the distribution of $\boldsymbol{x}_1 + \boldsymbol{x}_2$?

A. $\boldsymbol{x}_1 + \boldsymbol{x}_2 \sim \mathcal{N}(\mu_1 + \mu_2, (\sigma_1^2 + \sigma_2^2)\mathbf{I})$

Suppose $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\boldsymbol{x}_1 = \sigma_1 \boldsymbol{\varepsilon}_1$ and $\boldsymbol{x}_2 = \sigma_2 \boldsymbol{\varepsilon}_2$.

Q. What is the distribution of $\boldsymbol{x}_1 + \boldsymbol{x}_2$?

A. $\boldsymbol{x}_1 + \boldsymbol{x}_2 \sim \mathcal{N}(\mathbf{0}, (\sigma_1^2 + \sigma_2^2)\mathbf{I})$.

$\boldsymbol{x}_1 + \boldsymbol{x}_2 = \sqrt{\sigma_1^2 + \sigma_2^2}\,\boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is another standard normal sample.

$q(\mathbf{x}_t \mid \mathbf{x}_0)$ is Gaussian!

Data

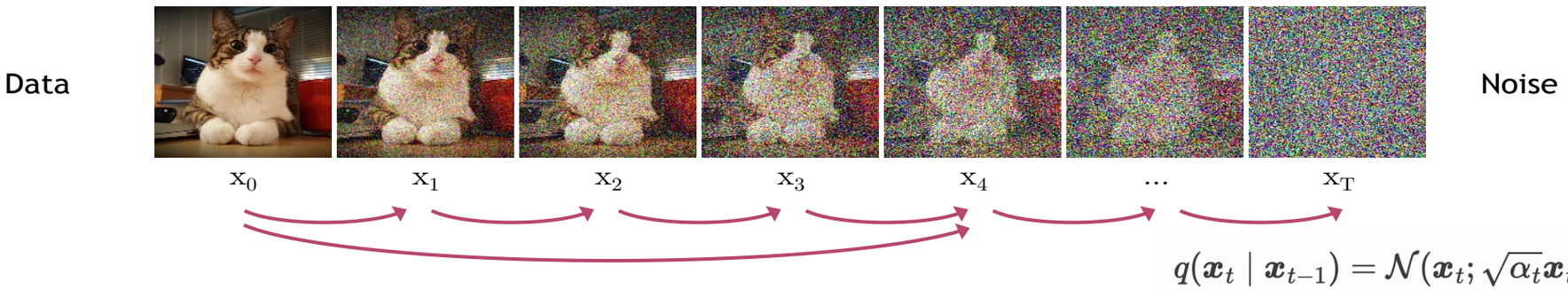
Noise

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

For an arbitrary sample $\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)$:

$$\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1}
\end{aligned}$$

$q(\mathbf{x}_t \mid \mathbf{x}_0)$ is Gaussian!



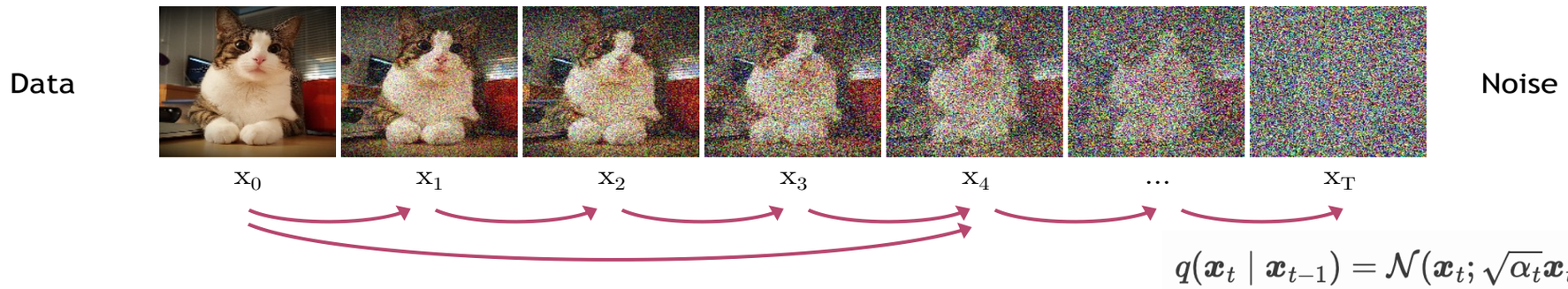
$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

For an arbitrary sample $\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)$:

$$\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}^2 + \sqrt{1-\alpha_t}^2}\,\boldsymbol{\epsilon}_{t-2}
\end{aligned}$$

The sum of two independent Gaussian random variables is a Gaussian with mean being the sum of the two means, and variance being the sum of the two variances

$q(\mathbf{x}_t \mid \mathbf{x}_0)$ is Gaussian!



$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

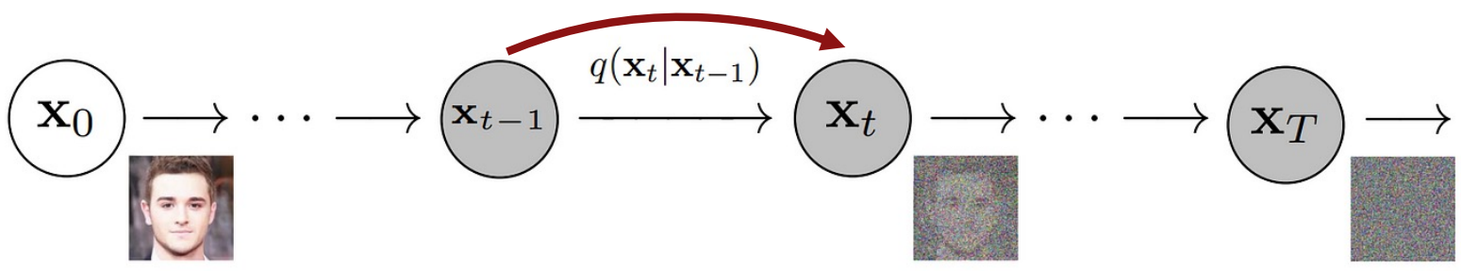
For an arbitrary sample $\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)$:

$$\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}^2 + \sqrt{1-\alpha_t}^2}\,\boldsymbol{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1} + 1 - \alpha_t}\boldsymbol{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} \\
&= \dots \\
&= \sqrt{\prod_{i=1}^{t}\alpha_i}\boldsymbol{x}_0 + \sqrt{1-\prod_{i=1}^{t}\alpha_i}\boldsymbol{\epsilon}_0 \\
&= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0
\end{aligned}$$

The sum of two independent Gaussian random variables is a Gaussian with mean being the sum of the two means, and variance being the sum of the two variances

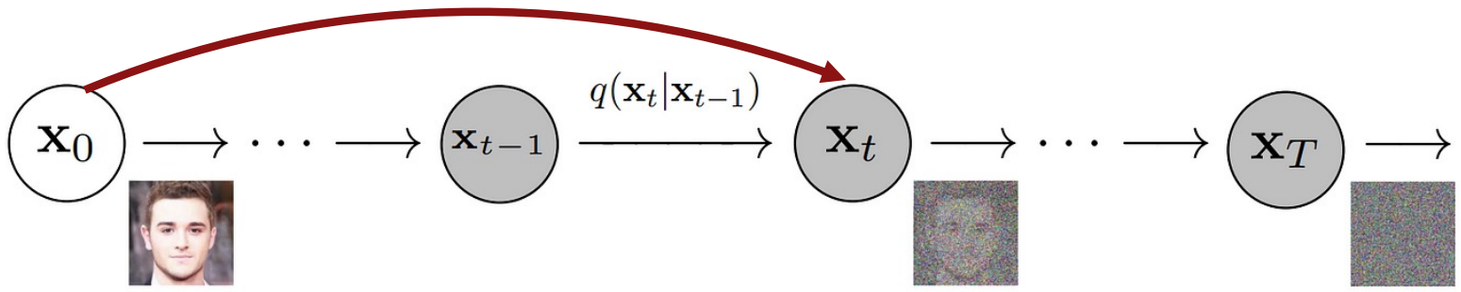
Fixed encoding process

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$



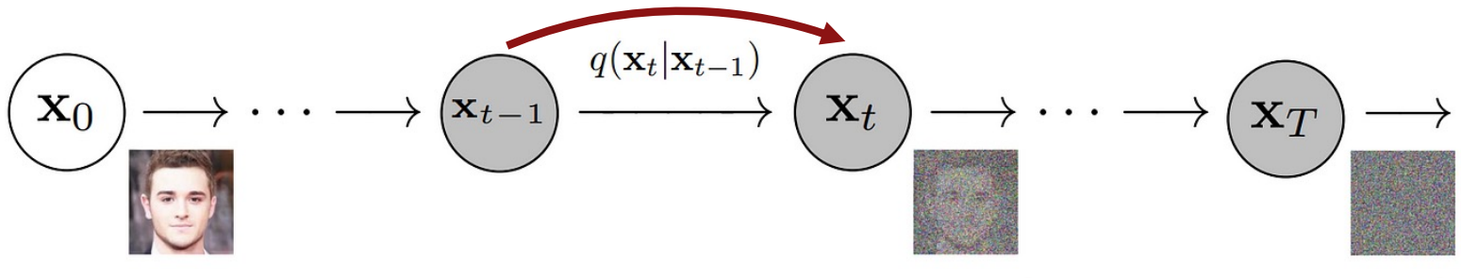
$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ *Also a normal distribution!*



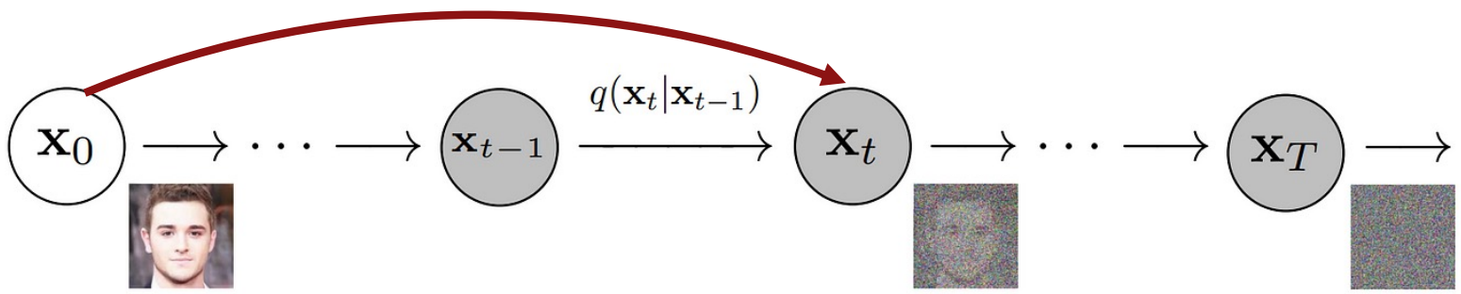
Fixed encoding process

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$



$$q(\boldsymbol{x}_T|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_T; \sqrt{\bar{\alpha}_T}\boldsymbol{x}_0, (1-\bar{\alpha}_T)\mathbf{I}\right)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ *Also a normal distribution!*

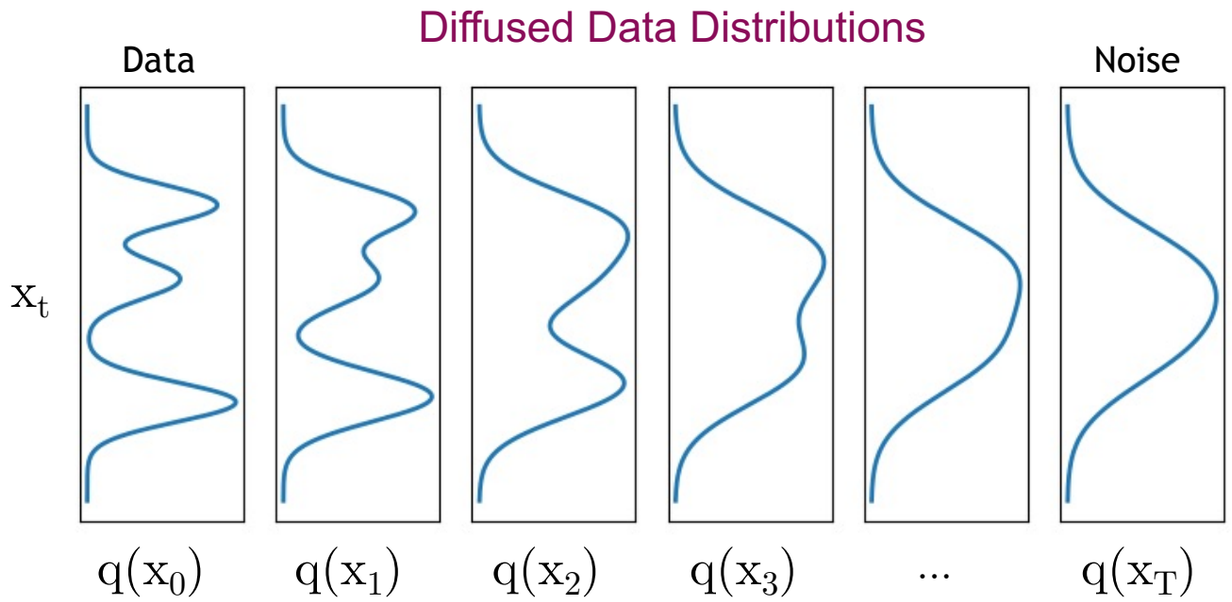


As $T \to \infty$, $q(\boldsymbol{x}_T|\boldsymbol{x}_0)$ converges to the standard normal distribution $\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \mathbf{I})$.

Fixed encoding process

$$\underbrace{q(\mathbf{x}_t)}_{\text{Diffused data dist.}} = \int \underbrace{q(\mathbf{x}_0, \mathbf{x}_t)}_{\text{Joint dist.}} d\mathbf{x}_0 = \int \underbrace{q(\mathbf{x}_0)}_{\text{Input data dist.}} \underbrace{q(\mathbf{x}_t|\mathbf{x}_0)}_{\text{Diffusion kernel}} d\mathbf{x}_0$$

The diffusion kernel is Gaussian convolution.



We can sample $\mathbf{x}_t \sim q(\mathbf{x}_t)$ by first sampling $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ and then sampling $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)$ (i.e., ancestral sampling).

21

Reverse denoising process

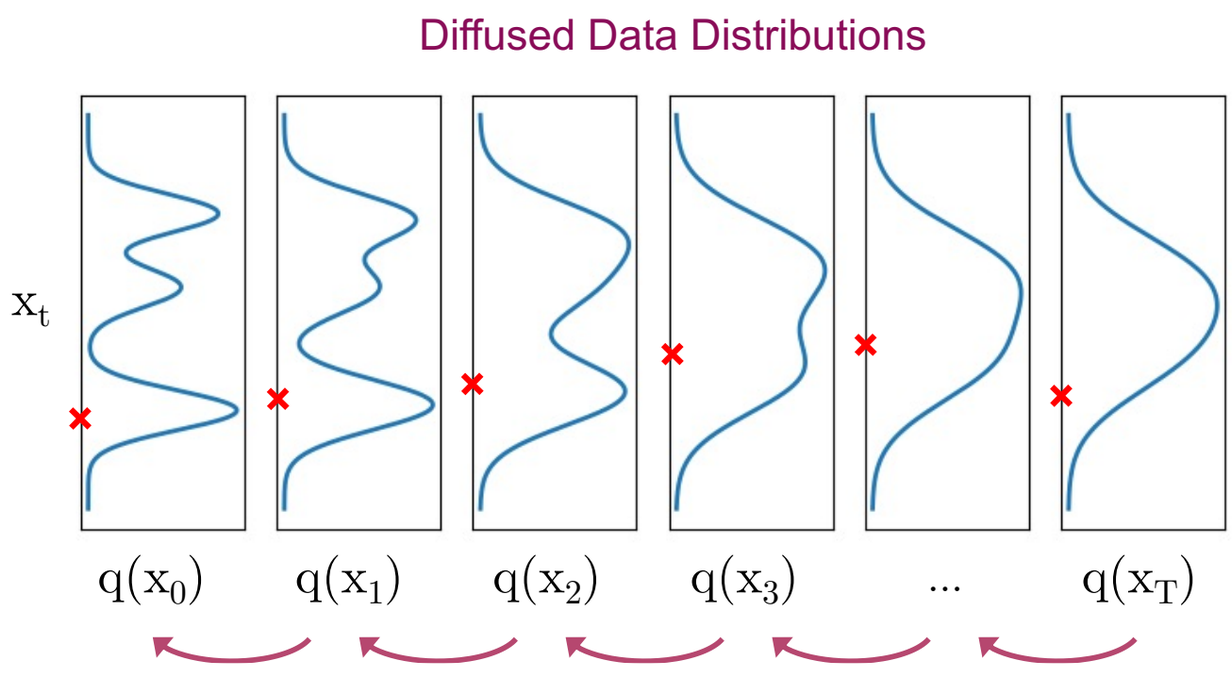
Data generation: Sample Gaussian noise from $p(\mathbf{x}_T)$ and iteratively run the denoising transitions $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ for T steps to generate a novel $\mathbf{x}_0$.

Generation:

Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$

Iteratively sample $\mathbf{x}_{t-1} \sim p(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$

True Denoising Dist.

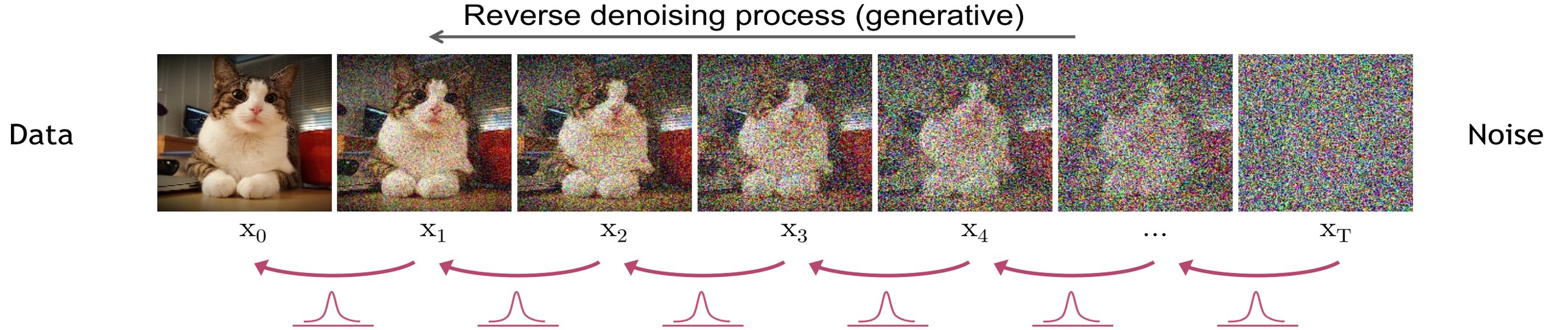


In general, $p(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ is intractable.

We will approximate it with a Gaussian distribution for small $1 - \alpha_t$.

Reverse denoising process

Data generation: Sample Gaussian noise from $p(\mathbf{x}_T)$ and iteratively run the denoising transitions $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ for T steps to generate a novel $\mathbf{x}_0$.



$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I})$$

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

Optimizing Diffusion Models by maximizing the ELBO

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=1}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} [\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \,||\, p(\boldsymbol{x}_T))]}_{\text{prior matching term}} \\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} [\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \,||\, p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}))]}_{\text{consistency term}}
\end{aligned}
$$

Optimizing Diffusion Models by maximizing the ELBO

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=1}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} [\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \,||\, p(\boldsymbol{x}_T))]}_{\text{prior matching term}} \\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} [\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \,||\, p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}))]}_{\text{consistency term}}
\end{aligned}
$$

Reconstuction term, same as vanilla VAEs given the first step latent.

Optimizing Diffusion Models by maximizing the ELBO

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=1}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} [\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \,||\, p(\boldsymbol{x}_T))]}_{\text{prior matching term}} \\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} [\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \,||\, p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}))]}_{\text{consistency term}}
\end{aligned}
$$

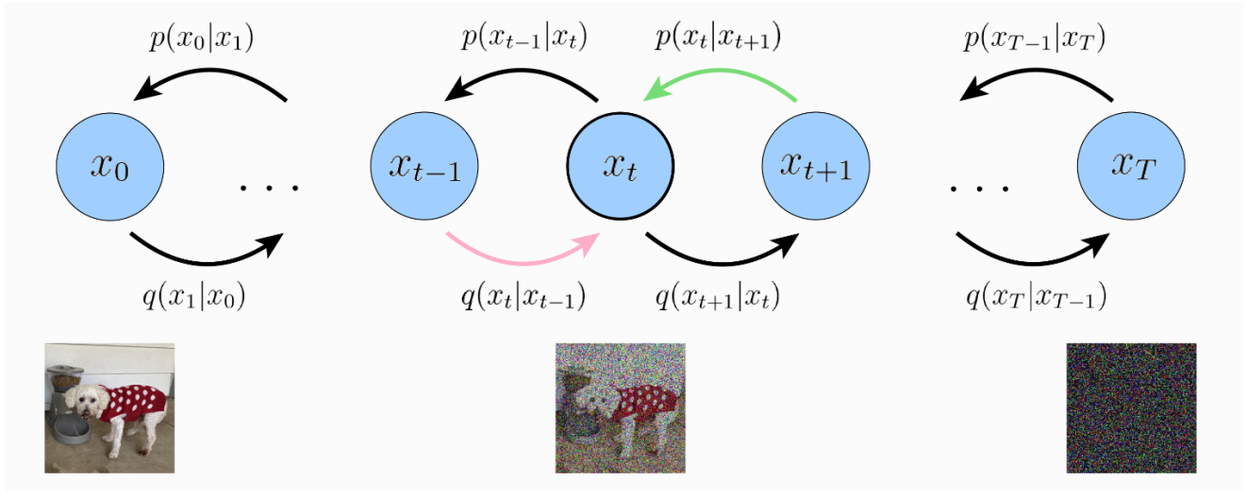
Prior matching term, no trainable parameters, the fial latent distribution need to match the Gaussian prior

Optimizing Diffusion Models by maximizing the ELBO

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=1}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_{\theta}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} [\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \,||\, p(\boldsymbol{x}_T))]}_{\text{prior matching term}} \\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} [\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \,||\, p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}))]}_{\text{consistency term}}
\end{aligned}
$$

The distribution over x_t obtained from transitioning from noisier samples x_{t+1} should match the distribution over x_t obtained from noising cleaner samples x_{t-1}, for every t.

Optimizing Diffusion Models by maximizing the ELBO

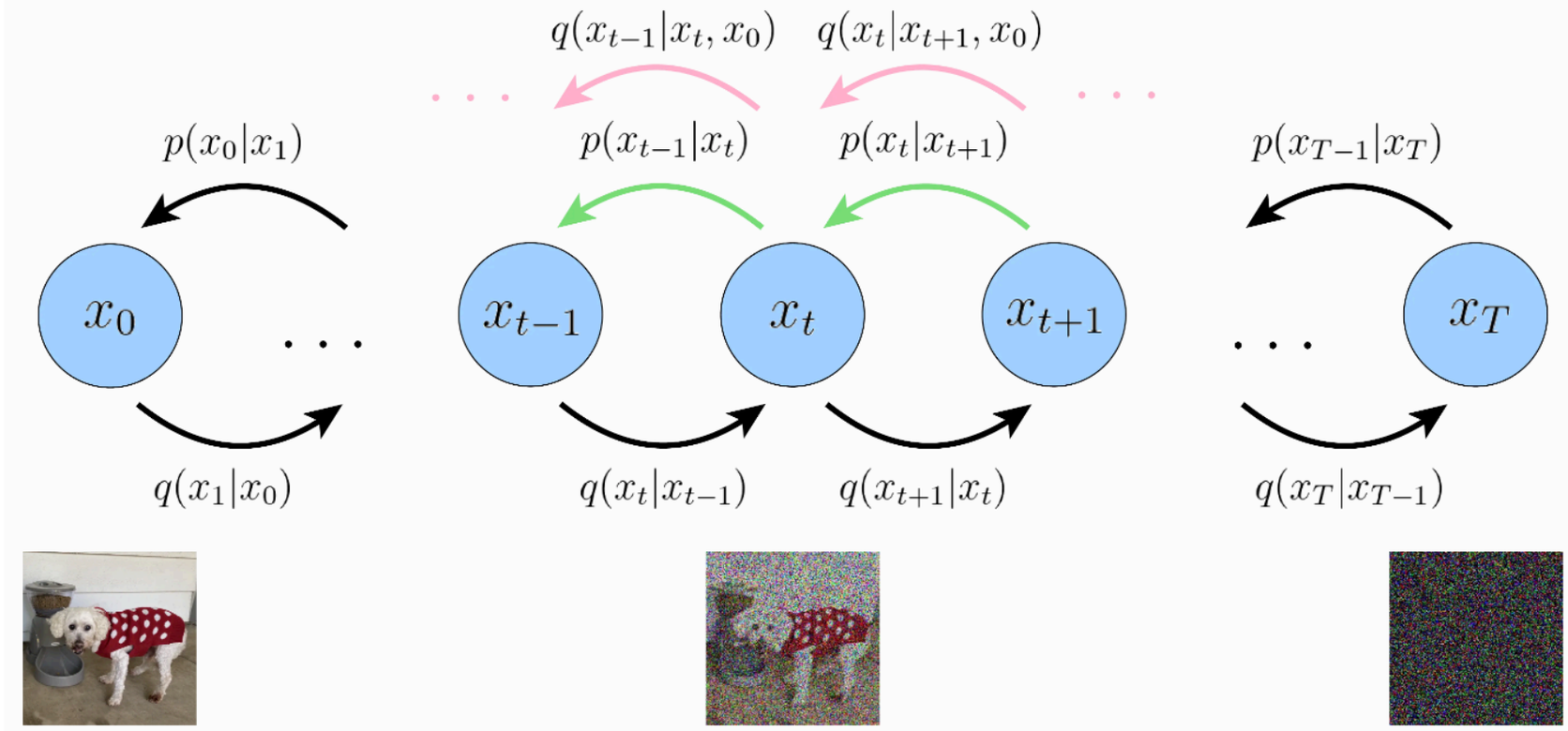


The distribution over x_t obtained from transitioning from noisier samples x_{t+1} should match the distribution over x_t obtained from transitioning from cleaner samples x_{t-1}, for every t.

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right]$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} \left[\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \,||\, p(\boldsymbol{x}_T)) \right]}_{\text{prior matching term}}$$

$$- \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \,||\, p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})) \right]}_{\text{consistency term}}$$



Key idea: the encoding process will use x_t and $x_0$to predisct x_{t-1} instead of x_{t-2} .

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right]$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_0) \,||\, p(\boldsymbol{x}_T))}_{\text{prior matching term}}$$

$$- \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

Gaussian form for $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$

$$
\begin{aligned}
q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}) \mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})} \\
&\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}, \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I})
\end{aligned}
$$

Gaussian form for $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$

Remember: $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$

$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$

$$
\begin{aligned}
q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}) \mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})} \\
&\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}, \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I})
\end{aligned}
$$

Gaussian form for $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$

Remember: $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$

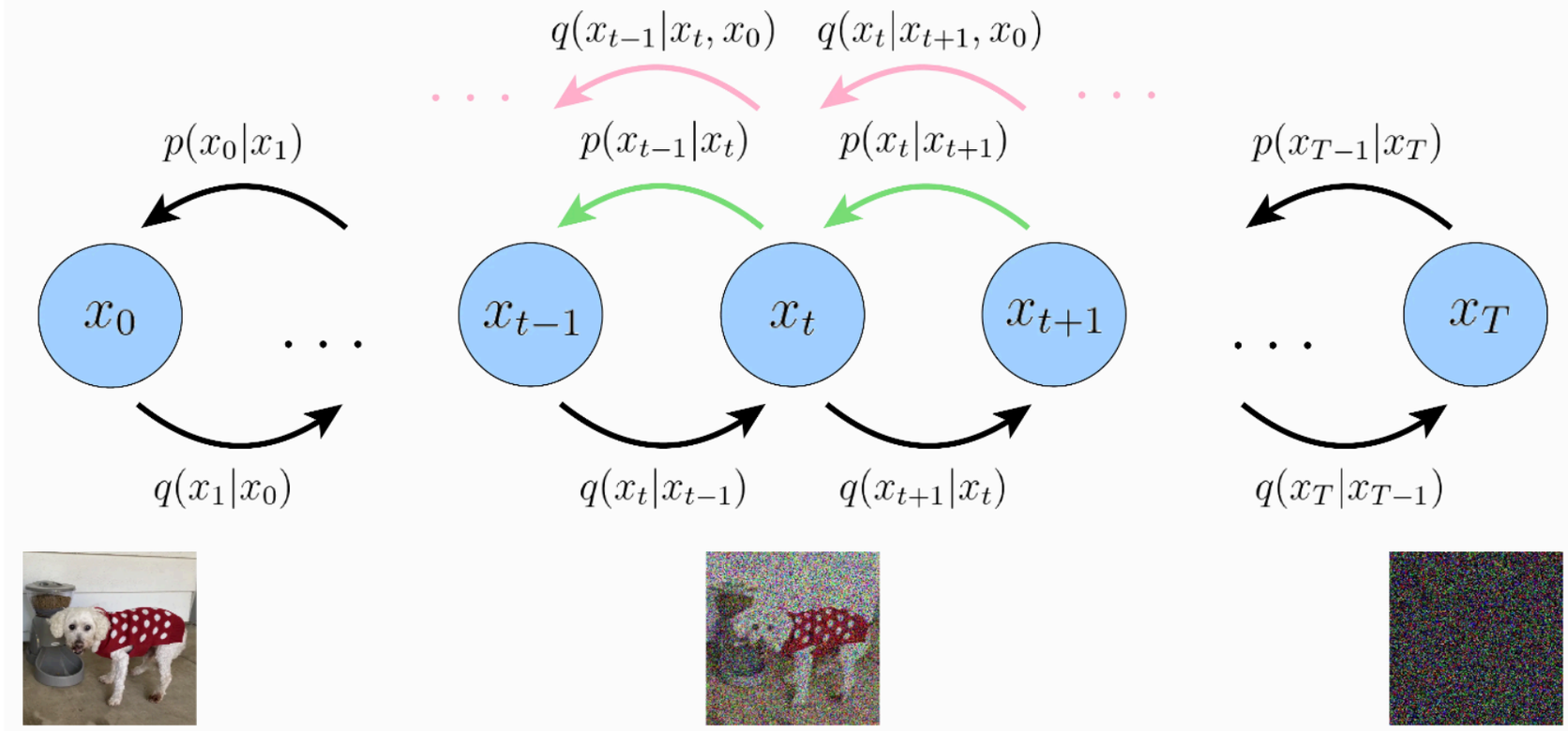
$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$

Noised distribution:

$$
\begin{aligned}
q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}) \mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})} \\
&\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})
\end{aligned}
$$

We will choose our denoising distribution to have the same covariance matrix and only parametrize its mean with a learnable function:

Denoising distribution: $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t), \Sigma_q(t))$



Key idea: the encoding process will use x_t and $x_0$to predisct x_{t-1} instead of x_{t-2} .

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right]$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_0) \,||\, p(\boldsymbol{x}_T))}_{\text{prior matching term}}$$

$$- \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

Matching the denoising and noised means

Noised: $$q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)})$$

Denoised: $$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t), \Sigma_q(t))$$

$$\sigma_q^2(t) = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$$

The KL takes between Gaussian distributions:

$$\mathcal{D}_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \,||\, \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)) = \frac{1}{2}\left[\log \frac{|\boldsymbol{\Sigma}_y|}{|\boldsymbol{\Sigma}_x|} - d + \operatorname{tr}(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_x) + (\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)^T \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)\right]$$

Since the variances match exactly, minimizing KL boils down to matching the means:

$$\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)) \\
&= \arg\min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}\left(\mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)\right) \,||\, \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_q(t)\right)\right) \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\|_2^2\right]
\end{aligned}$$

Matching the denoising and noised means

Noised: $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)})$

Denoised: $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t), \Sigma_q(t))$

We can rewrite the denoising transition mean:

$$\boldsymbol{\mu_\theta}(\boldsymbol{x}_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{1-\bar{\alpha}_t}$$

$$\arg\min_{\boldsymbol{\theta}} \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} \left[\mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)) \right]$$

$$= \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{t \sim U\{2,T\}} \left[\mathbb{E}_{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} \left[\frac{1}{2\sigma_q^2(t)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{(1-\bar{\alpha}_t)^2} \left[\left\| \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0 \right\|_2^2 \right] \right] \right]$$

Matching the denoising and noised means

Noised: $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)})$

Denoised: $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t), \Sigma_q(t))$

Remember: $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 \qquad \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}$$

Matching the denoising and noised means

Noised: $$q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)})$$

Denoised: $$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t), \Sigma_q(t))$$

Remember: $$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 \qquad \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}$$

Matching the denoising and noised means

Noised: $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)})$

Denoised: $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t), \Sigma_q(t))$

Remember: $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 \qquad \boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar{\alpha}_t}}$$

We can rewrite:

$$\begin{aligned}\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t} \\ &= \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0\end{aligned}$$

Matching the denoising and noised means

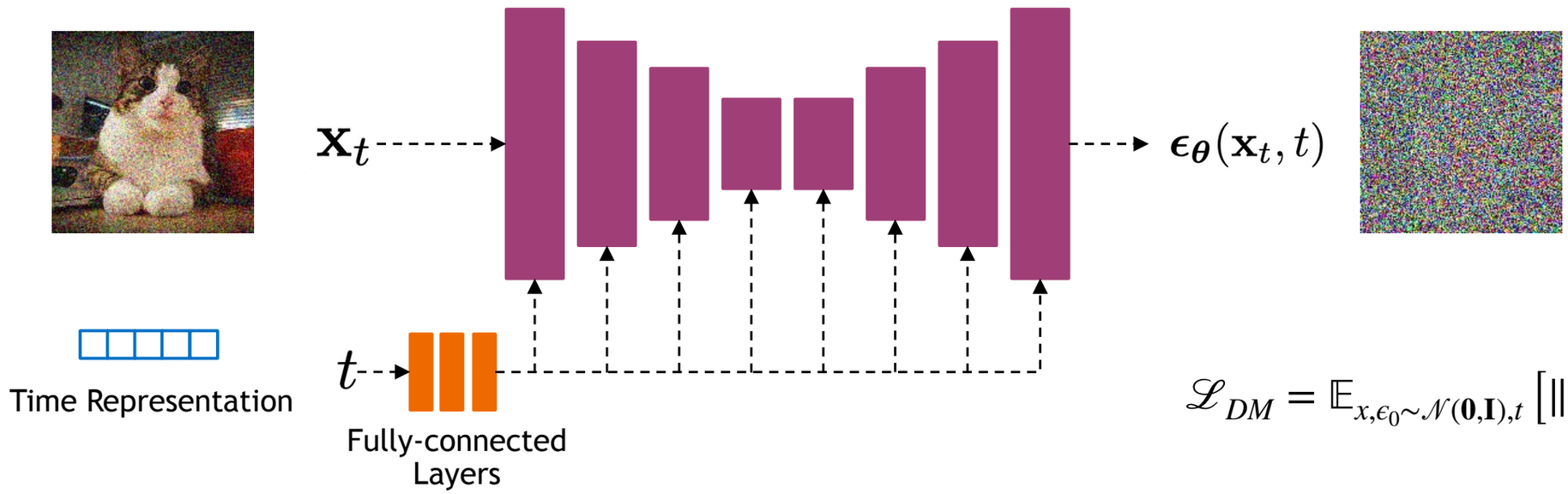
Noised: $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\Sigma_q(t)})$

Denoised: $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t), \Sigma_q(t))$

We can rewrite the denoising transition mean:

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t)$$

$$\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)) \\
&= \arg\min_{\boldsymbol{\theta}} \mathcal{D}_{\mathrm{KL}}\left(\mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)\right) \,||\, \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_q(t)\right)\right) \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t} \left[\left\| \boldsymbol{\epsilon}_0 - \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t) \right\|_2^2\right]
\end{aligned}$$



$$\mathscr{L}_{DM} = \mathbb{E}_{x,\epsilon_0 \sim \mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\epsilon_0 - \hat{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2\right]$$

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

What is this?

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

This is the estimated mean of the previous timestep: $\mu_\theta(\mathbf{x}_t)$

Conditional Diffusion Models

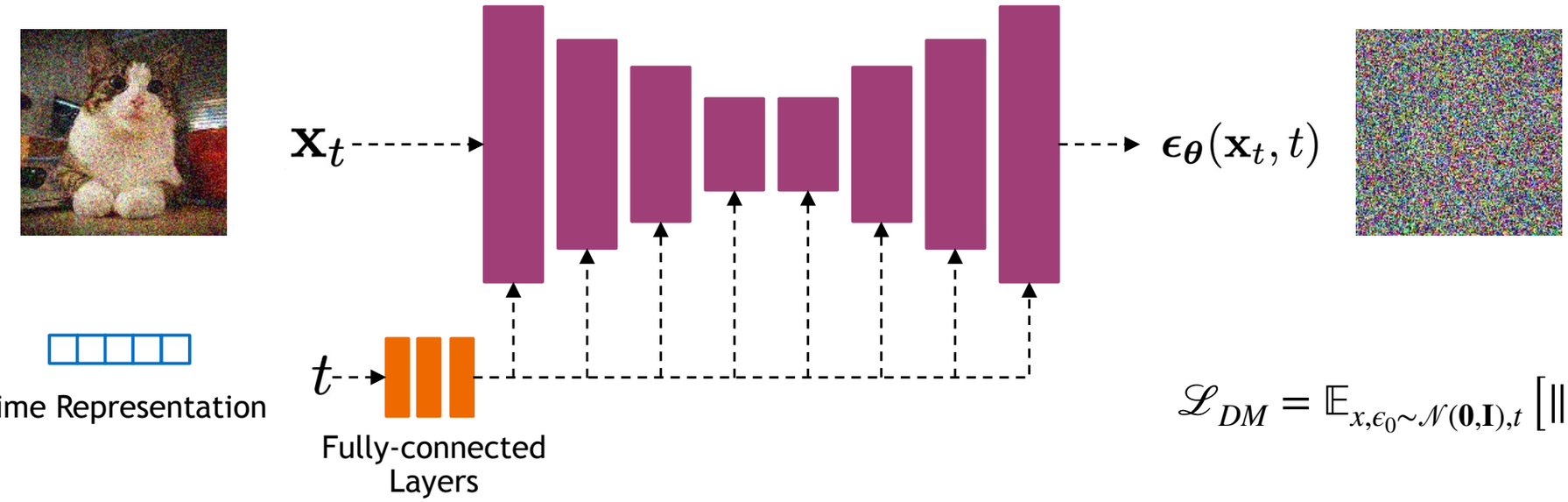
Conditional diffusion models add arbitrary conditioning information y at each transition step:

$$p(\mathbf{x}_{0:T}|y) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, y)$$

$\mathbf{x}_0$

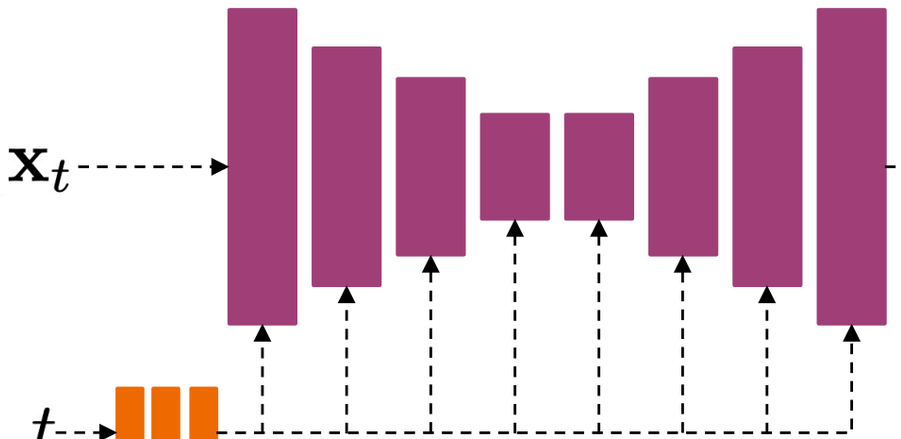
y

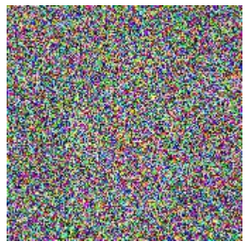
A medieval painting of the wifi not working



$$\mathscr{L}_{DM} = \mathbb{E}_{x,\epsilon_0 \sim \mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\epsilon_0 - \hat{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2\right]$$

y = "a cute cat"





Time Representation

$$\mathscr{L}_{DM} = \mathbb{E}_{x,\epsilon_0 \sim \mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\epsilon_0 - \hat{\epsilon}_\theta(\mathbf{x}_t, t, y)\|_2^2\right]$$

IMITATING HUMAN BEHAVIOUR WITH DIFFUSION MODELS

Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, Sam Devlin
Microsoft Research

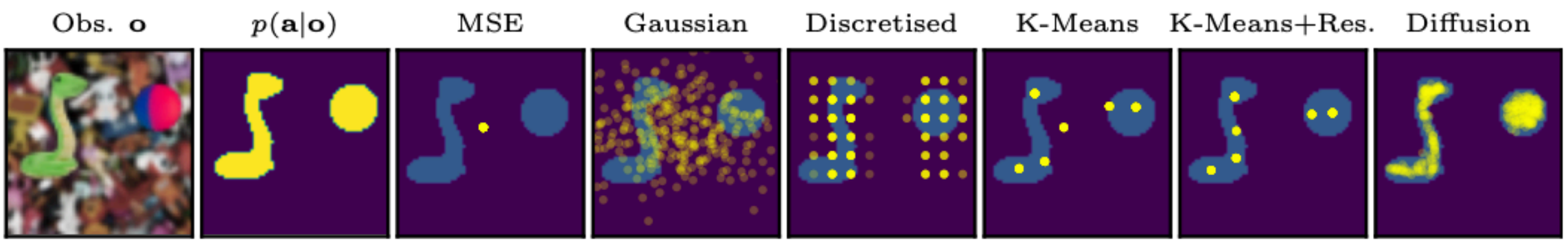
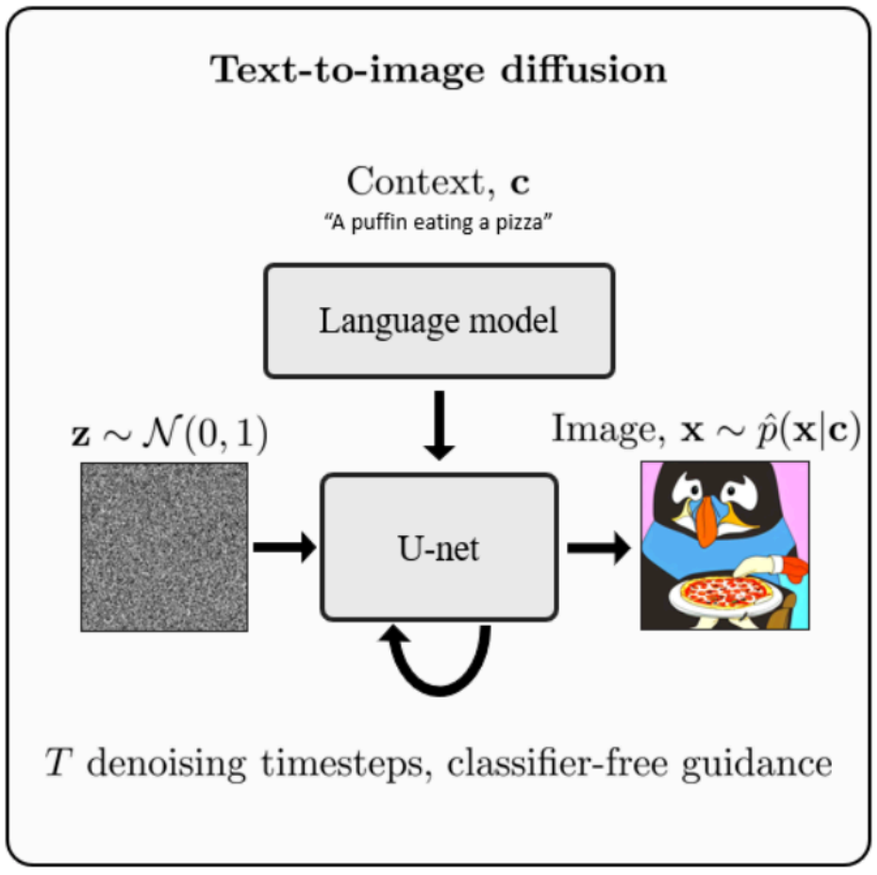
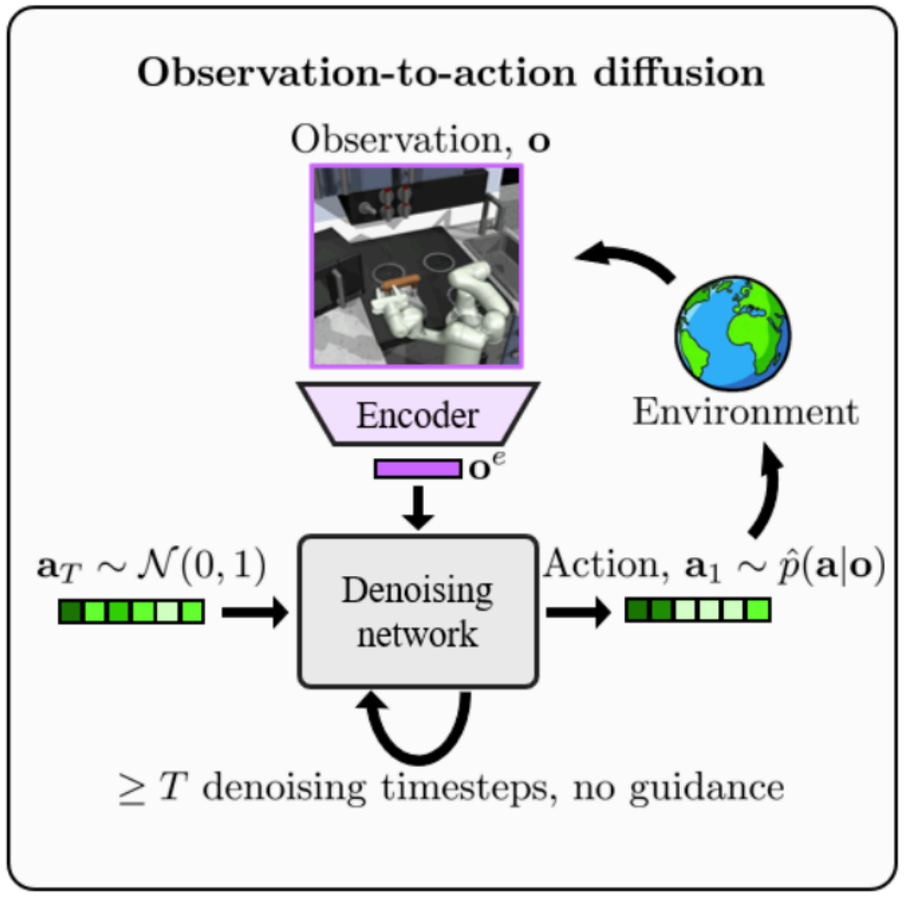


Figure 1: Expressiveness of a variety of models for behaviour cloning in a single-step, arcade claw game with two simultaneous, continuous actions. Existing methods fail to model the full action distribution, $p(\mathbf{a}|\mathbf{o})$, whilst diffusion models excel at covering multimodal & complex distributions.


Text-to-image diffusion
Context, c
"A puffin eating a pizza"
Language model
$\mathbf{z} \sim \mathcal{N}(0, 1)$
U-net
Image, $\mathbf{x} \sim \hat{p}(\mathbf{x}|\mathbf{c})$
T denoising timesteps, classifier-free guidance


Observation-to-action diffusion
Observation, o
Encoder
$\mathbf{o}^e$
Environment
$\mathbf{a}_T \sim \mathcal{N}(0, 1)$
Denoising network
Action, $\mathbf{a}_1 \sim \hat{p}(\mathbf{a}|\mathbf{o})$
$\geq T$ denoising timesteps, no guidance

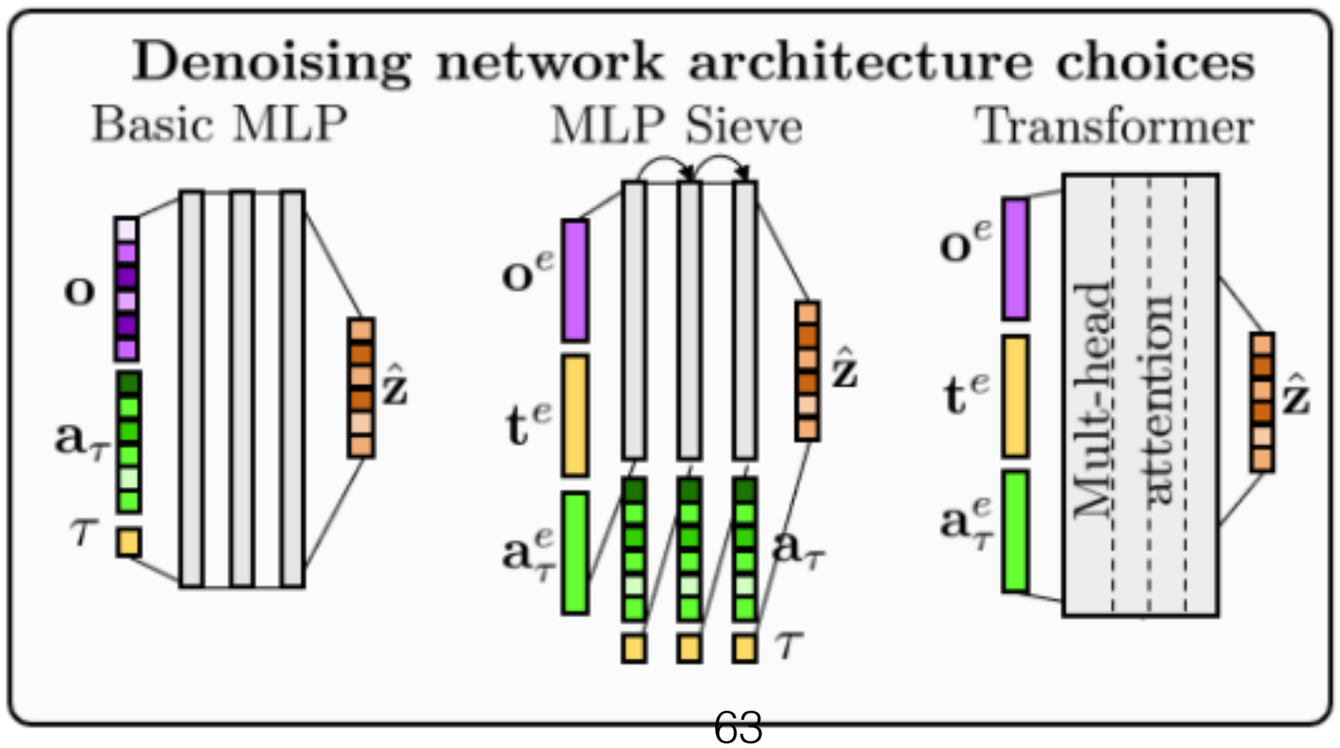
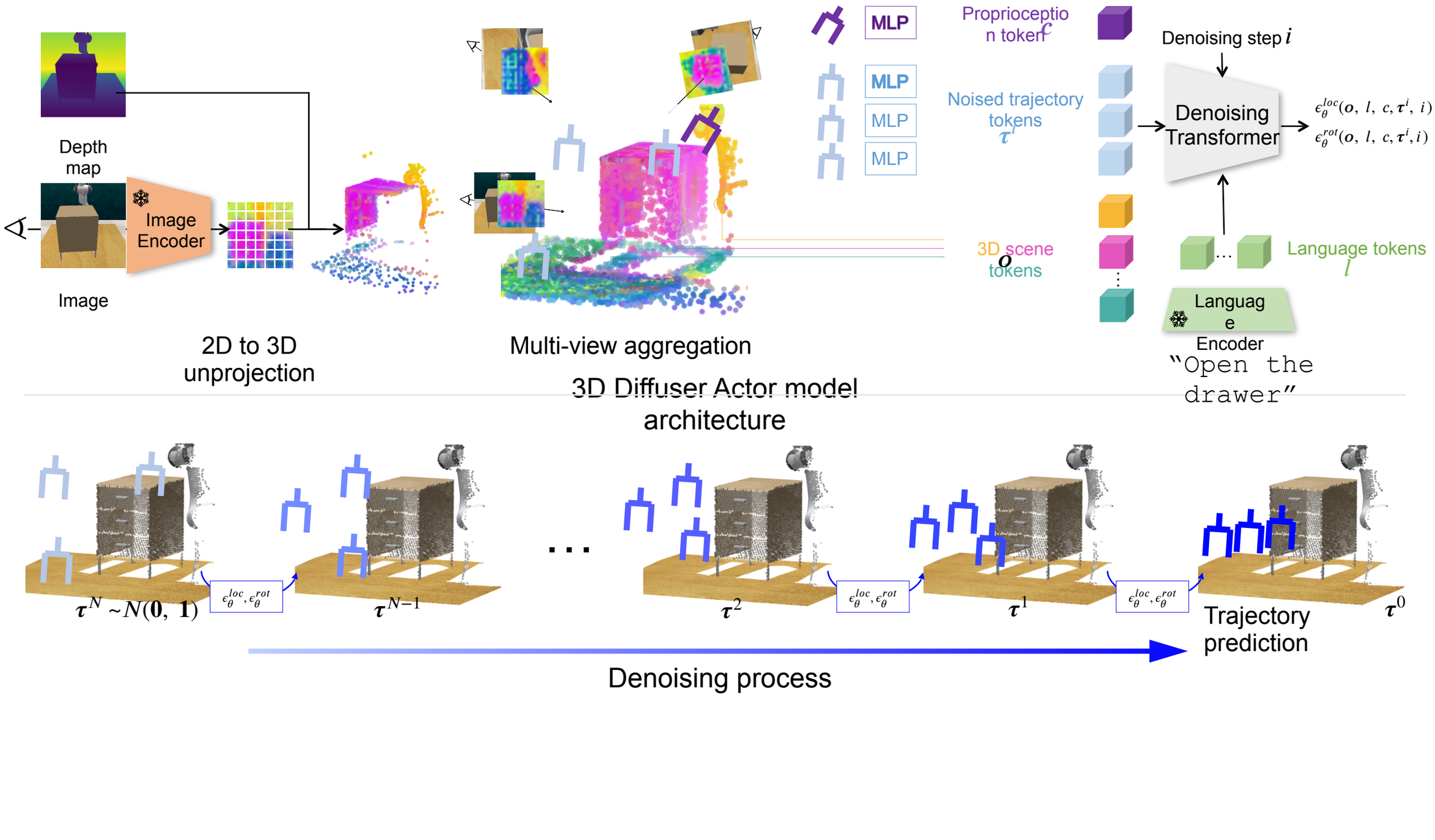

Denoising network architecture choices
Basic MLP
MLP Sieve
Transformer
$\mathbf{o}$
$\mathbf{a}_\tau$
τ
$\hat{\mathbf{z}}$
$\mathbf{o}^e$
$\mathbf{t}^e$
$\mathbf{a}^e_\tau$
Mult-head attention

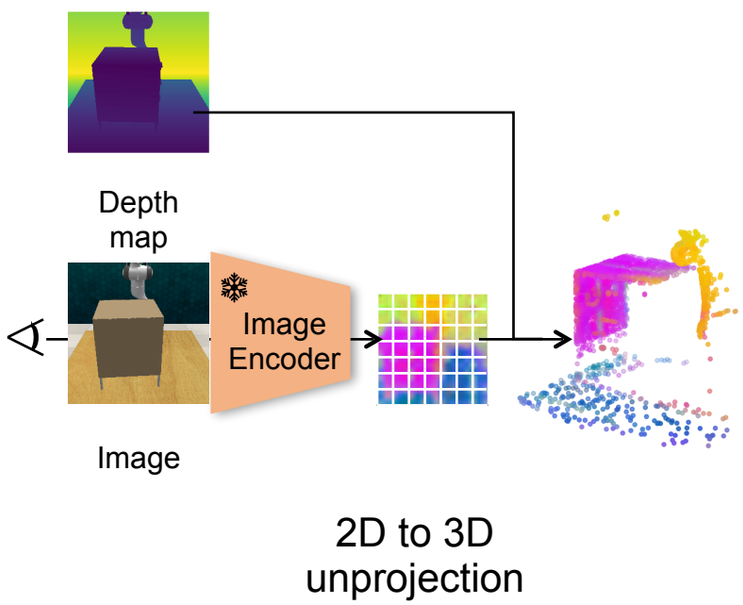
Table 1: Robotic control results. Mean ± one standard error over three training runs (100 rollouts). Methods marked with asterisk (*) are our proposed methods.

	Tasks $\geq 4 \uparrow$	Tasks Wasserstein $\downarrow$	Time Wasserstein $\downarrow$	State Wasserstein $\downarrow$	Density $\uparrow$	Coverage $\uparrow$
MLP Basic Architecture						
*Diffusion BC, Basic MLP	0.45 ± 0.03	1.96 ± 0.12	12.04 ± 2.20	0.463 ± 0.012	0.54 ± 0.02	0.38 ± 0.01
*Diffusion-KDE, Basic MLP	0.59 ± 0.01	1.72 ± 0.03	8.08 ± 0.24	0.481 ± 0.005	0.78 ± 0.00	0.37 ± 0.00
*Diffusion-X, Basic MLP	0.58 ± 0.02	1.51 ± 0.14	8.61 ± 0.14	0.424 ± 0.017	0.64 ± 0.00	0.41 ± 0.00
MLP Sieve Architecture						
MSE, MLP Sieve	0.5 ± 0.02	1.91 ± 0.07	6.40 ± 0.48	0.443 ± 0.021	0.71 ± 0.01	0.40 ± 0.01
Discretised, MLP Sieve	0.18 ± 0.02	3.43 ± 0.14	11.30 ± 1.29	0.651 ± 0.026	0.38 ± 0.02	0.31 ± 0.01
K-Means, MLP Sieve	0.0 ± 0.0	5.25 ± 0.0	–	1.469 ± 0.120	0.09 ± 0.00	0.06 ± 0.00
K-Means+Residual, MLP Sieve	0.23 ± 0.02	2.87 ± 0.16	11.60 ± 2.11	0.607 ± 0.027	0.51 ± 0.01	0.36 ± 0.00
EBM Deriv-Free, MLP Sieve	0.0	–	–	–	–	–
*Diffusion BC, MLP Sieve	0.68 ± 0.02	1.31 ± 0.05	6.06 ± 1.10	0.373 ± 0.012	0.66 ± 0.01	0.42 ± 0.00
*Diffusion-KDE, MLP Sieve	0.79 ± 0.04	1.6 ± 0.24	6.77 ± 0.64	0.439 ± 0.039	0.93 ± 0.02	0.41 ± 0.01
*Diffusion-X, MLP Sieve	0.77 ± 0.02	$\mathbf{1.06 \pm 0.05}$	5.24 ± 0.90	0.344 ± 0.004	0.78 ± 0.01	$\mathbf{0.45 \pm 0.00}$
Transformer Architecture						
MSE, Transformer	0.69 ± 0.02	1.47 ± 0.13	5.85 ± 0.27	0.397 ± 0.034	0.81 ± 0.01	0.42 ± 0.01
Discretised, Transformer	0.34 ± 0.02	2.54 ± 0.14	6.13 ± 0.49	0.512 ± 0.002	0.47 ± 0.01	0.36 ± 0.00
K-Means, Transformer	0.0	5.25	–	1.470	0.07	0.06
K-Means+Residual, Transformer	0.34 ± 0.02	2.25 ± 0.16	7.80 ± 0.87	0.426 ± 0.018	0.66 ± 0.02	0.38 ± 0.01
*Diffusion BC, Transformer	0.77 ± 0.01	1.35 ± 0.11	$\mathbf{4.11 \pm 0.05}$	$\mathbf{0.340 \pm 0.003}$	0.74 ± 0.01	0.44 ± 0.00
*Diffusion-KDE, Transformer	$\mathbf{0.89 \pm 0.01}$	1.31 ± 0.03	5.28 ± 0.41	0.418 ± 0.012	$\mathbf{0.97 \pm 0.02}$	0.43 ± 0.01
*Diffusion-X, Transformer	0.88 ± 0.01	1.17 ± 0.13	4.65 ± 0.47	0.365 ± 0.013	0.94 ± 0.02	$\mathbf{0.45 \pm 0.01}$
From Shafiullah et al. (2022)						
Behaviour Transformers	0.44					
Implicit BC	0.24					
SPiRL VAE	0.0					
PARROT Normalizing Flow	0.0					
Dataset						
Human	0.98					
Human sub-sampled	–	–	–	0.223 ± 0.006	1.00 ± 0.01	0.56 ± 0.01

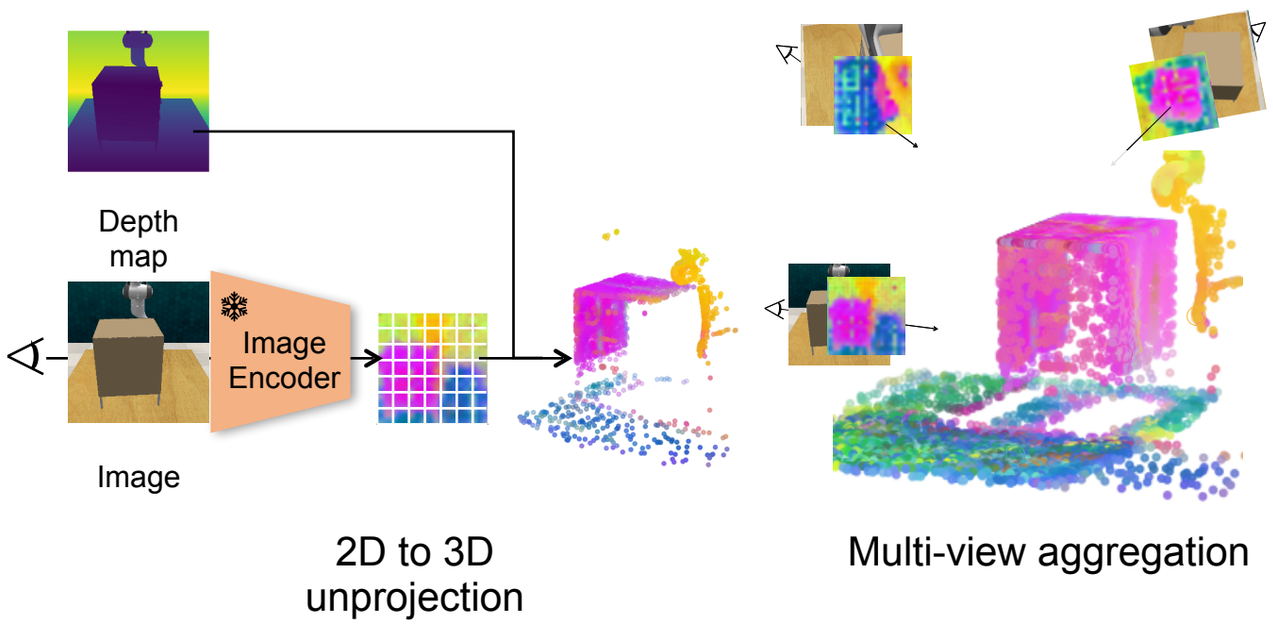
3D DiffuserActor: 3D scene representations + action diffusion



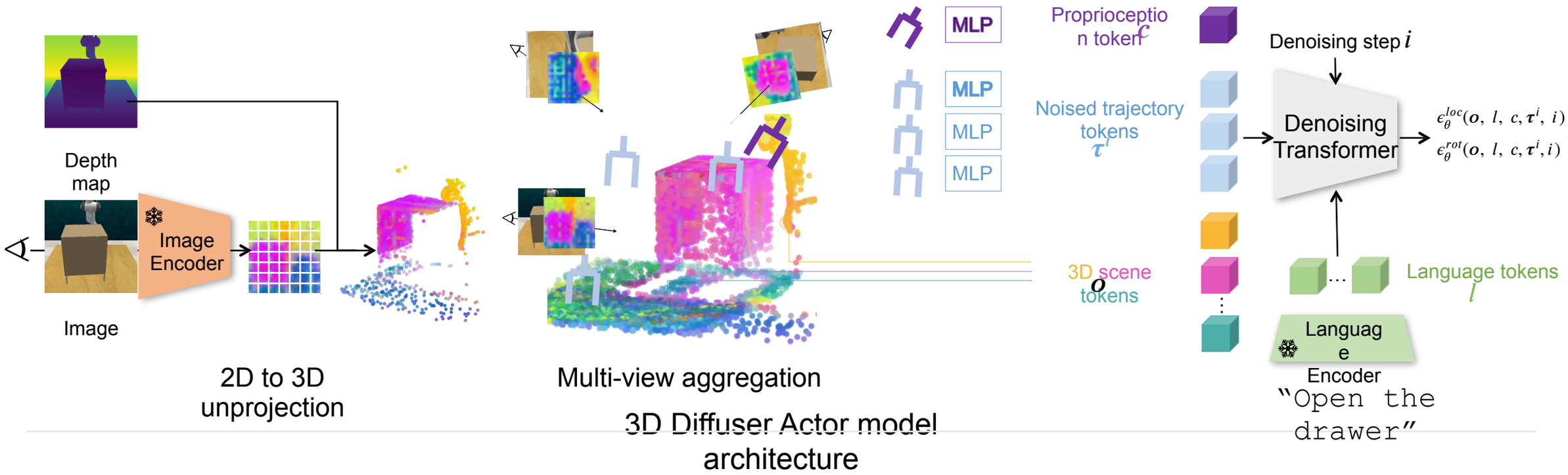
3D DiffuserActor: 3D scene representations + action diffusion



3D DiffuserActor: 3D scene representations + action diffusion

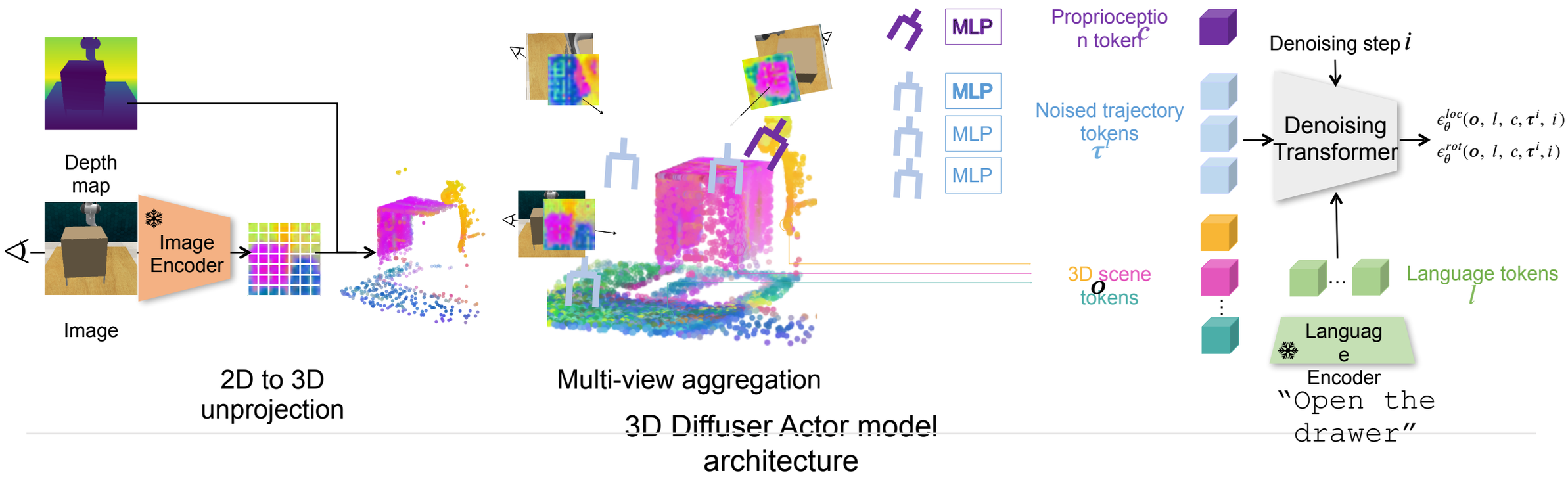


3D DiffuserActor: 3D scene representations + action diffusion



3D Diffuser Actor model architecture

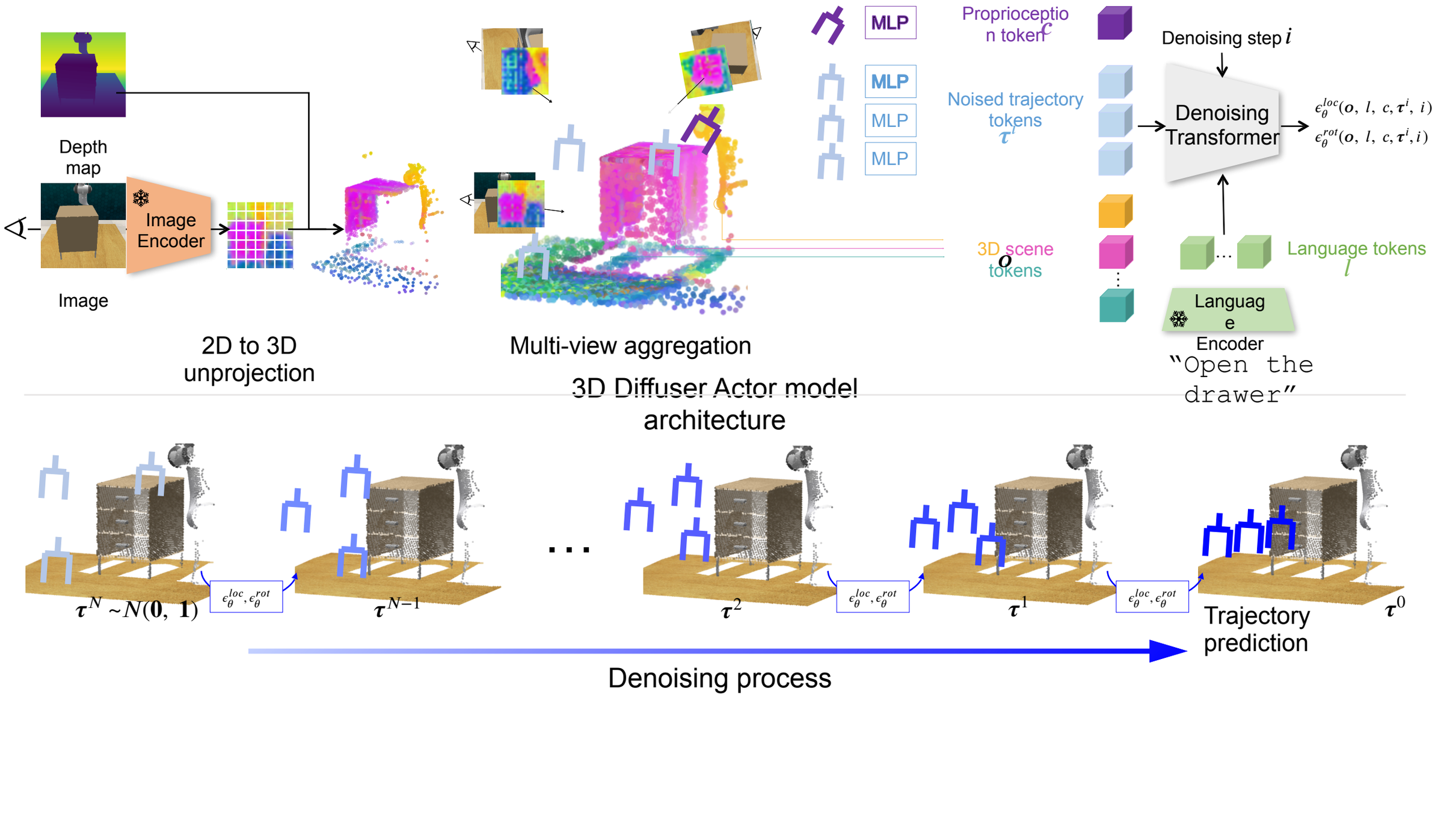
3D DiffuserActor: 3D scene representations + action diffusion



3D Diffuser Actor model architecture

$$\text{RelativeAttention}(\mathbf{q}_i, \mathbf{k}_j, \mathbf{v}_j) = \frac{\exp e_{i,j}}{\sum_l \exp e_{i,l}} \mathbf{v}_j \quad \text{where } e_{i,j} = \mathbf{q}_i^T \mathbf{M}(\mathbf{p}_j - \mathbf{p}_i)\mathbf{k}_j$$

3D DiffuserActor: 3D scene representations + action diffusion

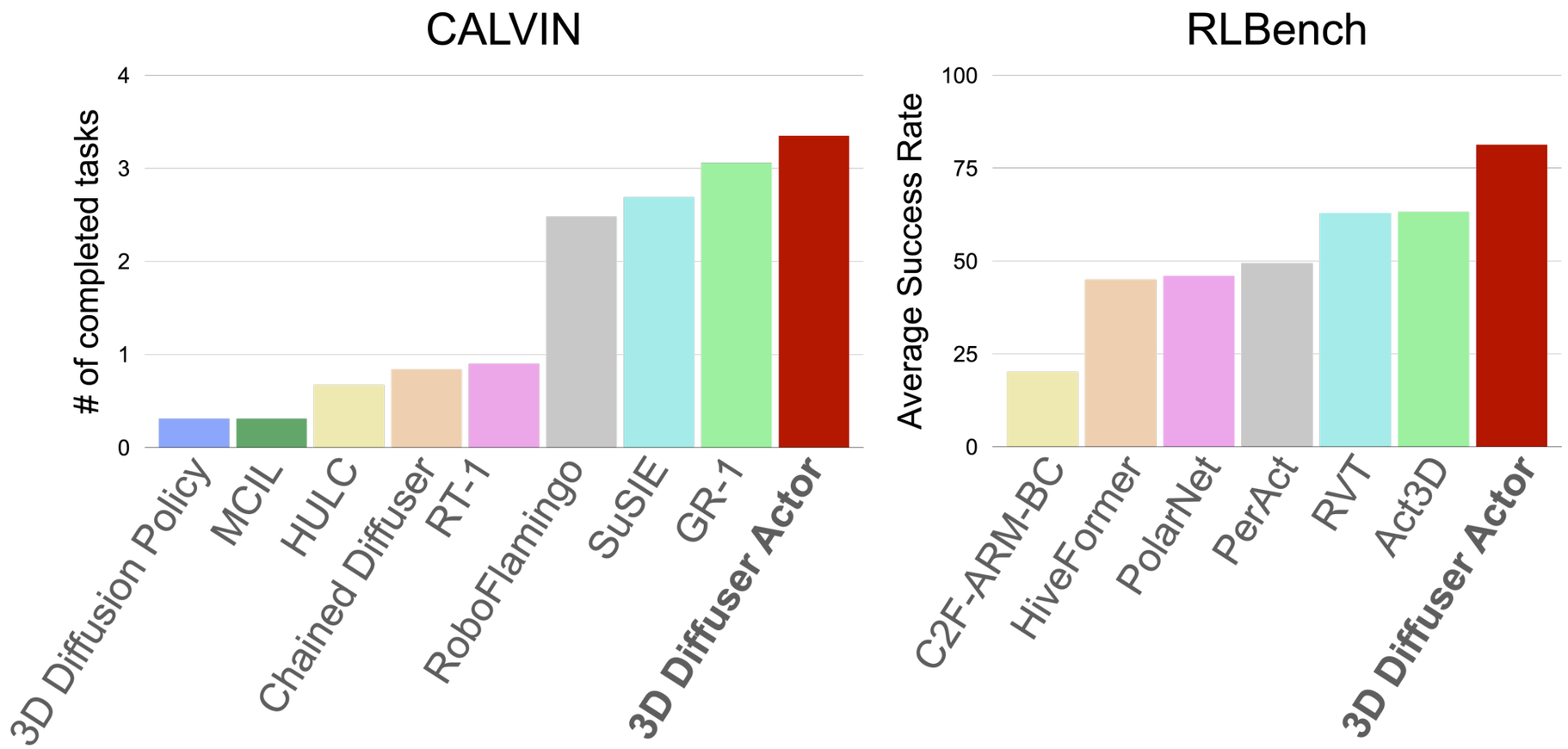


Trajectory samples

Keypose samples

"put grapes in the bowl"

3D Diffuser Actor sets new a SOTA on RLBench and CALVIN



		Avg. Success ↑	Avg. Rank ↓	open drawer	slide block	sweep to dustpan	meat off grill	turn tap	put in drawer	close jar	drag stick
	PerAct [57]	49	4.5	88	74	52	70	88	51	55	90
	RVT [18]	63	3.5	71	82	72	88	94	88	52	99
	Act3D [15]	65	3.4	93	93	92	94	94	90	92	92
Ablations	2D Diffuser Actor	44	5.9	68	65	90	75	72	55	94	79
Ablations	3D Diffuser Actor w/o Rel. Attn.	68	2.6	96	**99**	87	98	97	91	94	98
Ablations	3D Diffuser Actor (ours)	**78** (+13)	**1.3**	**97** (+4)	**94** (+1)	**99** (+7)	95 (+1)	**99** (+3)	**92** (+2)	97 (+5)	97 (+5)

		stack blocks	screw bulb	put in safe	place wine	put in cupboard	sort shape	push buttons	insert peg	stack cups	place cups
	PerAct [57]	26	18	86	45	28	17	93	6	2	2
	RVT [18]	29	48	91	91	50	36	100	11	26	4
	Act3D [15]	12	47	95	80	51	8	99	27	9	3
Ablations	2D Diffuser Actor	1	48	57	68	3	0	9	6	0	0
Ablations	3D Diffuser Actor w/o Rel. Attn.	35	48	97	94	64	7	98	6	12	0
Ablations	3D Diffuser Actor (ours)	**72** (+60)	**89** (+42)	**100** (+5)	**95** (+15)	**69** (+18)	**13** (+5)	**100** (+1)	**32** (+5)	**55** (+46)	**16** (+13)

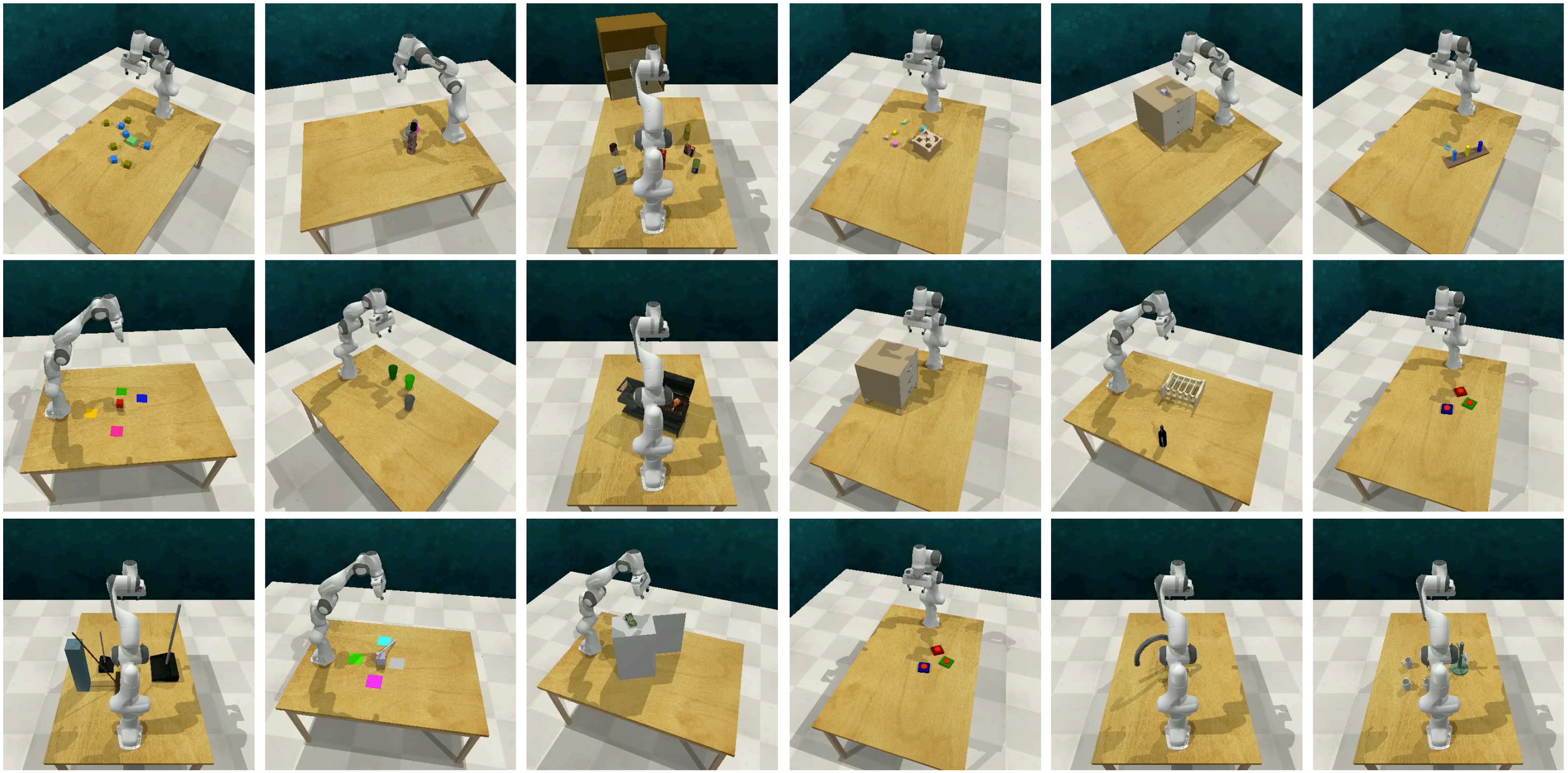
	Avg. Success ↑	Avg. Rank ↓	open drawer	slide block	sweep to dustpan	meat off grill	turn tap	put in drawer	close jar	drag stick
PerAct [57]	49	4.5	88	74	52	70	88	51	55	90
RVT [18]	63	3.5	71	82	72	88	94	88	52	99
Act3D [15]	65	3.4	93	93	92	94	94	90	92	92
Ablations: 2D Diffuser Actor	44	5.9	68	65	90	75	72	55	94	79
Ablations: 3D Diffuser Actor w/o Rel. Attn.	68	2.6	96	**99**	87	98	97	91	94	98
Ablations: 3D Diffuser Actor (ours)	**78** (+13)	**1.3**	**97** (+4)	**94** (+1)	**99** (+7)	95 (+1)	**99** (+3)	**92** (+2)	97 (+5)	97 (+5)

	stack blocks	screw bulb	put in safe	place wine	put in cupboard	sort shape	push buttons	insert peg	stack cups	place cups
PerAct [57]	26	18	86	45	28	17	93	6	2	2
RVT [18]	29	48	91	91	50	36	100	11	26	4
Act3D [15]	12	47	95	80	51	8	99	27	9	3
Ablations: 2D Diffuser Actor	1	48	57	68	3	0	9	6	0	0
Ablations: 3D Diffuser Actor w/o Rel. Attn.	35	48	97	94	64	7	98	6	12	0
Ablations: 3D Diffuser Actor (ours)	**72** (+60)	**89** (+42)	**100** (+5)	**95** (+15)	**69** (+18)	**13** (+5)	**100** (+1)	**32** (+5)	**55** (+46)	**16** (+13)

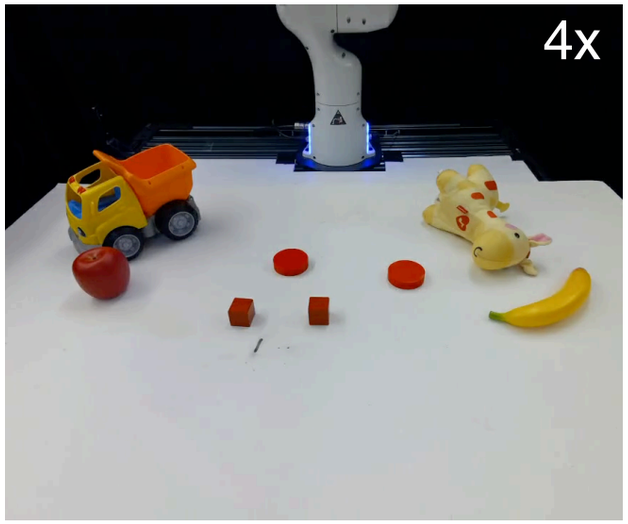
Single camera

	Avg. Success.	close jar	open drawer	sweep to dustpan	meat off grill	turn tap	slide block	put in drawer	drag stick	push buttons	stack blocks
GNFactor [67]	31.7	25.3	76.0	28.0	57.3	50.7	20.0	0.0	37.3	18.7	4.0
Act3D [15]	65.3	52.0	84.0	80.0	66.7	64.0	100.0	54.7	86.7	64.0	0.0
3D Diffuser Actor	78.4 (+13.1)	82.7 (**+29.3**)	89.3 (**+5.3**)	94.7 (**+14.7**)	88.0 (**+21.3**)	80.0 (**+16.0**)	92.0 (**-8.0**)	77.3 (**+22.6**)	98.7 (**+12.0**)	69.3 (**+5.3**)	12.0 (**+8.0**

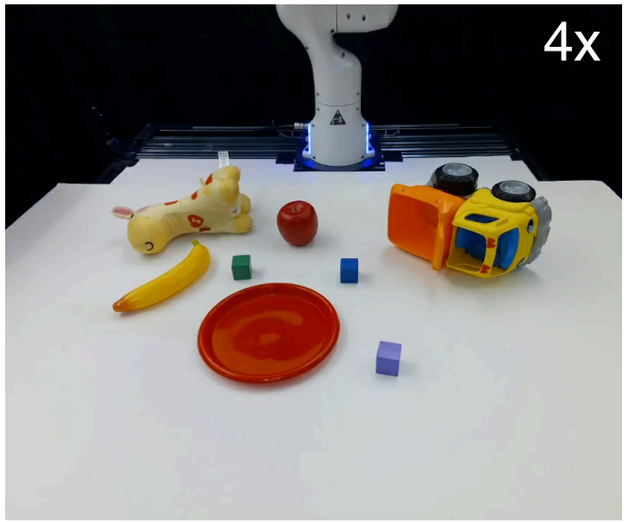
RLBench



Visualization results in the real world

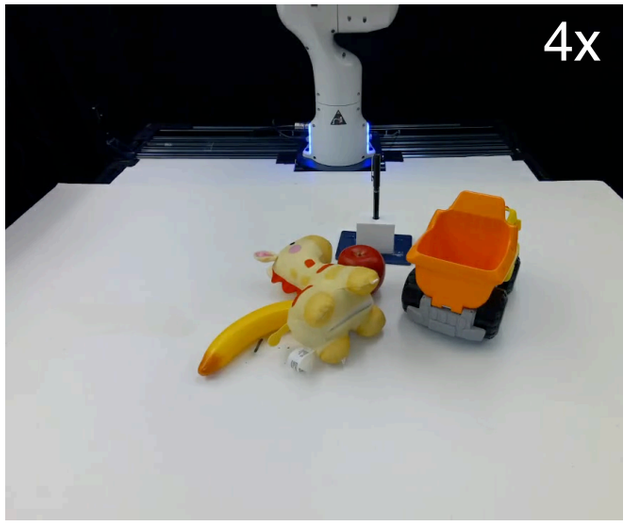


"stack blocks with the same shape"

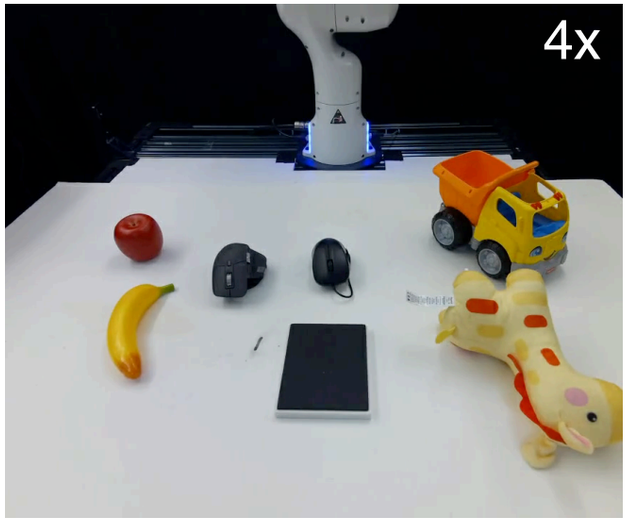


"put block in a triangle on the plate"

"put grapes in the bowl"



"open the pen"



"put a computer mouse on the pad"

"insert a peg horizontally into the torus"

"insert a peg vertically into the hole"

"close a box"