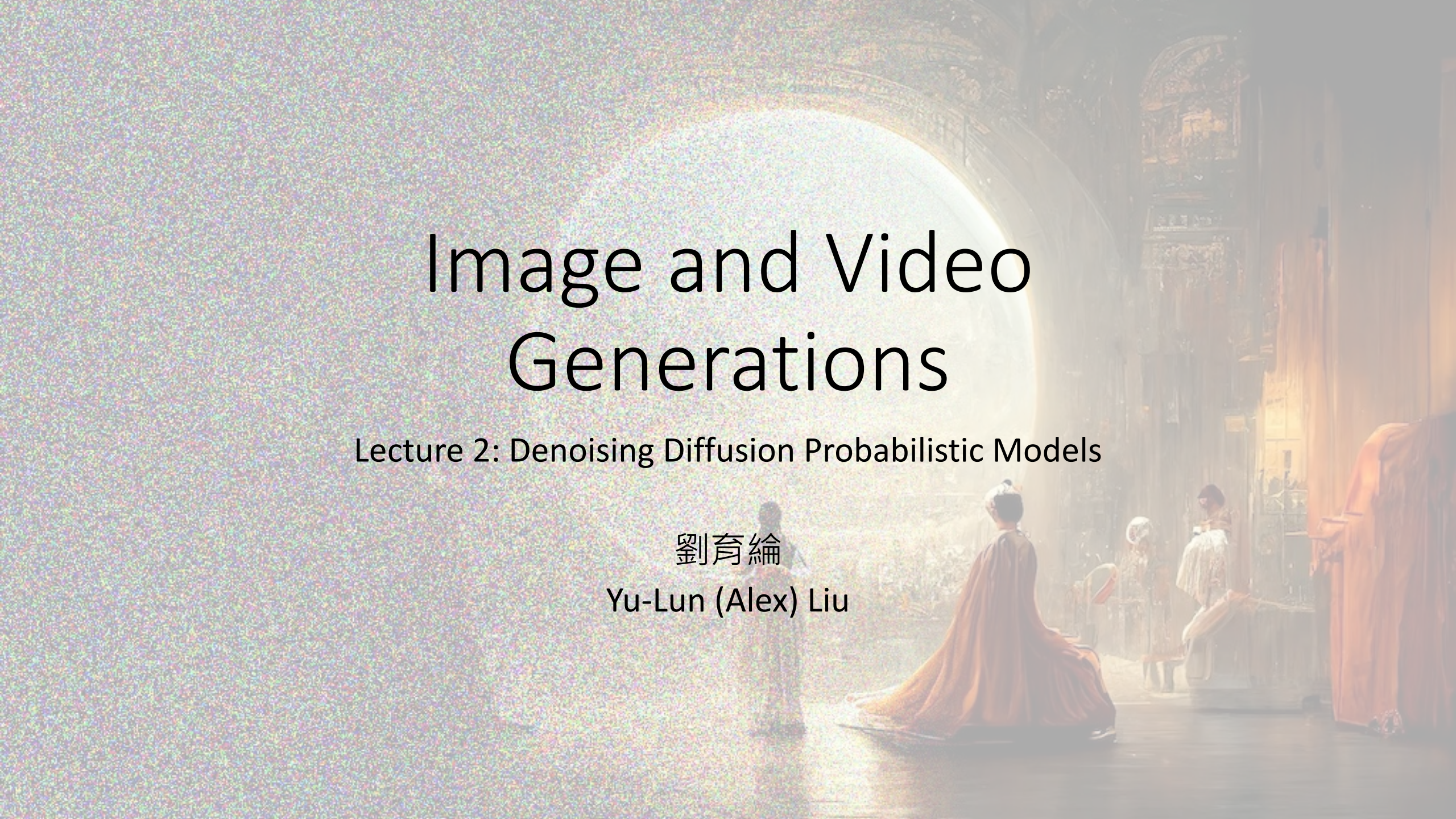
Image and Video Generations

Lecture 2: Denoising Diffusion Probabilistic Models

劉育綸

Yu-Lun (Alex) Liu



Week	Date	Topic	Assignments
1	2025-09-04		
2	2025-09-11	GAN / VAE	
3	2025-09-18	DDPM	#1 – DDPM
4	2025-09-25	DDIM	
5	2025-10-02	CFG / Latent Diffusion / ControlNet / LoRA / Zero-Shot Applications **(attending ICCV 2025)**	#2 – DDIM & LoRA
6	2025-10-09	DDIM Inversion / Score Distillation	
7	2025-10-16	Diffusion Synchronization / Inverse Problems	#3 – Distillation
8	2025-10-23	Probability Flow ODE / DPM-Solver **(attending ICCV 2025)**	
9	2025-10-30	Flow Matching	#4 – Flow Matching
10	2025-11-06	Final Project Proposal	
11	2025-11-13		
12	2025-11-20		
13	2025-11-27	Paper Presentation	
14	2025-12-04	Guest Lecture – Ta-Ying (Tim) Cheng	
15	2025-12-11	Guest Lecture – Chieh (Hubert) Lin	
16	2025-12-18	Guest Lecture – Chih-Hao Lin	
17	2025-12-25	Final Project Presentation	

Denoising Diffusion Probabilistic Models (DDPM)

Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020.

Denoising Diffusion Probabilistic Models

Consider a special case of the Markovian hierarchical VAEs where:

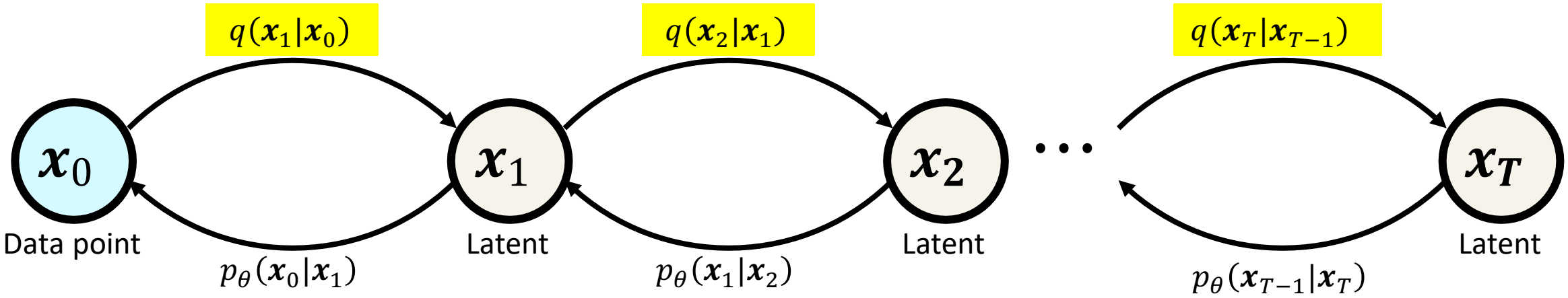
- the latent dimension is the same as the data dimension, and
- the variational posteriors $q_\phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ are not learned but predefined:

$$q_\phi(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \rightarrow q(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$$

Terminology

Forward process (predefined):

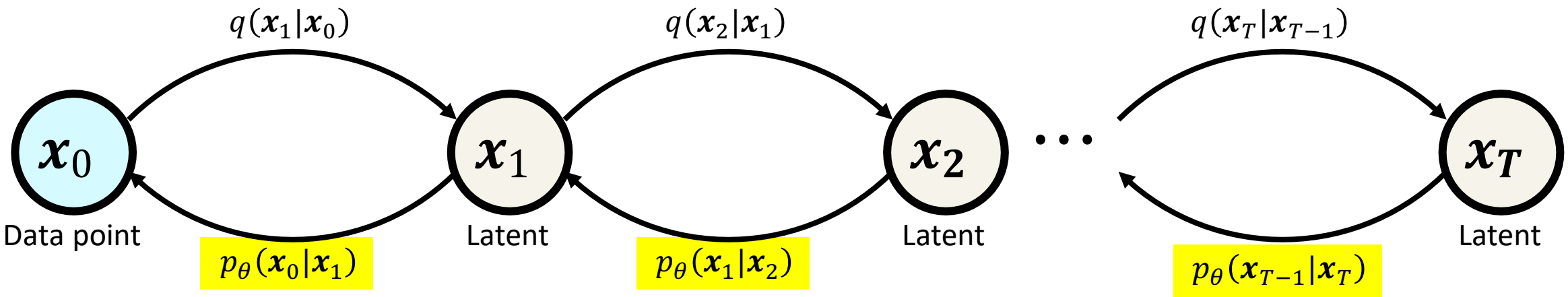
$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$



Terminology

Reverse process (learned):

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



Forward Process (Data → Latent)　正向加噪

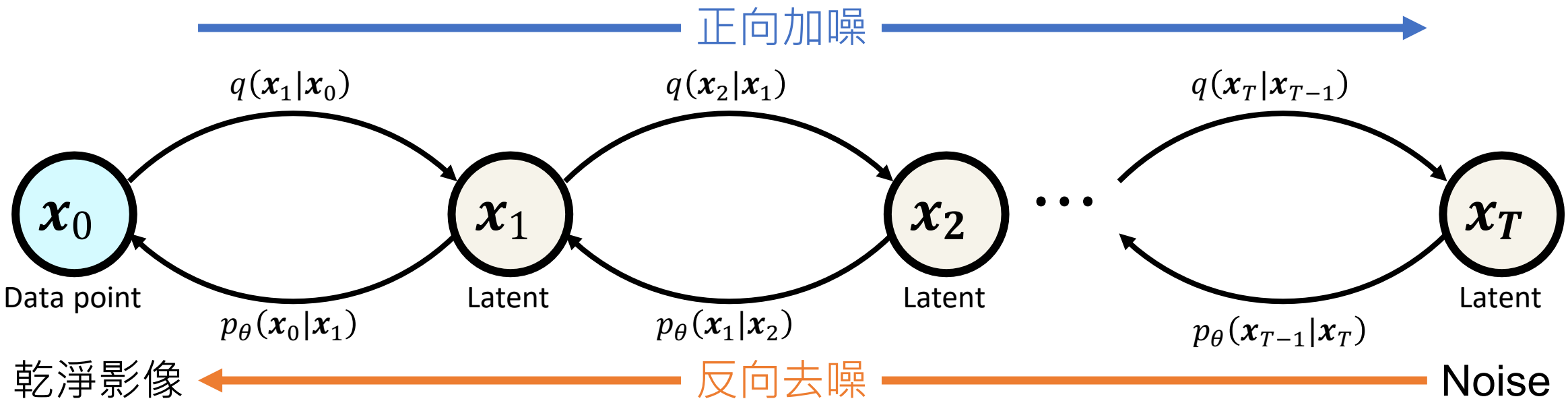
In the forward process, the transition distribution $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is specifically predefined as:

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

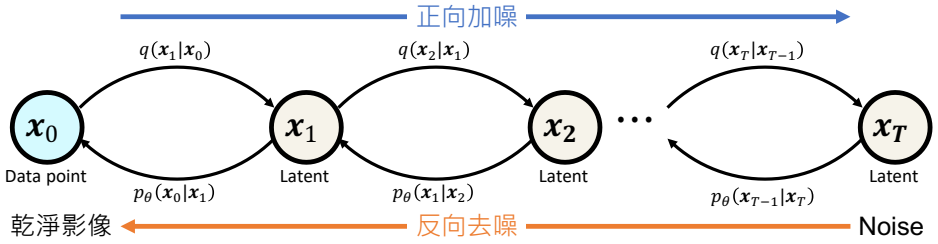
Time-dependent parameter

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_T$.

"Adding Gaussian noise iteratively!"


正向加噪
$q(\boldsymbol{x}_1|\boldsymbol{x}_0)$
$q(\boldsymbol{x}_2|\boldsymbol{x}_1)$
$q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})$
$\boldsymbol{x}_0$
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\cdots$
$\boldsymbol{x}_T$
Data point
Latent
Latent
Latent
$p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)$
$p_\theta(\boldsymbol{x}_1|\boldsymbol{x}_2)$
$p_\theta(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)$
乾淨影像
反向去噪
Noise

VP-SDE vs. VE-SDE



- $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}\right)$,
  is called Variance Preserving (VP) form.

- There are other options. For example:

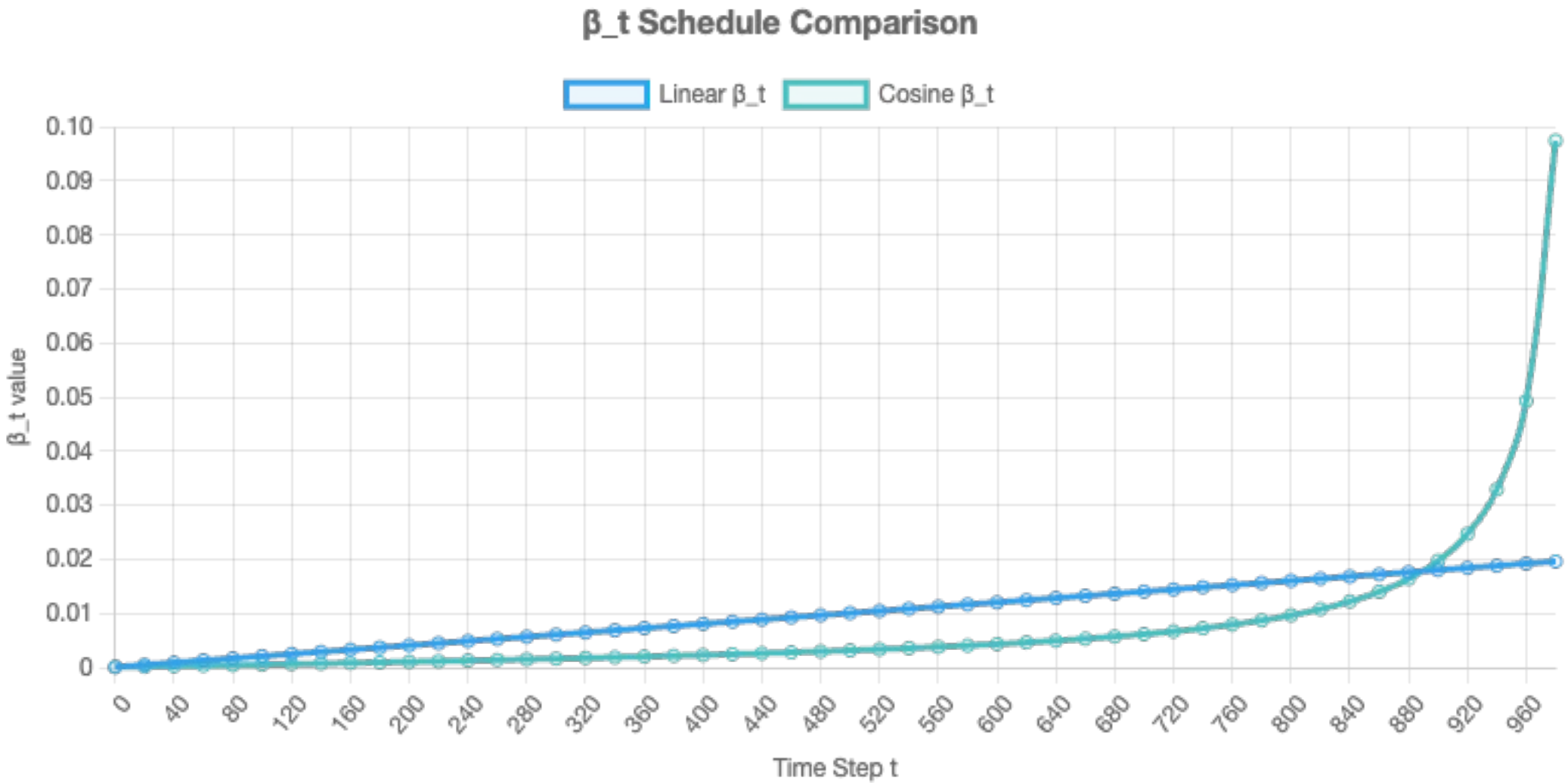
$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{x}_{t-1}, \left(\sigma_i^2 - \sigma_{i-1}^2\right)\mathbf{I}\right),$$

  which is called Variance Exploding form.

Choice of β_t

- Learned.
- Constant.
- Linearly or quadratically increased.
- Follows a cosine function
  (Nichol and Dhariwal, Improved Denoising Diffusion Probabilistic Models, ICML 2021).

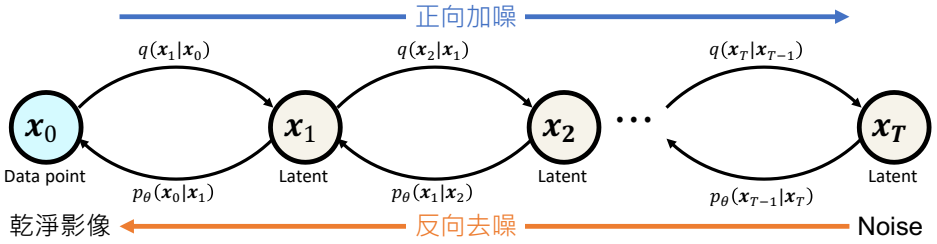
Choice of β_t



How to maximize ELBO in this case?

Disclaimer: We'll skip some complicated equations in the following slides.

ELBO



How to minimize the *negative* ELBO in this case?

$$-\log p(\boldsymbol{x}_0) = -\log \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[-\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

$$= \cdots =$$

$$-\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$$

$$+\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})\|p(\boldsymbol{x}_T))]$$

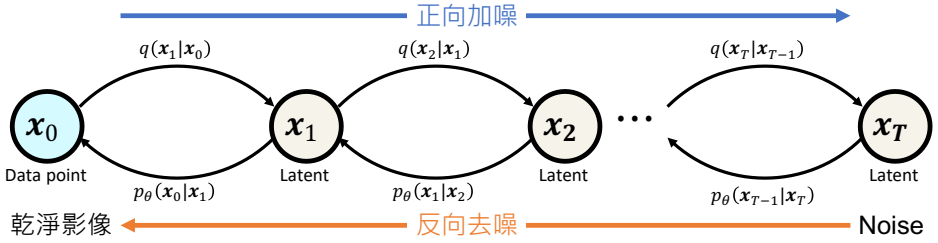
$$+\sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\|p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]$$

ELBO: Complex Derivations

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} && (34)\\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} && (35)\\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] && (36)\\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] && (37)\\
&= \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}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (38)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (39)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=1}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (40)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\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|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (41)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\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|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (42)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\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|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (43)\\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{T-1},\boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\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|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (44)\\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T)) \right]}_{\text{prior matching term}} && (45)\\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})) \right]}_{\text{consistency term}}
\end{aligned}
$$

https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
https://zhuanlan.zhihu.com/p/558937247

ELBO



How to minimize the *negative* ELBO in this case?

$$-\log p(\boldsymbol{x}_0) = -\log \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[-\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

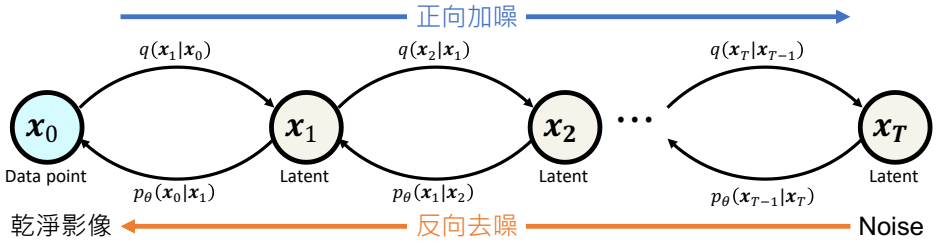
$$= \cdots =$$

$$-\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$$ Reconstruction term

$$+\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})\|p(\boldsymbol{x}_T))]$$ Prior matching term

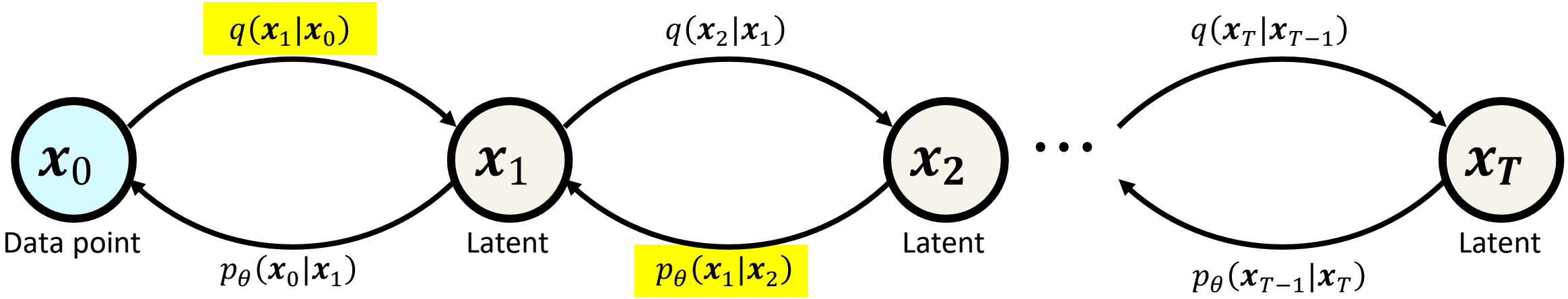
$$+\sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})\|p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]$$ **Consistency term**

Consistency Term



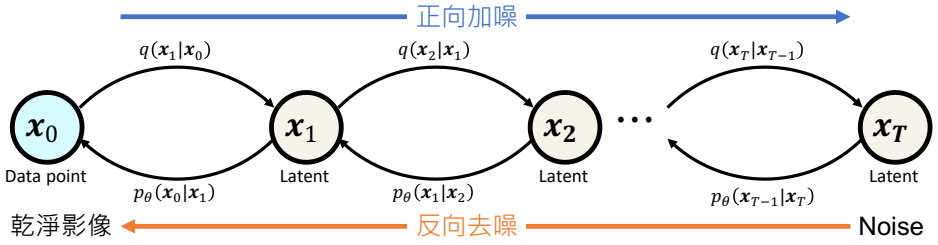
$$\sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \| p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))]$$

Make the forward and reverse steps be consistent at each time step.



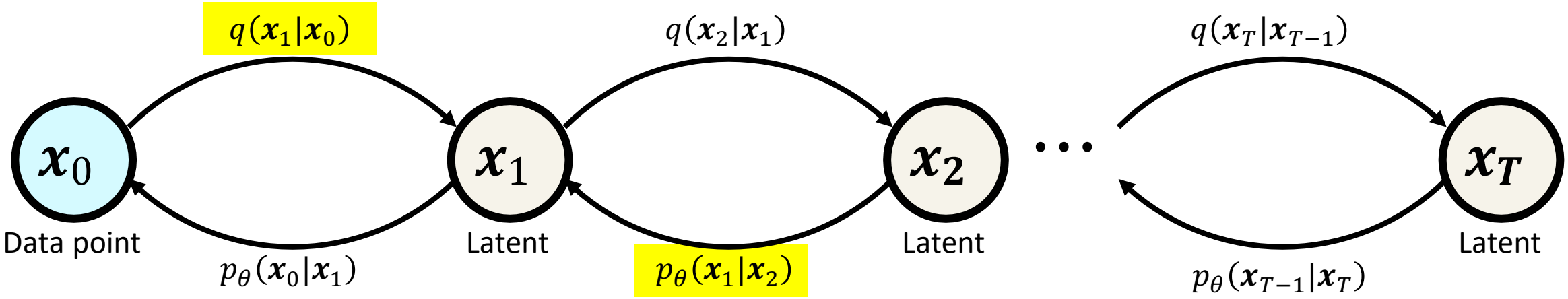
這兩個到達 $\boldsymbol{x}_t$ 的路徑要一致

Consistency Term



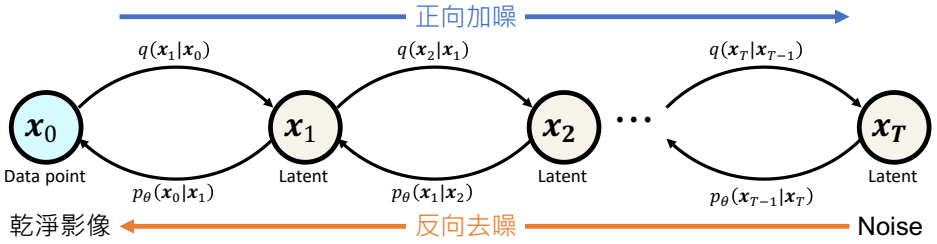
$$\sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x_{t-1}, x_{t+1}}|\boldsymbol{x_0})}[D_{KL}(q(\boldsymbol{x_t}|\boldsymbol{x_{t-1}})\|p_\theta(\boldsymbol{x_t}|\boldsymbol{x_{t+1}}))]$$

Expectation over two random variables; computationally expensive.



這兩個到達 x_t 的路徑要一致

ELBO



Can we avoid having two random variables in an expectation?

Let's re-decompose the ELBO using the fact that

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0).$$

Q. Why $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)$?

馬可夫過程的條件機率僅僅與系統的當前狀態相關，而與它的過去歷史或未來狀態，都是獨立、不相關的

ELBO: Complex Derivations

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)} \tag{46}$$

https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
https://zhuanlan.zhihu.com/p/558937247

ELBO: Complex Derivations

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \tag{47}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{48}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0) \prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \tag{49}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0) \prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)} \right] \tag{50}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)} \right] \tag{51}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}} \right] \tag{52}$$

https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
https://zhuanlan.zhihu.com/p/558937247

ELBO: Complex Derivations

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \cancel{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}}{\cancel{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}} \right] \tag{53}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}} + \log \frac{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right] \tag{54}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \sum_{t=2}^{T} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right] \tag{55}$$

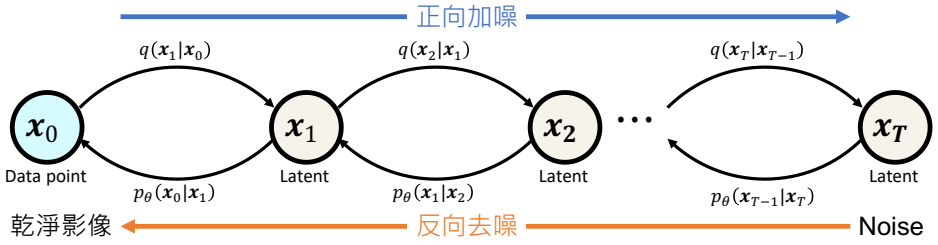
$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right] \tag{56}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} \right] + \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)} \right] \tag{57}$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \right]}_{\text{denoising matching term}} \tag{58}$$

https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
https://zhuanlan.zhihu.com/p/558937247

ELBO



Decompose the negative ELBO in a different way:

$$-\log p(\boldsymbol{x}_0) = -\log \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[-\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

$$= \cdots =$$

$$-\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$$

Reconstruction term $\mathcal{L}_0$

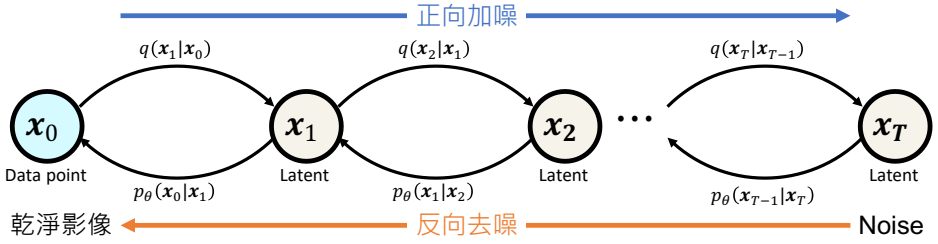
$$+D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$$

New prior matching term $\mathcal{L}_T$

$$+\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

Denoising matching term $\mathcal{L}_{t-1}$

Reconstruction Term $\mathcal{L}_0$

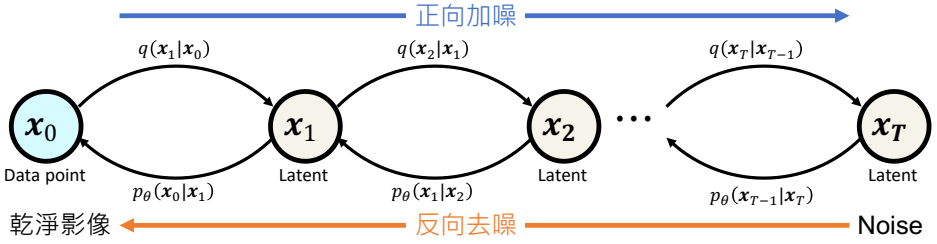


$$-\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$$

下標表示的是對哪個隨機變數取期望值，以及這個隨機變數的分佈
期望值下標 $q(\boldsymbol{x}_1|\boldsymbol{x}_0)$ 代表這個 loss 是對所有可能的 noise $\boldsymbol{x}_1$ 取平均
而 $\boldsymbol{x}_1$的分佈由 forward process $q(\boldsymbol{x}_1|\boldsymbol{x}_0)$ 決定

The same loss term in VAE, but applied only to the final reverse step.

ELBO



Decompose the negative ELBO in a different way:

$$-\log p(\boldsymbol{x}_0) = -\log \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[-\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

$$= \cdots =$$

$$-\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$$

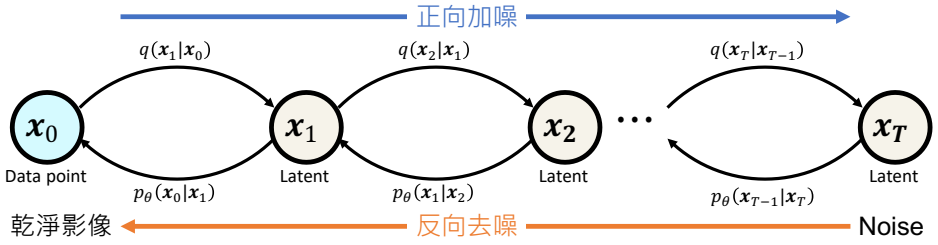
$$+D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$$

New prior matching term $\mathcal{L}_T$

$$+\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

Prior Matching Term $\mathcal{L}_T$

Prior Matching Term $\mathcal{L}_T$



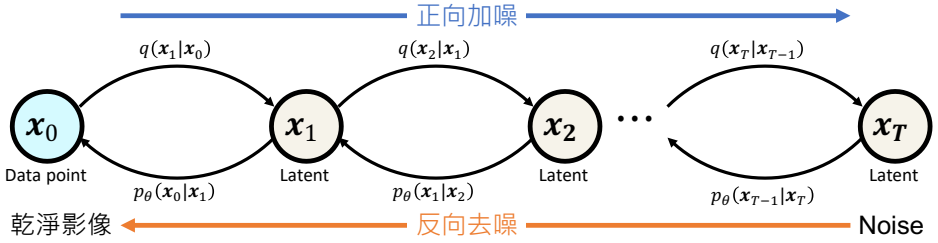
$$D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$$

- Identical to the KL divergence term in VAE.

- Note that there is nothing to be optimized; $q(\boldsymbol{x}_T|\boldsymbol{x}_0)$ and $p(\boldsymbol{x}_T)$ are predefined.

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\mathbf{I}\right) \qquad p(\boldsymbol{x}_T) \sim \mathcal{N}(\boldsymbol{x}; \mathbf{0}, \mathbf{I})$$

Forward Convergence



Then, $q(\boldsymbol{x}_T|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}; \mathbf{0}, \mathbf{I})$?

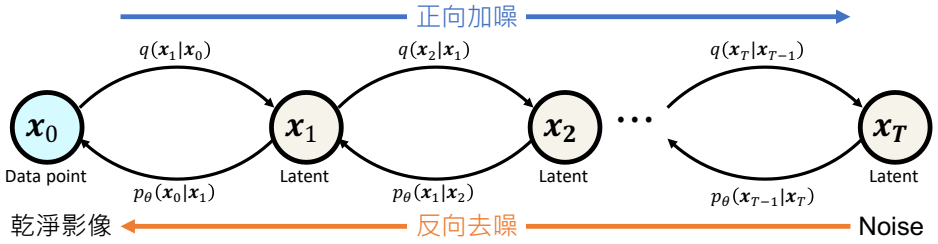
Yes, under certain assumptions.

Recall

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

where $\{\beta_t \in (0,1)\}_{t=1}^T$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_T$.

Forward Convergence



Then, $q(\boldsymbol{x}_T|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}; \mathbf{0}, \mathbf{I})$?

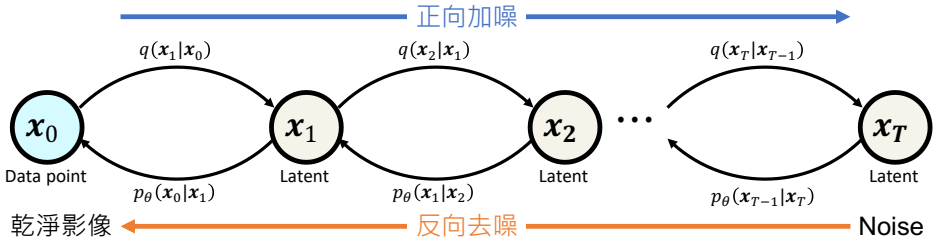
Yes, under certain assumptions.

Let $\alpha_t = 1 - \beta_t$.

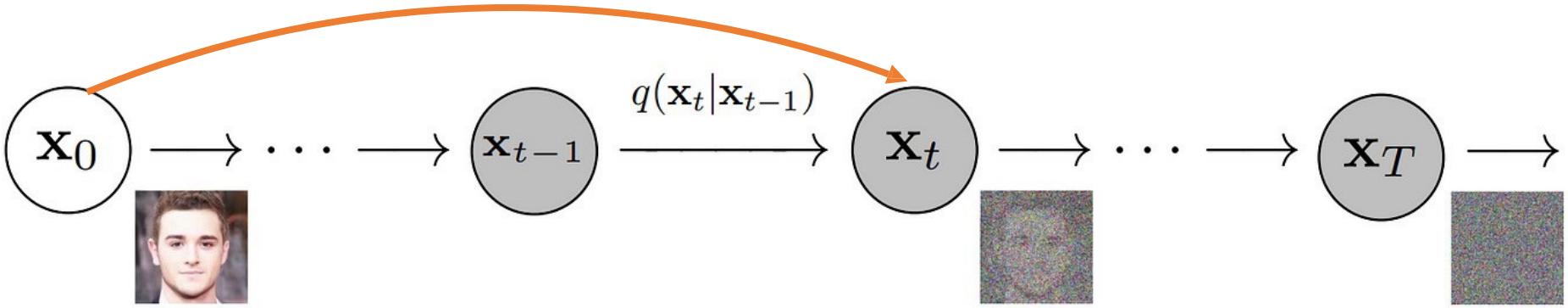
$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\,\mathbf{I})$$

where $\{\alpha_t \in (0,1)\}_{t=1}^T$ and $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_T$.

$q(\boldsymbol{x}_t|\boldsymbol{x}_0)$



Can we derive $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ from the sequence of $q(\boldsymbol{x}_{t'}|\boldsymbol{x}_{t'-1})$ for $t = 1, \ldots, t'$?



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

Basics: 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 $\boldsymbol{x}_1 + \boldsymbol{x}_2$?

Basics: Combination of Gaussian Variables

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

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

Basics: Combination of Gaussian Variables

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 $\boldsymbol{x}_1 + \boldsymbol{x}_2$?

Basics: Combination of Gaussian Variables

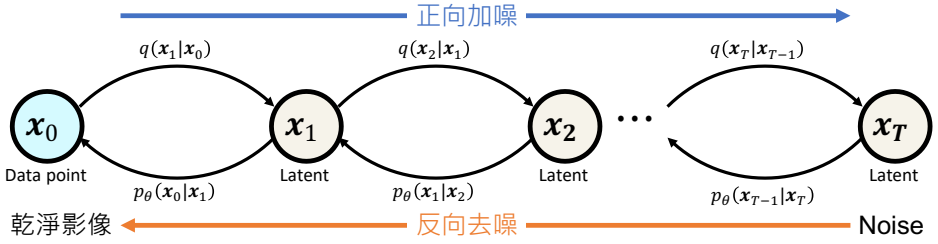
Suppose $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and

$\boldsymbol{x}_1 = \sigma_1 \boldsymbol{\varepsilon}_1$ and $\boldsymbol{x}_2 = \sigma_2 \boldsymbol{\varepsilon}_2$.

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.

Forward Convergence



$$q(\boldsymbol{x}_1|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_1; \sqrt{\alpha_1}\boldsymbol{x}_0, (1-\alpha_1)\,\mathbf{I})$$

$$q(\boldsymbol{x}_2|\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_2; \sqrt{\alpha_2}\boldsymbol{x}_1, (1-\alpha_2)\,\mathbf{I})$$

Q. What is the distribution of $q(\boldsymbol{x}_2|\boldsymbol{x}_0)$?

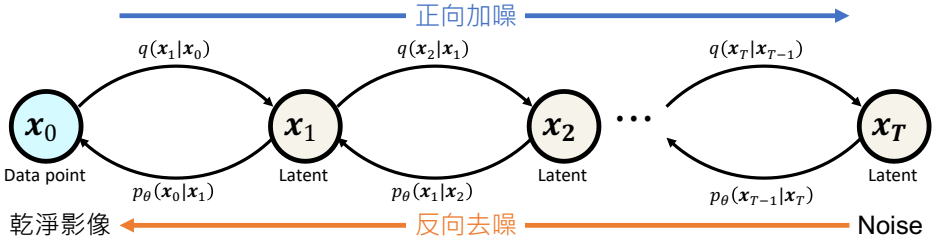
Hint. Let's use the reparameterization trick:

$$\boldsymbol{x}_1 = \sqrt{\alpha_1}\boldsymbol{x}_0 + \sqrt{1-\alpha_1}\boldsymbol{\epsilon}_0$$

$$\boldsymbol{x}_2 = \sqrt{\alpha_2}\boldsymbol{x}_1 + \sqrt{1-\alpha_2}\boldsymbol{\epsilon}_1$$

$\boldsymbol{\epsilon}_0, \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Forward Convergence



A. $\boldsymbol{x}_2 = \sqrt{\alpha_2}\boldsymbol{x}_1 + \sqrt{1-\alpha_2}\boldsymbol{\epsilon}_1$

$= \sqrt{\alpha_2}\left(\sqrt{\alpha_1}\boldsymbol{x}_0 + \sqrt{1-\alpha_1}\boldsymbol{\epsilon}_0\right) + \sqrt{1-\alpha_2}\boldsymbol{\epsilon}_1$

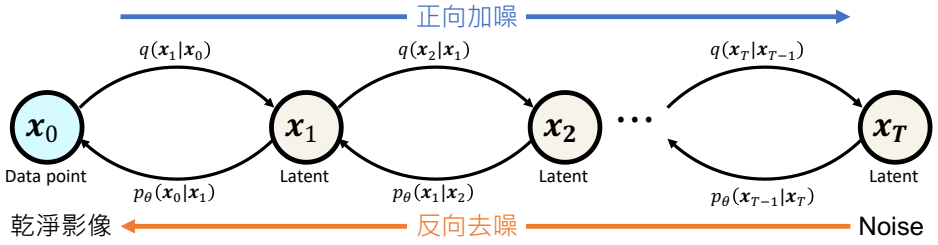
$= \sqrt{\alpha_2\alpha_1}\boldsymbol{x}_0 + \sqrt{\alpha_2(1-\alpha_1)}\boldsymbol{\epsilon}_0 + \sqrt{1-\alpha_2}\boldsymbol{\epsilon}_1$

$= \sqrt{\alpha_2\alpha_1}\boldsymbol{x}_0 + \sqrt{(1-\alpha_2\alpha_1)}\bar{\boldsymbol{\epsilon}}_0$

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

$$\therefore q(\boldsymbol{x}_2|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_2\alpha_1}\boldsymbol{x}_0, (1-\alpha_2\alpha_1)\mathbf{I})$$

Forward Convergence



$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}$$

$$= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}$$

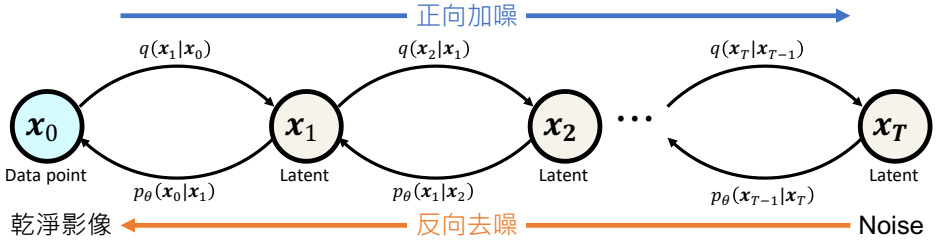
$$= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\boldsymbol{\epsilon}_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{(1-\alpha_t\alpha_{t-1})}\bar{\boldsymbol{\epsilon}}_{t-2}$$

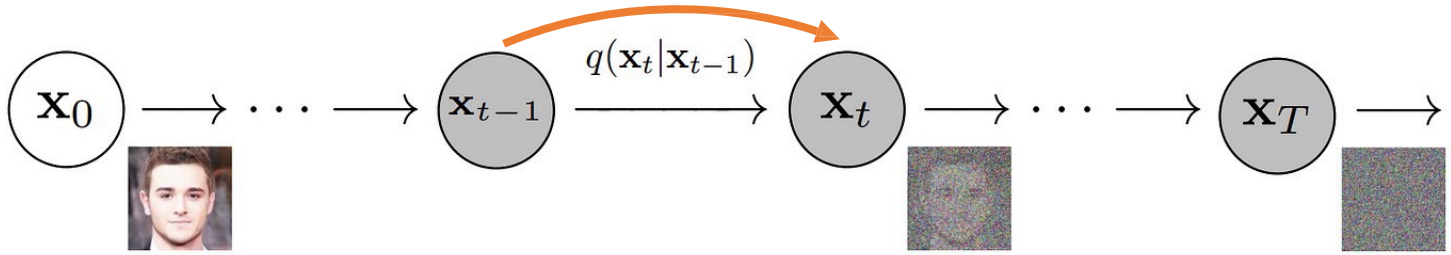
$$= \cdots$$

$$= \sqrt{\textstyle\sum_{i=1}^{t}\alpha_i}\,\boldsymbol{x}_0 + \sqrt{\left(1-\textstyle\prod_{i=1}^{t}\alpha_i\right)}\bar{\boldsymbol{\epsilon}}_0$$

Forward Convergence

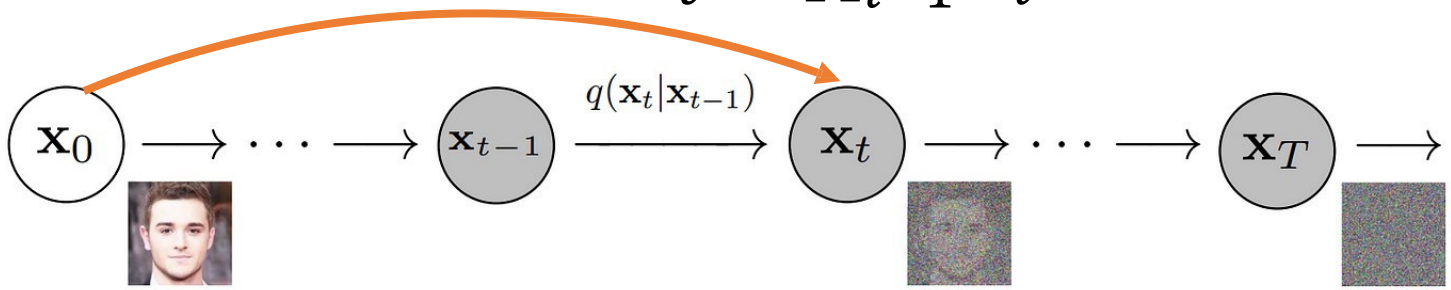


$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$



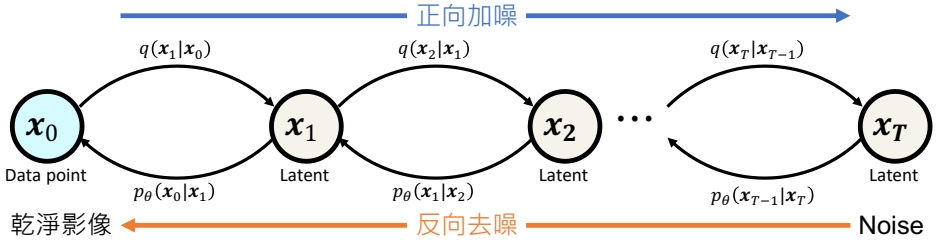
$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$$

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



Forward jump/正向跳躍

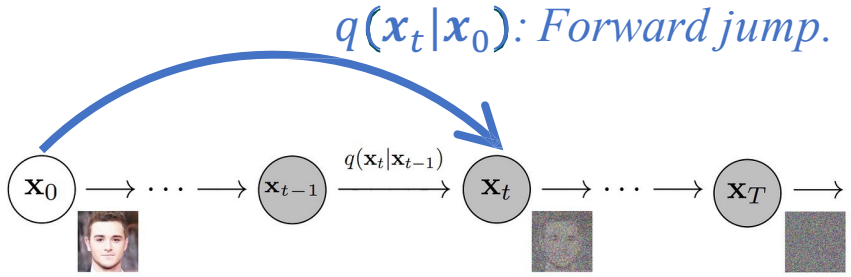
$q(\boldsymbol{x}_t|\boldsymbol{x}_0)$



Given $\boldsymbol{x}_0$, $\boldsymbol{x}_t$ at any arbitrary timestep t can be directly sampled from a Gaussian distribution without a Markov chain:

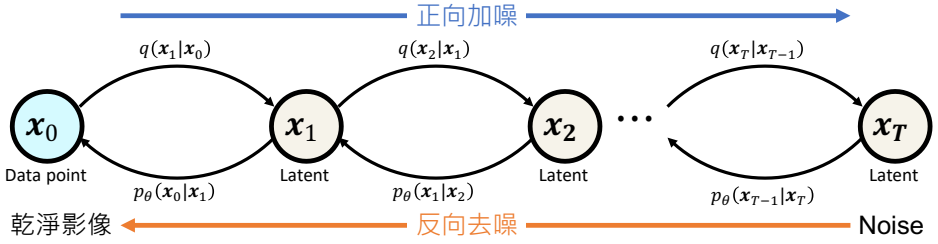
$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right).$$

Note that $\bar{\alpha}_1 > \bar{\alpha}_2 > \cdots > \bar{\alpha}_T$.



Forward jump/正向跳躍

Forward Convergence



$$\overset{0}{} \qquad \overset{1}{}$$

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

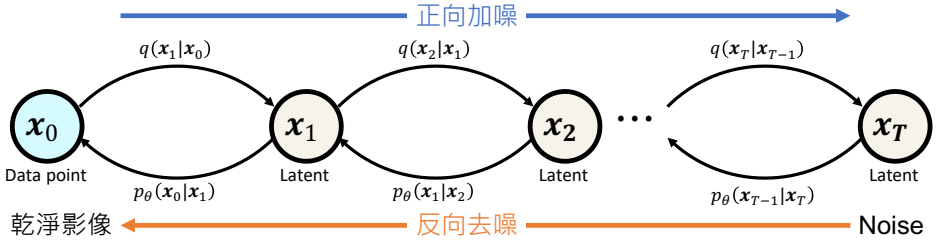
where $\bar{\alpha}_T = \prod_{t=1}^{T} \alpha_t = \prod_{t=1}^{T}(1-\beta_t)$.

Q. When $\{\beta_t \in (0,1)\}_{t=1}^{T}$, What is

$$\lim_{T\to\infty} \bar{\alpha}_T = \lim_{T\to\infty} \prod_{t=1}^{T}(1-\beta_t)?$$

$$= 0$$

Forward Convergence



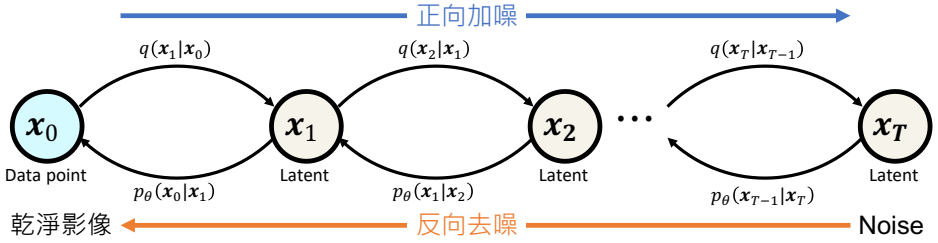
$$q(\boldsymbol{x}_T|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_T; \sqrt{\bar{\alpha}_T}\boldsymbol{x}_0, (1-\bar{\alpha}_T)\mathbf{I}\right)$$

0 0

where $\bar{\alpha}_T = \prod_{t=1}^{T} \alpha_t = \prod_{t=1}^{T}(1-\beta_t)$.

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

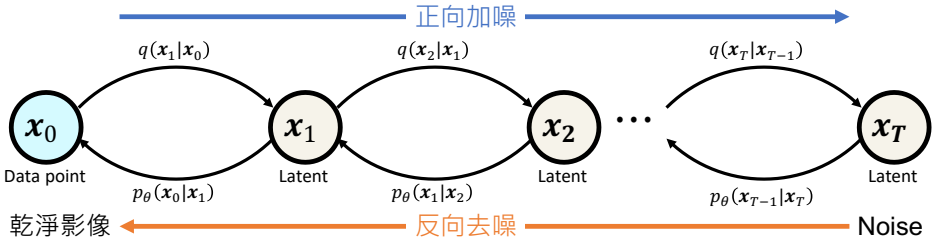
Prior Matching Term $\mathcal{L}_T$



$$D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$$

Close to zero by the definition of the forward transition distribution $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. *Nothing to do for the optimization*.

ELBO



Decompose the negative ELBO in a different way:

$$-\log p(\boldsymbol{x}_0) = -\log \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[-\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

$$= \cdots =$$

$$-\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$$

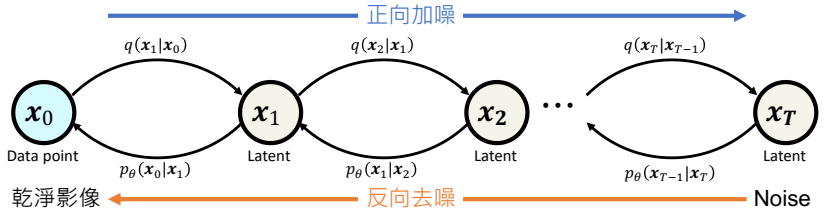
$$+D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T)) \quad \rightarrow 0$$

$$+\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

Denoising matching term $\mathcal{L}_{t-1}$

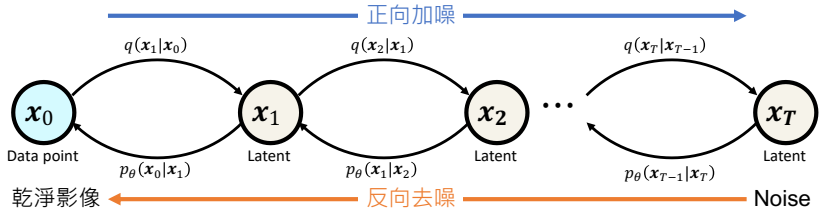
Denoising Matching Term $\mathcal{L}_{t-1}$

Denoising Matching Term $\mathcal{L}_{t-1}$



$$\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

The variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ should be close to $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ for each t.



Denoising Matching Term $\mathcal{L}_{t-1}$

What is $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$?

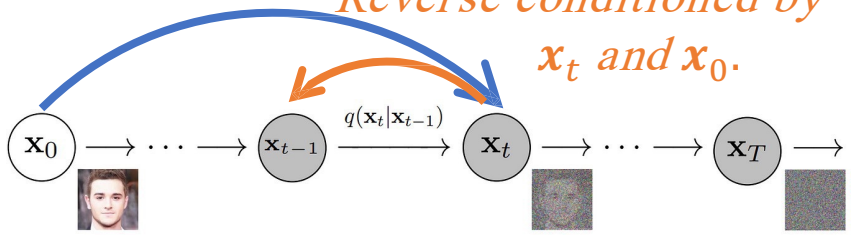
$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}$$

Same as $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$, the forward transition, since it's a Markovian process.

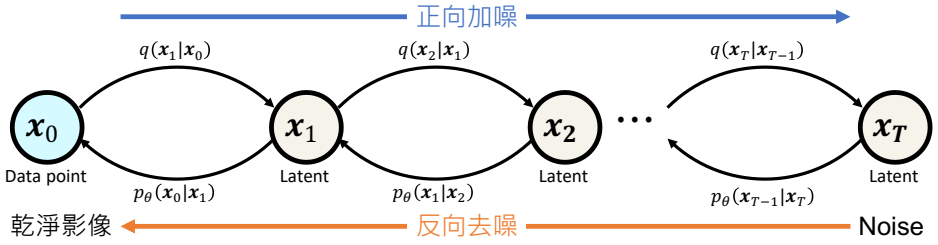
We have seen how to compute these.

Forward jump/正向跳躍

$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$: *Reverse conditioned by* $\boldsymbol{x}_t$ *and* $\boldsymbol{x}_0$.



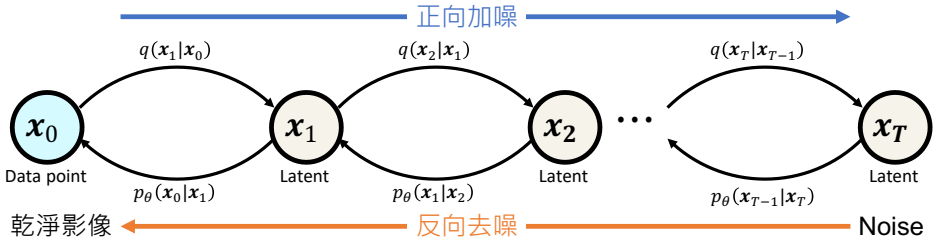
$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$



$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}$$

Q. What are $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$, $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)$, and $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$?

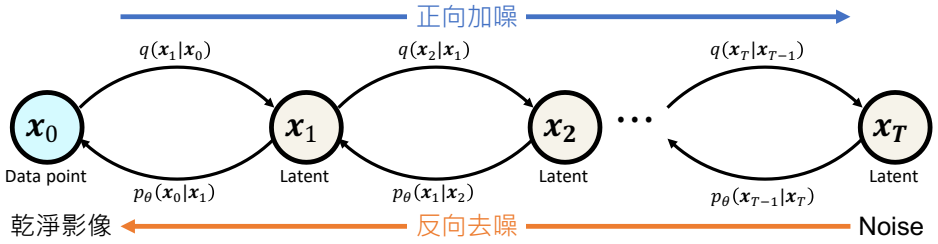
$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$



$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}$$

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

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



$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{\left(\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_{t-1}\right)}{1-\alpha_t} + \frac{\left(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0\right)}{1-\bar{\alpha}_{t-1}} - \frac{\left(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0\right)}{1-\bar{\alpha}_t}\right)\right)$$

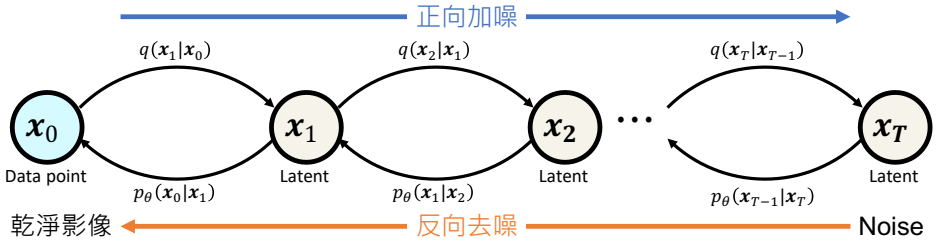
$$= \cdots$$

$$= \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$$ *Another normal distribution!*

where $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0$ and $\tilde{\sigma}_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$

https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
https://zhuanlan.zhihu.com/p/558937247

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

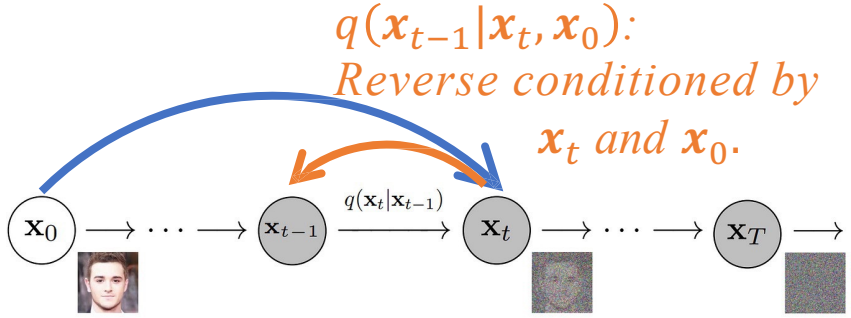


- The mean $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0$ is a function of both $\boldsymbol{x}_t$ and $\boldsymbol{x}_0$.

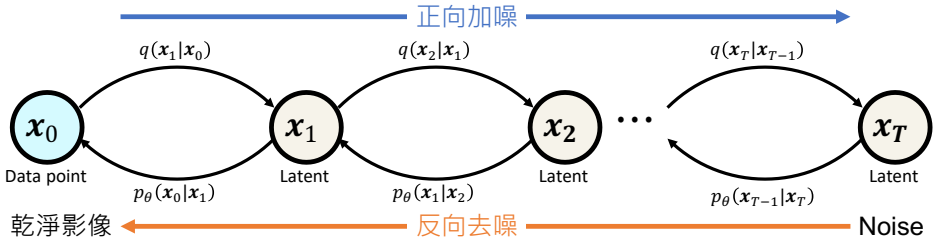
$\boldsymbol{x}_{t-1}$ 的 mean $\tilde{\mu}$ 是 $\boldsymbol{x}_t$ 與 $\boldsymbol{x}_0$ 的線性內插

- The covariance $\tilde{\sigma}_t^2\mathbf{I} = \left(\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\right)\mathbf{I}$ is predefined from the user-defined $\{\beta_t\}_{t=1}^T$.

$\boldsymbol{x}_{t-1}$ 的 variance $\tilde{\sigma}_t^2$ 與 $\boldsymbol{x}_t$ 或 $\boldsymbol{x}_0$ 無關，
僅由 time step t 與 β_t 預先定義



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



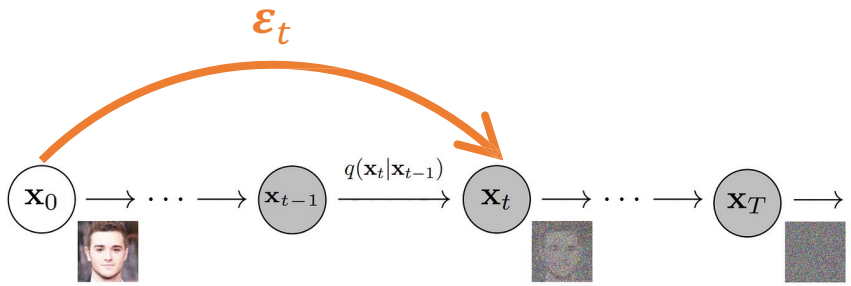
From the forward jump $q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$,

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t \text{ where } \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

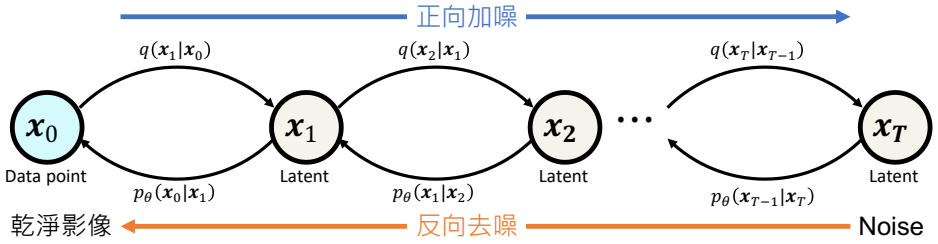
If $\boldsymbol{x}_t$ and $\boldsymbol{x}_0$ are given, define $\boldsymbol{\varepsilon}_t$ as

$$\boldsymbol{\varepsilon}_t = \frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_0$$

Q. Rewrite $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0)$ as a function of $\boldsymbol{x}_t$ and $\boldsymbol{\varepsilon}_t$.



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

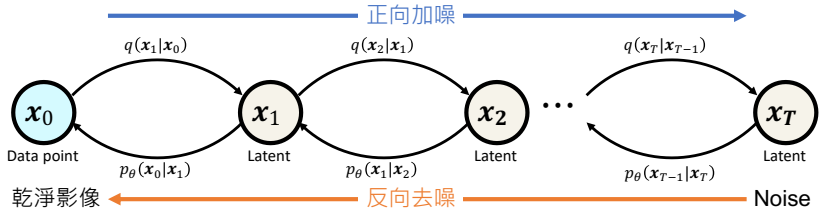


A.

$$\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_0$$

$$= \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_t\right)$$

Denoising Matching Term $\mathcal{L}_{t-1}$

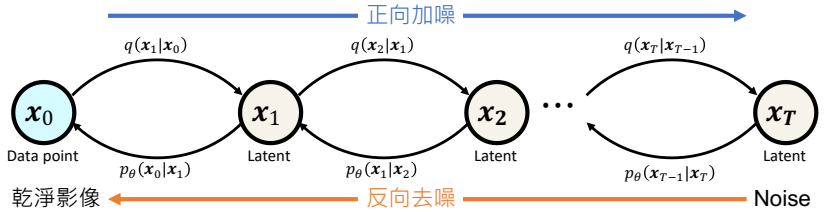


Back to the denoising matching term…

$$\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

How to model the variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$?

Denoising Matching Term $\mathcal{L}_{t-1}$



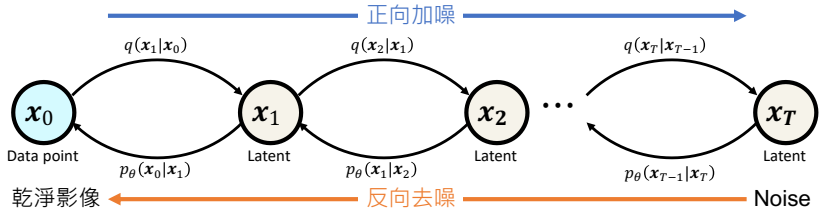
- For $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$, the variance $\tilde{\sigma}_t^2$ is *not* a function of $\boldsymbol{x}_t$ and $\boldsymbol{x}_0$.

- Hence, define the variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ as

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

where $\mu_\theta(\boldsymbol{x}_t, t)$ is the mean predictor.

Denoising Matching Term $\mathcal{L}_{t-1}$



- For $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$, the variance $\tilde{\sigma}_t^2$ is *not* a function of $\boldsymbol{x}_t$ and $\boldsymbol{x}_0$.

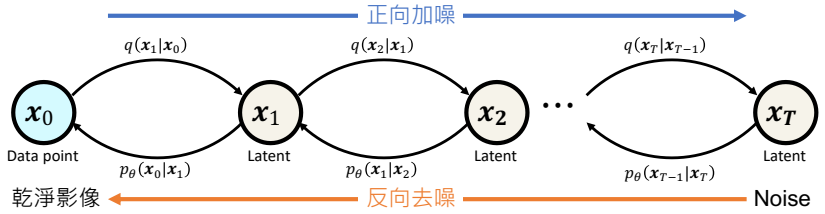
- Hence, define the variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ as

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

  where $\mu_\theta(\boldsymbol{x}_t, t)$ is the mean predictor.

Q. 為什麼我們要把 likelihood (p_θ) 的 variance 定義成與 posterior (q) 相同？

Denoising Matching Term $\mathcal{L}_{t-1}$



- For $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$, the variance $\tilde{\sigma}_t^2$ is *not* a function of $\boldsymbol{x}_t$ and $\boldsymbol{x}_0$.

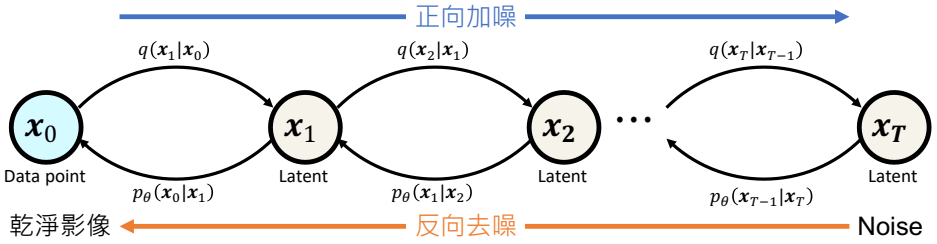
- Hence, define the variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ as

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

where $\mu_\theta(\boldsymbol{x}_t, t)$ is the mean predictor.

A. 簡化 optimization 目標：當我們最小化 KL divergence，如果兩個分佈的 variance 相同，KL divergence 會簡化為只需要匹配 mean，這讓訓練變成一個簡單的 mean 預測問題

Denoising Matching Term

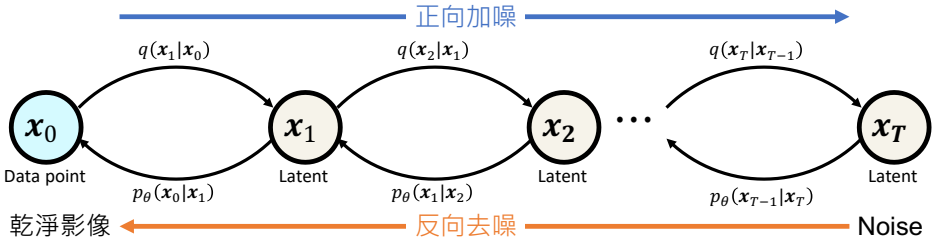


How to compute

$$\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]?$$

Q. When $p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_p,\sigma^2\mathbf{I}\right)$ and $q(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_q,\sigma^2\mathbf{I}\right)$, What is $D_{KL}(p\|q)$?

Denoising Matching Term



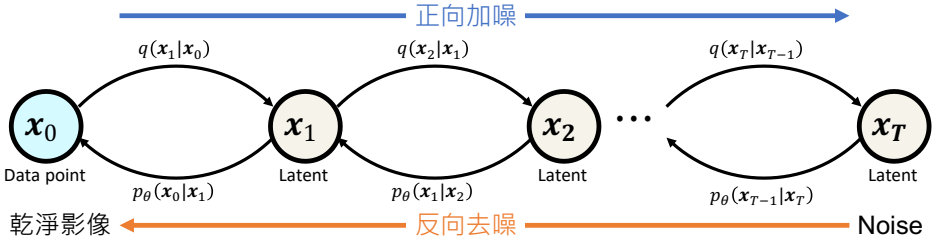
A.

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_p, \sigma^2 \mathbf{I}\right)$$

$$q(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_q, \sigma^2 \mathbf{I}\right)$$

$$D_{KL}(p\|q) = \frac{1}{2\sigma^2}\left\|\boldsymbol{\mu}_q - \boldsymbol{\mu}_p\right\|^2$$

Denoising Matching Term



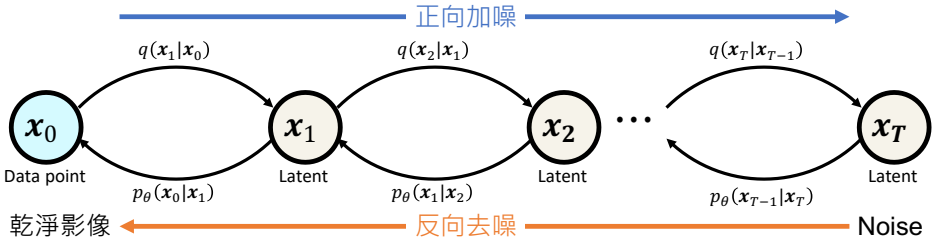
How to compute

$$\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]?$$

$$\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

$$= \frac{1}{2\tilde{\sigma}_t^2}\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\mu_\theta(\boldsymbol{x}_t,t)-\tilde{\mu}(\boldsymbol{x}_t,\boldsymbol{x}_0)\|^2]$$

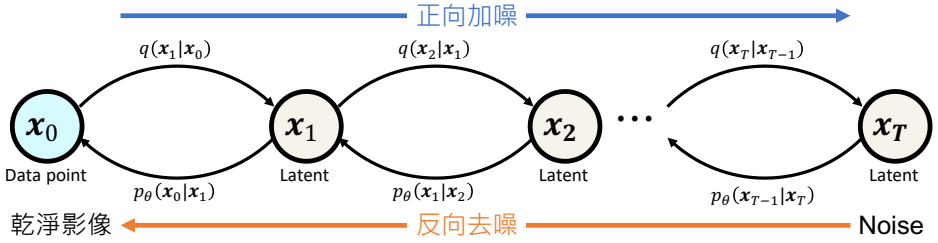
$\boldsymbol{x}_0$ Predictor



Q. What if we have a $\boldsymbol{x}_0$ predictor $\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, t)$ instead of the mean predictor $\mu_\theta(\boldsymbol{x}_t, t)$? Note that

$$\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0$$

$\boldsymbol{x}_0$ Predictor



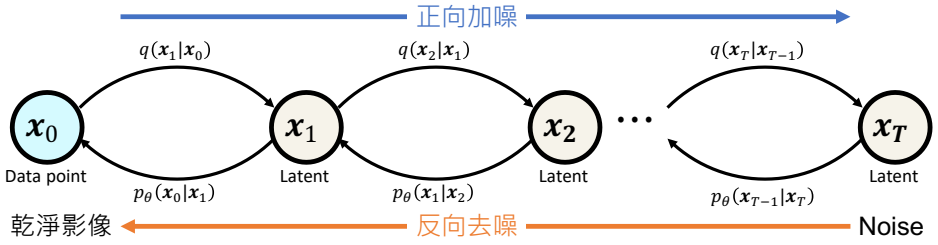
A. $\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$

$$= \frac{1}{2\tilde{\sigma}_t^2}\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\mu_\theta(\boldsymbol{x}_t,t)-\tilde{\mu}(\boldsymbol{x}_t,\boldsymbol{x}_0)\|^2]$$

$$= \frac{1}{2\tilde{\sigma}_t^2}\frac{\bar{\alpha}_{t-1}\beta_t^2}{(1-\bar{\alpha}_t)^2}\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|^2]$$

$$= \omega_t\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|^2]$$

x_0 Predictor



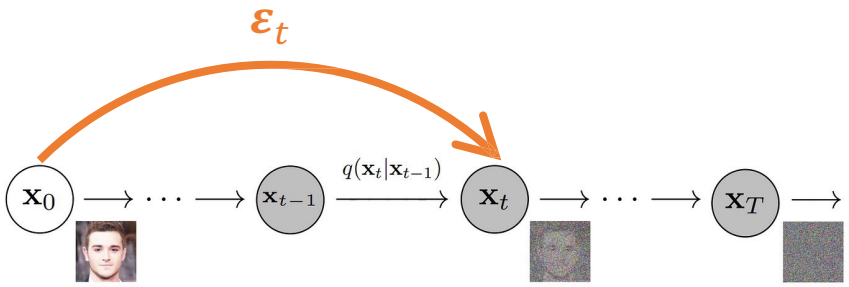
$$\omega_t \mathbb{E}_{q(x_t|x_0)}[\|\hat{x}_\theta(x_t, t) - x_0\|^2]$$

- x_t is sampled from x_0.

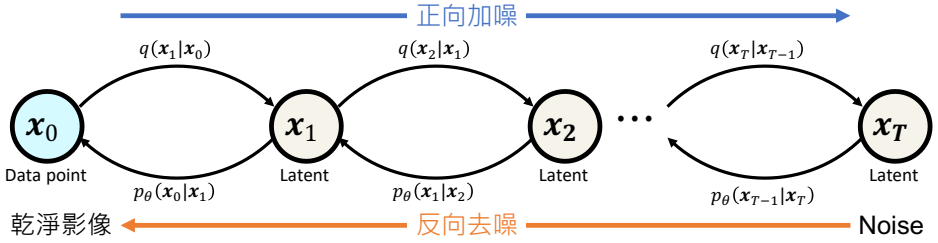
給定一張 noisy 影像 x_t，理論上有無數個可能的原始影像 x_0 都能通過加 noise 產生這個 x_t

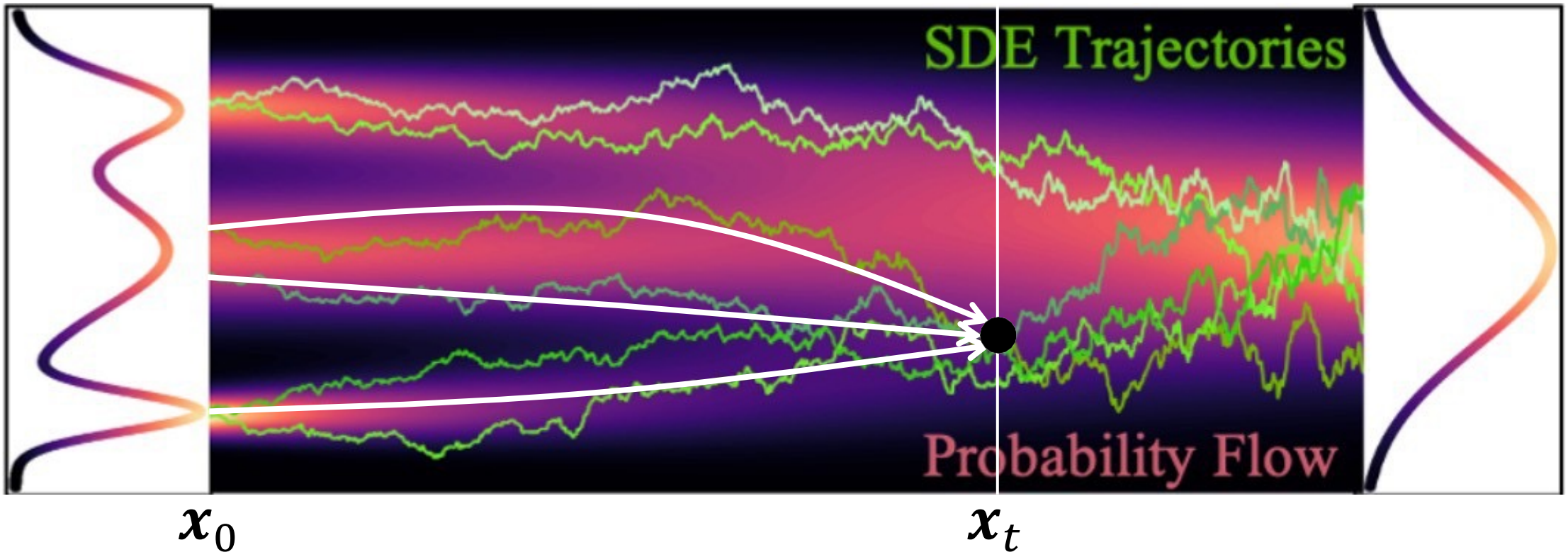
- From x_t, predict the *expected* value of x_0 that would result in sampling x_t from it through the forward jump.

x_0 Predictor 在做的事情是：看到一張 noisy 影像 x_t，猜測原始影像 x_0，但因為有多種可能，所以預測「平均來說最可能」（期望值）的那個

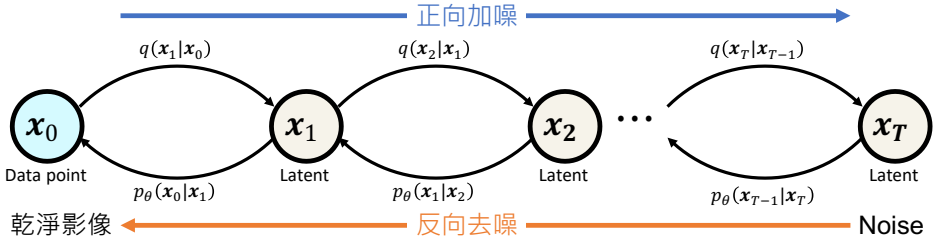


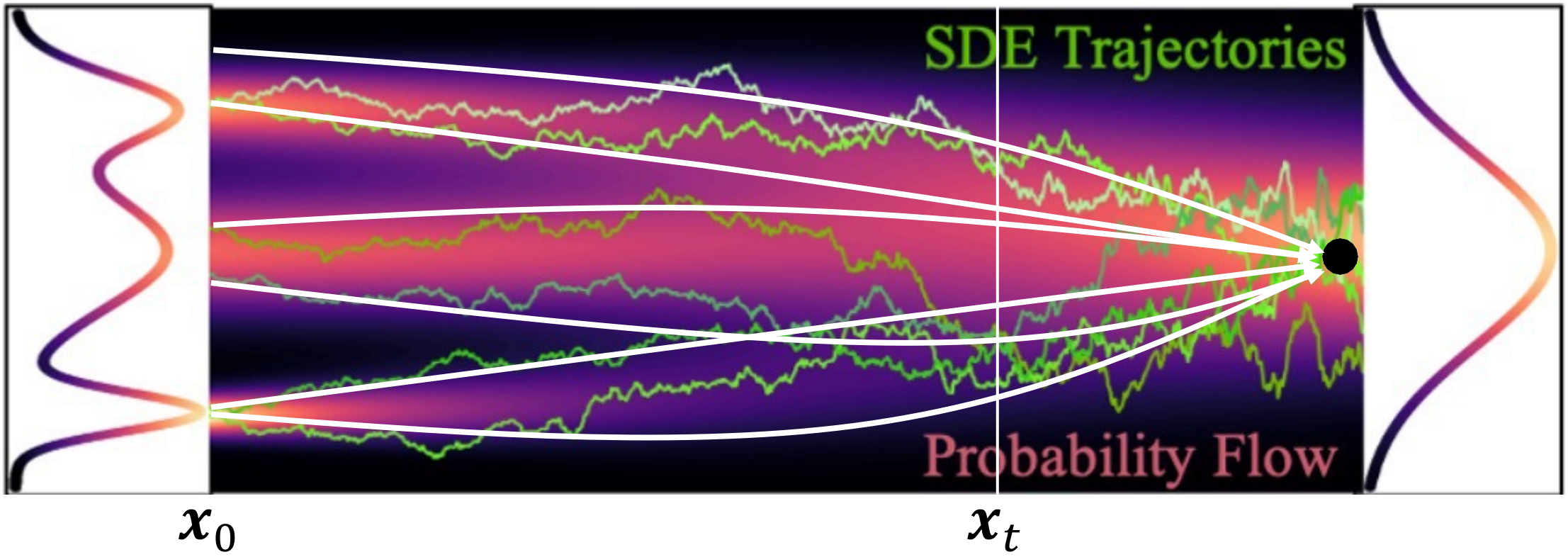
x_0 Predictor



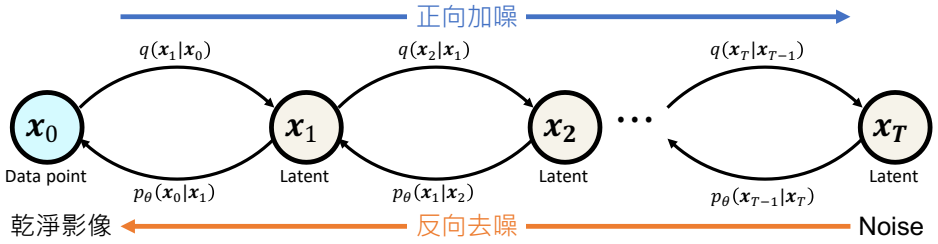


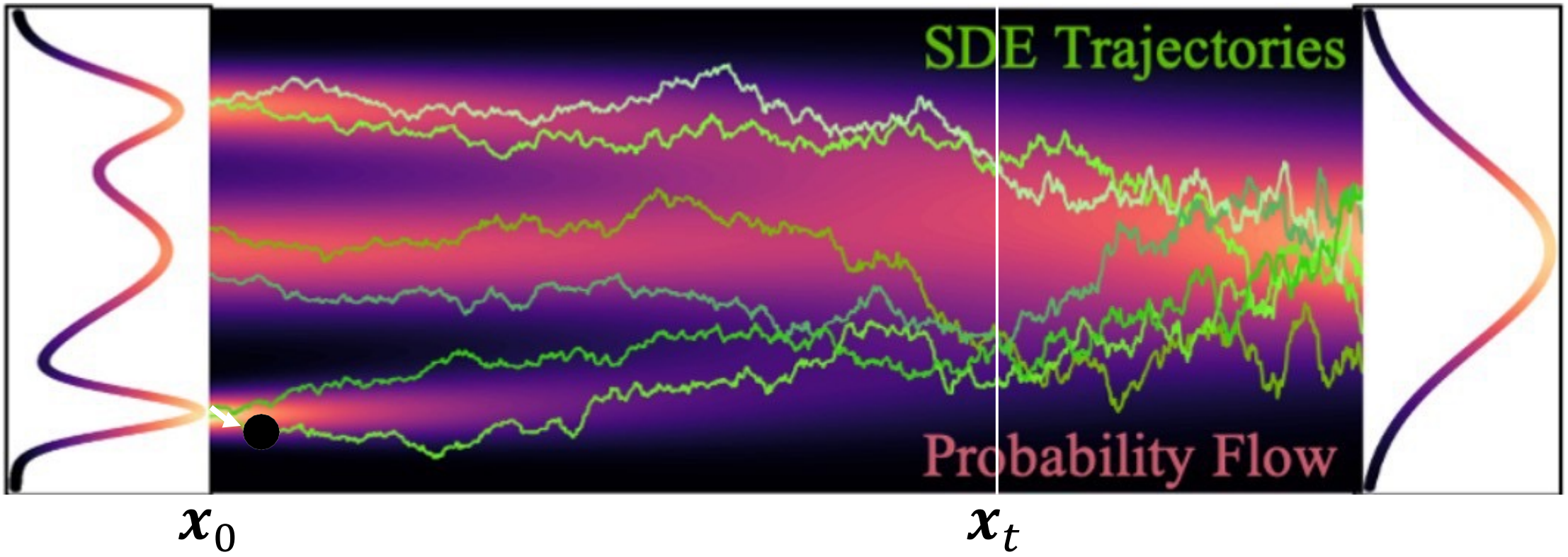
x_0 Predictor



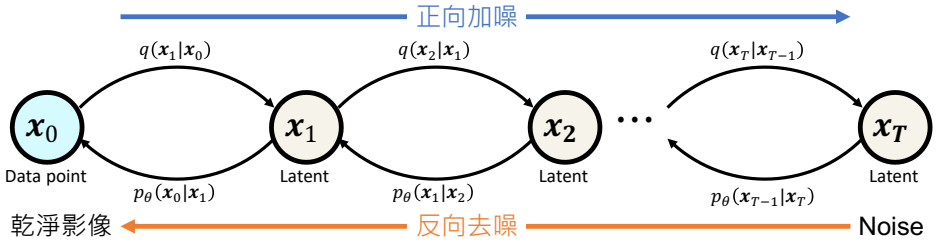


x_0 Predictor





$\boldsymbol{x}_0$ Predictor



- Note that our goal is to sample $\boldsymbol{x}_0$ from a standard normal sample $\boldsymbol{x}_T$ and through latent variables $\boldsymbol{x}_{T-1}, \boldsymbol{x}_{T-2}, \cdots, \boldsymbol{x}_1$.

- But for every $\boldsymbol{x}_t$, we directly predict the expected value of $\boldsymbol{x}_0$ from $\boldsymbol{x}_t$.

雖然生成過程是逐步的，但訓練時我們其實是直接預測原始數據 $\boldsymbol{x}_0$

- 表面上看起來像是每步只預測「前一步」
- 實際上在每個 time step t，network 都在嘗試直接預測 $\boldsymbol{x}_0$

為什麼這樣設計？

因為 posterior 的 mean 依賴於 $\boldsymbol{x}_0$: $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$

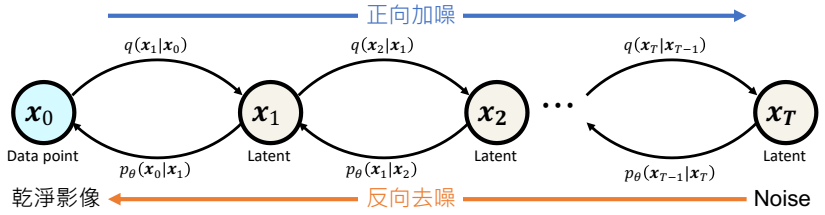
（$\boldsymbol{x}_{t-1}$ 的 mean $\tilde{\mu}$ 是 $\boldsymbol{x}_t$ 與 $\boldsymbol{x}_0$ 的線性內差）所以需要預測 $\boldsymbol{x}_0$

不是盲目地去噪，而是基於對最終目標的理解這種「**全局視野**」，每一步都在糾正對原始圖像的估計，讓模型能更好地保持圖像的結構

Denoising Diffusion Probabilistic Models (DDPM)

Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020.

Basic Idea of Generative Models



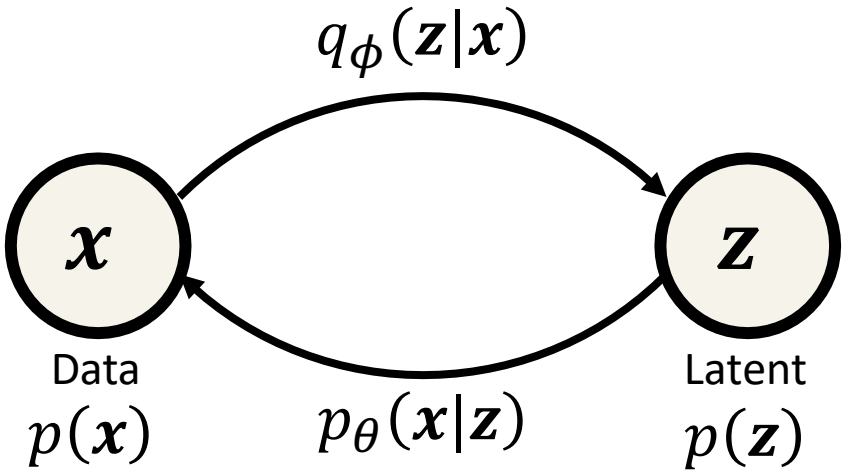
Likelihood
Decoder

Posterior
Encoder

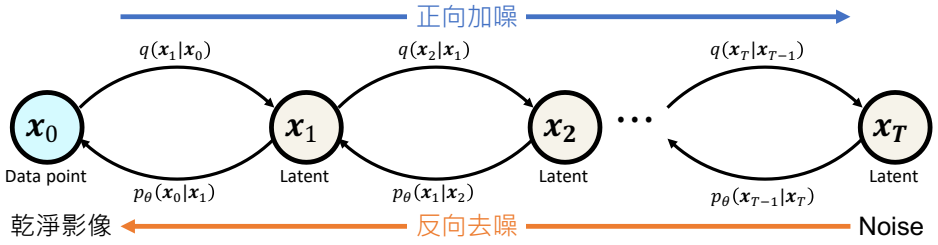
Prior
Latent

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

Marginal
Data

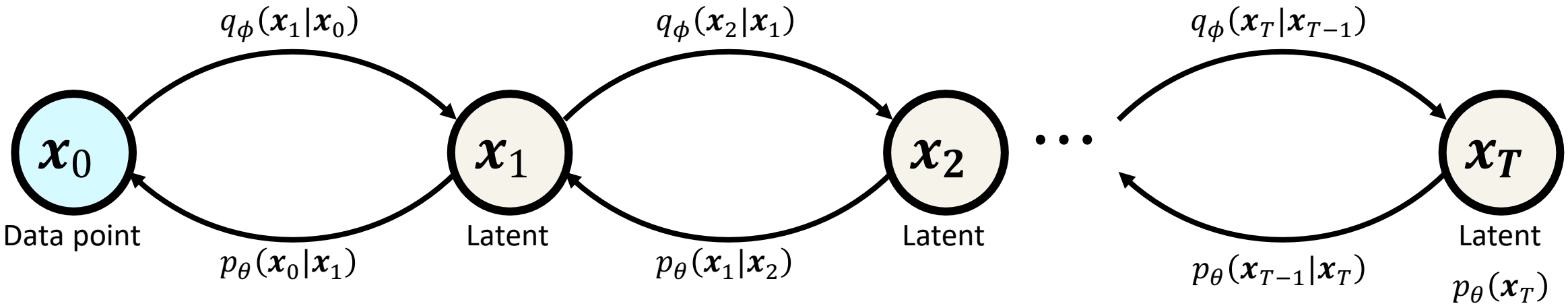


Variational Diffusion Models

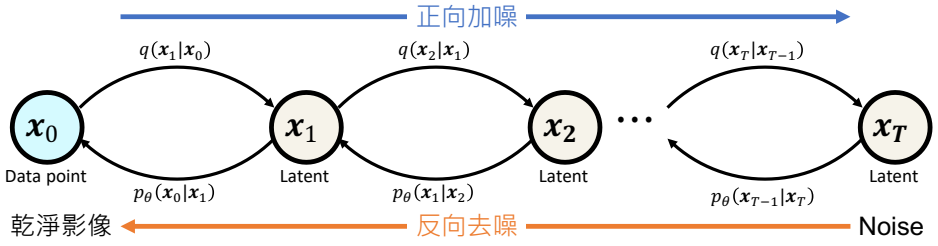


Key features:

1. Both forward (encoding) and reverse (decoding) processes are sequential processes (with a sequence of latent variables).

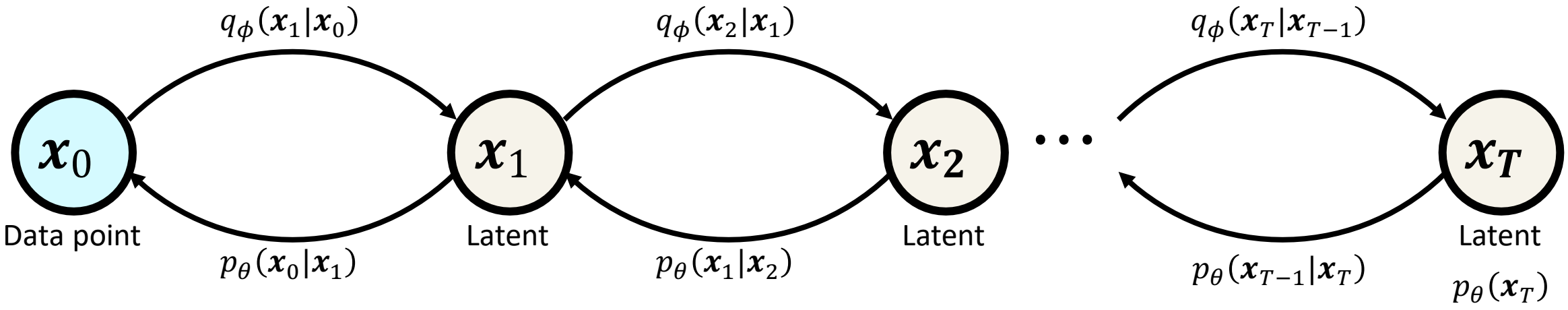


Variational Diffusion Models

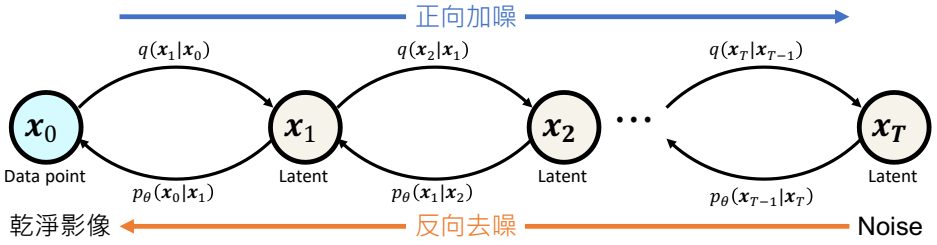


Key features:

1. Both forward (encoding) and reverse (decoding) processes are sequential processes (with a sequence of latent variables).
2. The forward process is not learned but predefined.



Variational Diffusion Models



Key features:

1. Both forward (encoding) and reverse (decoding) processes are sequential processes (with a sequence of latent variables).
2. The forward process is not learned but predefined.
3. The dimensions of the latent variables and the input data are the same.

Denoising Diffusion Probabilistic Models (DDPM)

Key features of DDPM (specifically):

1. The sequential forward and reverse process are Markovian processes:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)$$

Denoising Diffusion Probabilistic Models (DDPM)

Key features of DDPM (specifically):

1. The sequential forward and reverse process are Markovian processes:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)$$

2. The forward transitional distribution is defined as follows:

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

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_T$.

Denoising Diffusion Probabilistic Models (DDPM)

Key features of DDPM (specifically):

1. The sequential forward and reverse process are Markovian processes:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)$$

2. The forward transitional distribution is defined as follows:

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

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_T$.

Q. 為什麼正向加噪要這樣定義？

Denoising Diffusion Probabilistic Models (DDPM)

Key features of DDPM (specifically):

1. The sequential forward and reverse process are Markovian processes:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)$$

2. The forward transitional distribution is defined as follows:

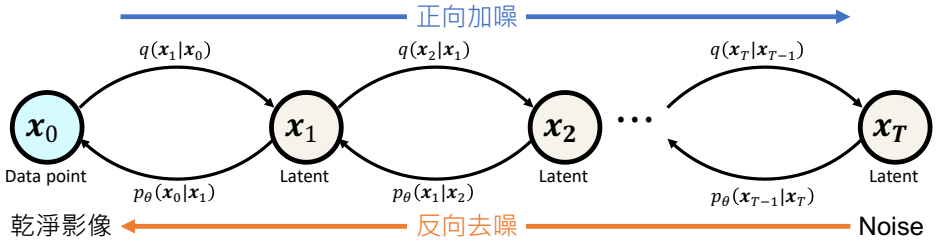
$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\mathbf{I}\right)$$

where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_T$.

A. 保證收斂到 standard normal distribution，$q(\boldsymbol{x}_T|\boldsymbol{x}_0) \to \mathcal{N}(\boldsymbol{x}; \mathbf{0}, \mathbf{I})$

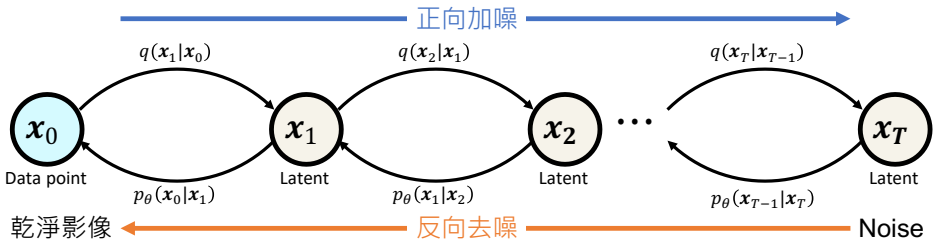
$$\boldsymbol{x}_T \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$$

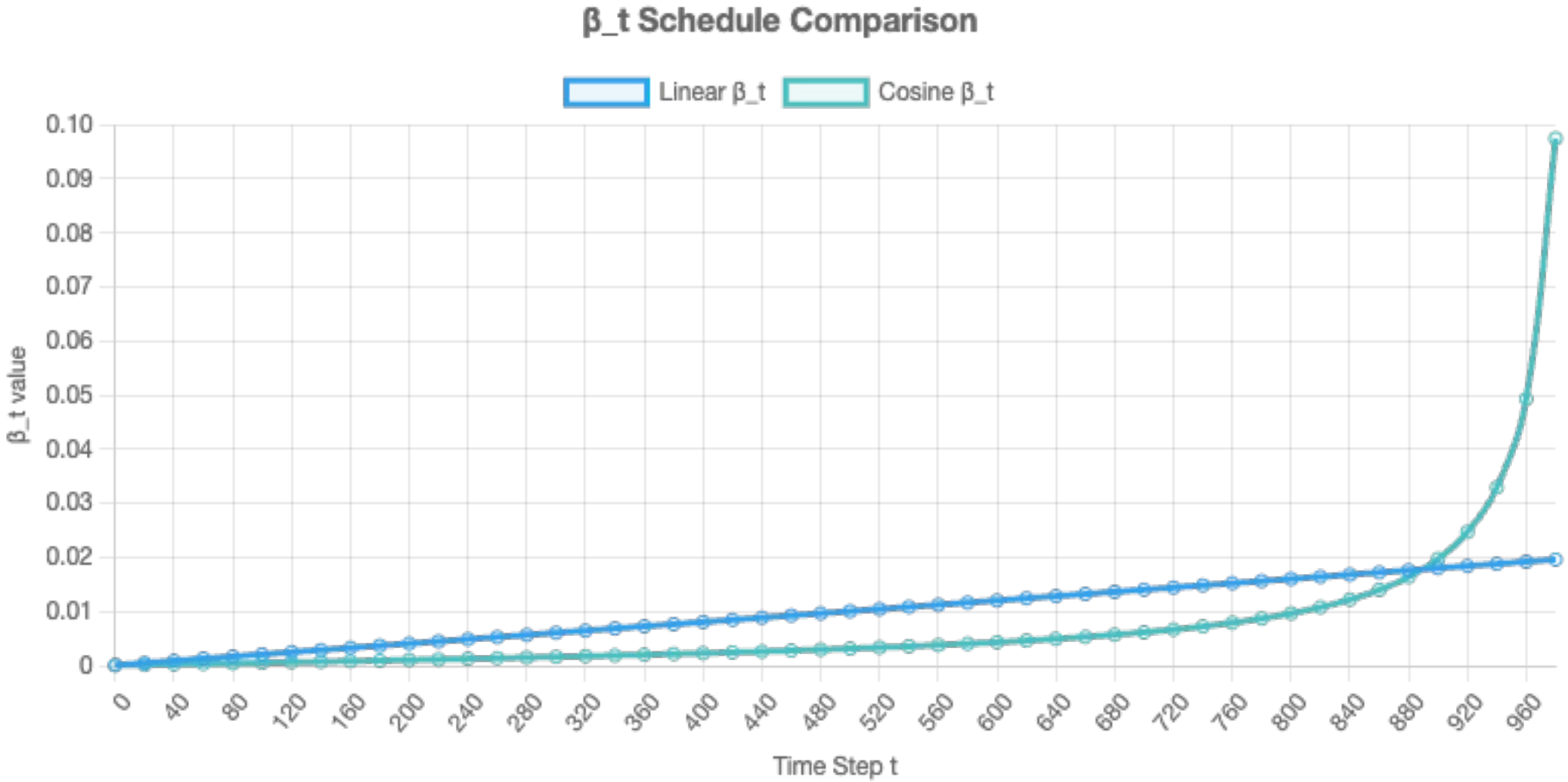
Choice of β_t



- Learned.
- Constant.
- Linearly or quadratically increased.
- Follows a cosine function
  (Nichol and Dhariwal, Improved Denoising Diffusion Probabilistic Models, ICML 2021).
- Note that the reverse step $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ becomes a Gaussian form only when β_t is small ($\beta_t \ll 1$).

Choice of β_t

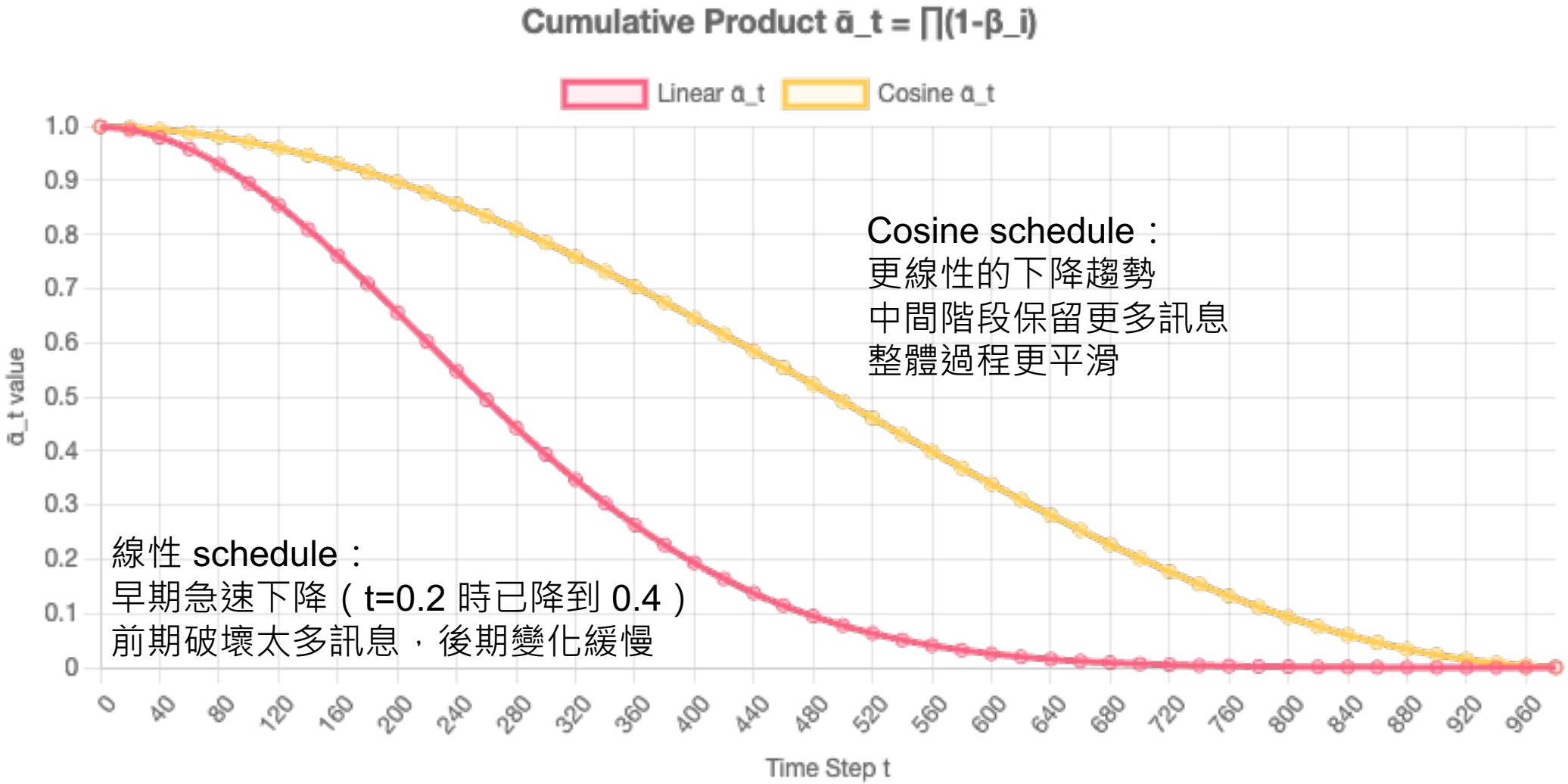




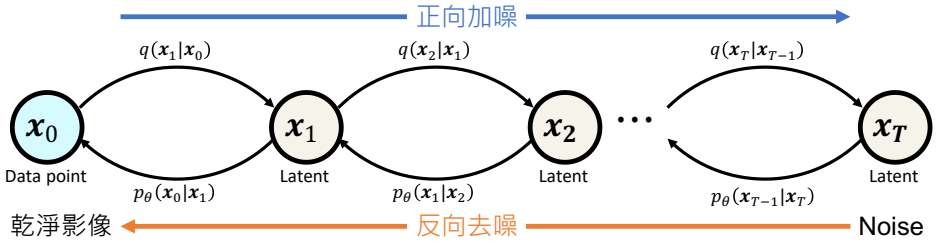
Choice of β_t

$\bar{\alpha}_t$ 代表原始影像的保留比例

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



Choice of β_t



Linear Schedule

$$\beta_t = 0.0001 + \frac{t}{T}(0.02 - 0.0001)$$

簡單直接的線性增長，早期階段 noise 增加較快，可能導致訓練不穩定

Cosine Schedule

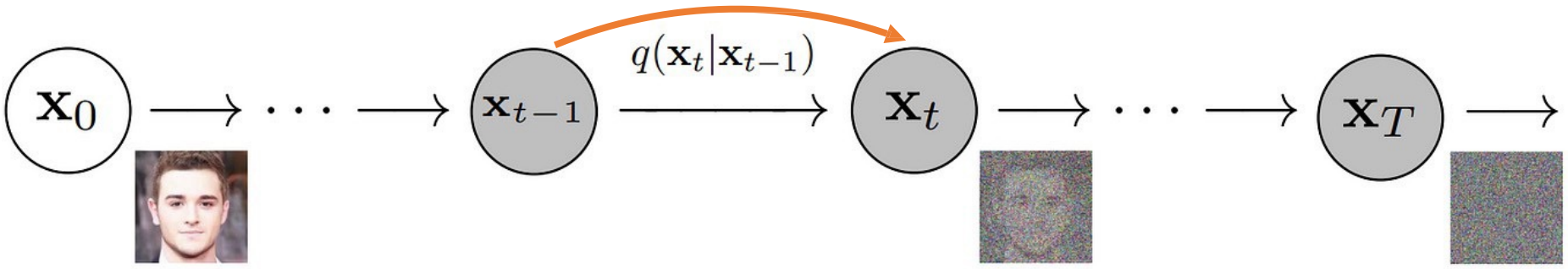
$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, f(t) = \cos\left(\frac{t/T + s}{1+s} \cdot \frac{\pi}{2}\right)^2$$

更平滑的噪音增加，中間階段保留更多資訊，訓練更穩定，生成品質通常更好，適合高解析度影像

Denoising Diffusion Probabilistic Models (DDPM)

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

where $\alpha_t = 1 - \beta_t$.

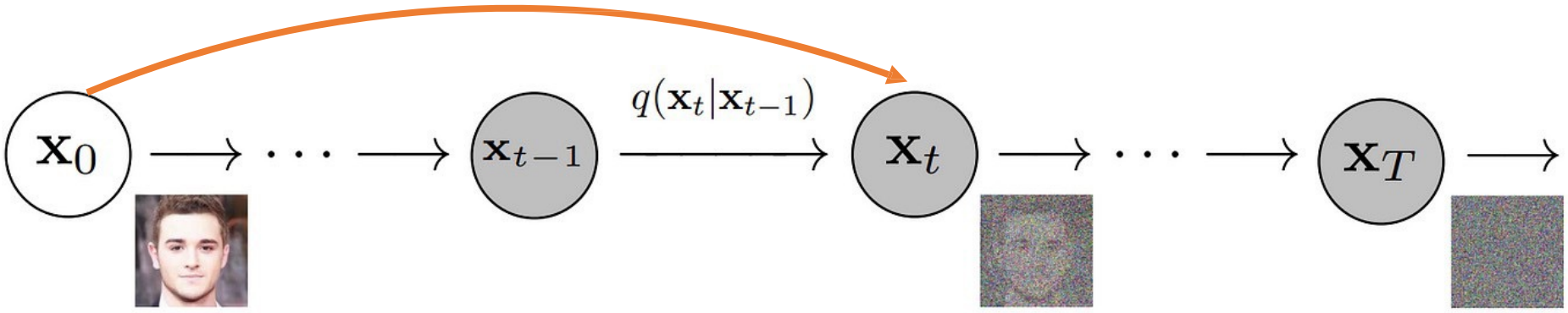


Denoising Diffusion Probabilistic Models (DDPM)

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

串起來推導可得

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

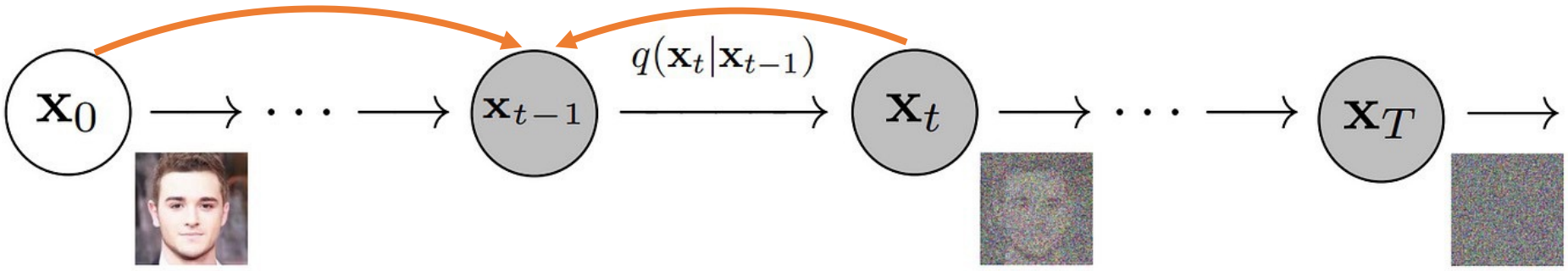


Denoising Diffusion Probabilistic Models (DDPM)

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$$

where $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_0$ and $\tilde{\sigma}_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

$\boldsymbol{x}_{t-1}$ 的 mean $\tilde{\mu}$ 是 $\boldsymbol{x}_t$ 與 $\boldsymbol{x}_0$ 的線性內差



Denoising Diffusion Probabilistic Models (DDPM)

1. $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}\right)$
   where $\alpha_t = 1 - \beta_t$.

   正向加噪過程，是我們精心定義的

2. $q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$
   where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

   正向跳躍，串起來推導可得

3. $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2\mathbf{I})$
   where $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_0$ and $\tilde{\sigma}_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

$\boldsymbol{x}_{t-1}$ 是 Gaussian, mean 是 $\boldsymbol{x}_t$ 與 $\boldsymbol{x}_0$ 的線性內差，variance 由 time step t 與 β_t 預先定義

Denoising Diffusion Probabilistic Models (DDPM)

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$$

1. $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0$ or

mean 可以由 $\boldsymbol{x}_t$ 與 $\boldsymbol{x}_0$ 線性內差得到

2. $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{\varepsilon}_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_t\right)$

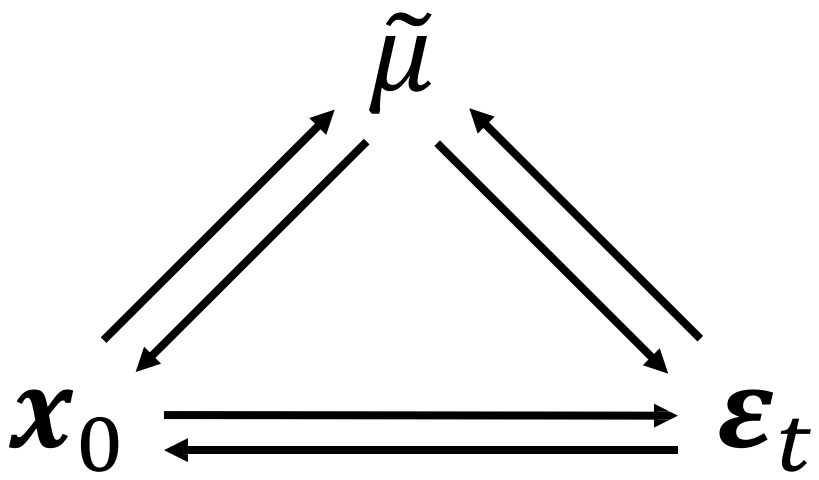
where $\boldsymbol{\varepsilon}_t = \frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_0$

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t$$

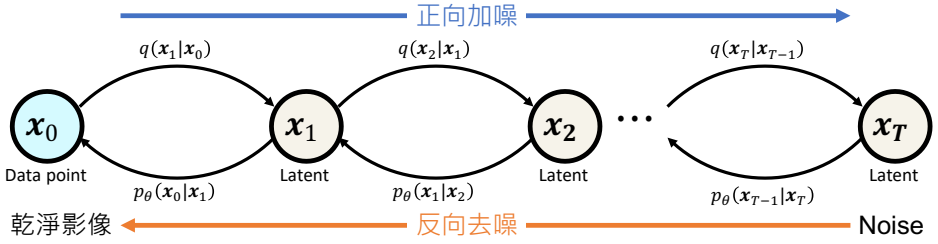
mean 也可以由 $\boldsymbol{x}_t$ 與 noise $\boldsymbol{\varepsilon}_t$ 得到

Denoising Diffusion Probabilistic Models (DDPM)

Given $\boldsymbol{x}_t$, each of $\tilde{\mu}$, $\boldsymbol{x}_0$, and $\boldsymbol{\varepsilon}_t$ can be computed from any of the others.



ELBO



The negative ELBO is decomposed into three terms:

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

$$-\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$$

Reconstruction term $\mathcal{L}_0$

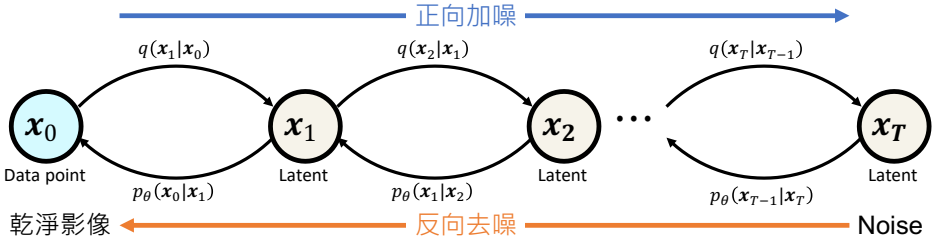
$$+D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$$

Prior matching term $\mathcal{L}_T$

$$+\sum_{t=2}^{T}\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

Denoising matching term $\mathcal{L}_{t-1}$

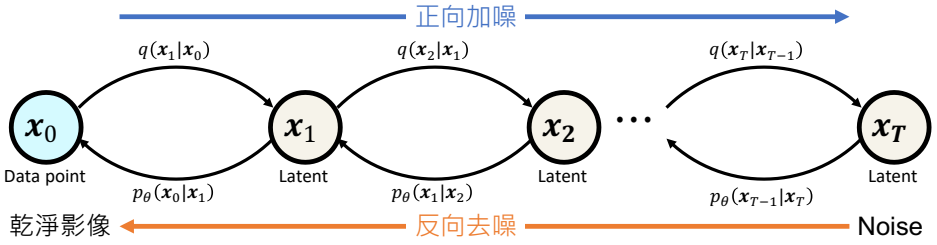
ELBO



- Reconstruction term $\mathcal{L}_0 = -\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$: The same as with VAEs; it is also negligible.

Q. 為什麼這一項可以忽略？

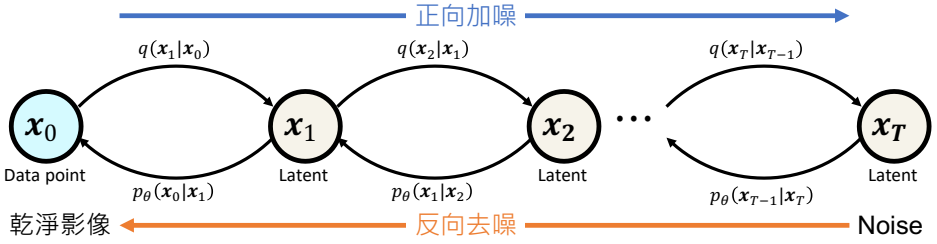
ELBO



- Reconstruction term $\mathcal{L}_0 = -\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$: The same as with VAEs; it is also negligible.

A. 因為 β_1 通常很小，所以 $\boldsymbol{x}_1 \approx \boldsymbol{x}_0$，重建變得微不足道，且相對於其他項很小

ELBO



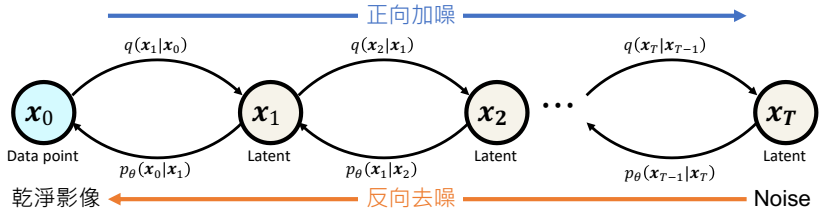
- Reconstruction term $\mathcal{L}_0 = -\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$: The same as with VAEs; it is also negligible.

- Prior matching term $\mathcal{L}_T = D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$: Converges to zero when $T \to \infty$.

- Denoising matching term $\mathcal{L}_{t-1}$

$$\mathcal{L}_{t-1} = \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

Denoising Matching Term $\mathcal{L}_{t-1}$



$$\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

$$= \mathbb{E}_{t>1, q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

How to model the variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$?

Variational Distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$$

Three options to define variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$:

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

   with the mean predictor $\mu_\theta(\boldsymbol{x}_t, t)$.

希望 variational distribution 的 mean 與 posterior distribution 的 mean 越像越好

Variational Distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$$

Three options to define variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$:

2. $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, t), \tilde{\sigma}_t^2 \mathbf{I}\right)$

   with the $\boldsymbol{x}_0$ predictor $\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, t)$.

Variational Distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2 \mathbf{I})$$

Three options to define variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$:

3. $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\varepsilon}}_t(\boldsymbol{x}_t, t)\right), \tilde{\sigma}_t^2 \mathbf{I}\right)$

   with the $\boldsymbol{\varepsilon}_t$ predictor $\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)$.

Variational Distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

1. Mean predictor $\mu_\theta(\boldsymbol{x}_t, t)$:

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

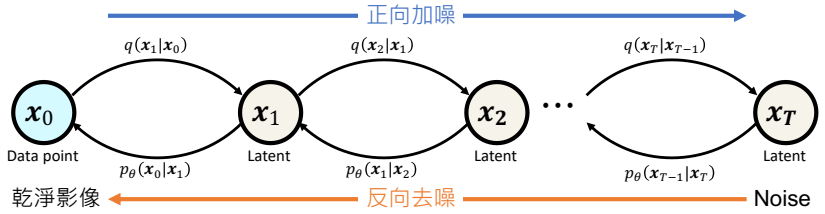
2. $\boldsymbol{x}_0$ predictor $\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, t)$:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, t), \tilde{\sigma}_t^2 \mathbf{I}\right)$$

3. $\boldsymbol{\varepsilon}_t$ predictor $\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)$:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)\right), \tilde{\sigma}_t^2 \mathbf{I}\right)$$

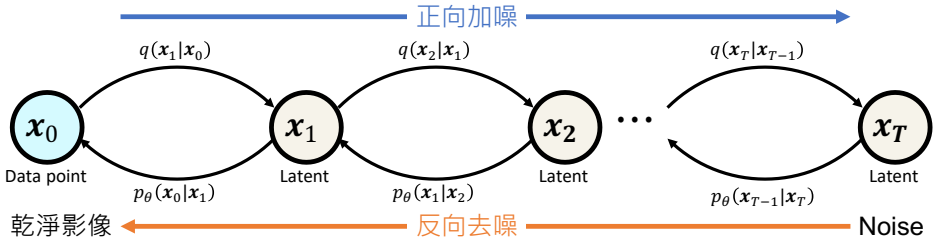
Denoising Matching Term $\mathcal{L}_{t-1}$



$$\mathbb{E}_{t>1,q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

How can we minimize the denoising matching term for each predictor?

Diffusion Matching Term



Rewrite the diffusion matching term for each case:

1. Mean predictor $\mu_\theta(\boldsymbol{x}_t, t)$:

$$\mathbb{E}_{t>1,q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\frac{1}{2\tilde{\sigma}_t^2}\|\mu_\theta(\boldsymbol{x}_t,t)-\tilde{\mu}(\boldsymbol{x}_t,\boldsymbol{x}_0)\|^2\right]$$

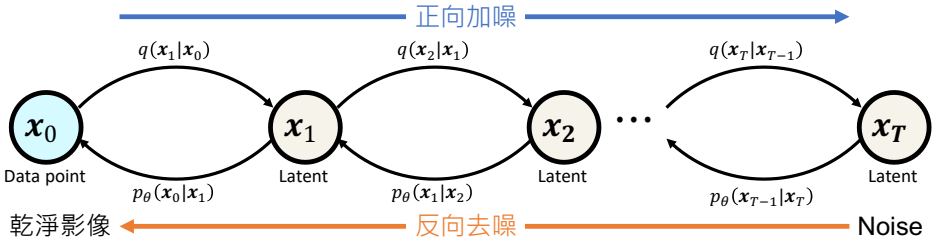
2. $\boldsymbol{x}_0$ predictor $\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, t)$:

$$\mathbb{E}_{t>1,q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\frac{1}{2\tilde{\sigma}_t^2}\omega_t\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|^2\right]$$

3. $\boldsymbol{\varepsilon}_t$ predictor $\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)$:

$$\mathbb{E}_{t>1,q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\frac{1}{2\tilde{\sigma}_t^2}\omega_t'\|\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{\varepsilon}_t\|^2\right]$$

Diffusion Matching Term



In practice, we can simply drop the weight term in training:

1. Mean predictor $\mu_\theta(\boldsymbol{x}_t, t)$:

$$\mathbb{E}_{t>1,q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\mu_\theta(\boldsymbol{x}_t, t) - \tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0)\|^2]$$

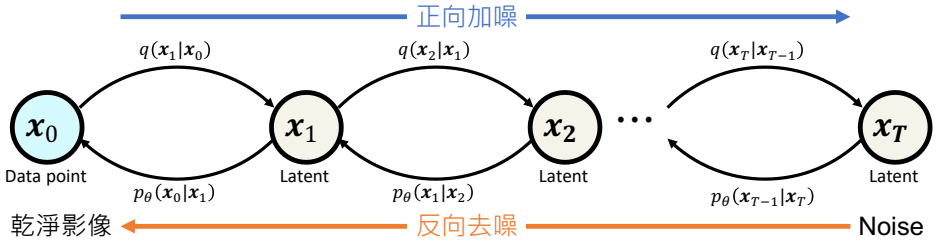
2. $\boldsymbol{x}_0$ predictor $\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, t)$:

$$\mathbb{E}_{t>1,q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\|^2]$$

3. $\boldsymbol{\varepsilon}_t$ predictor $\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)$:

$$\mathbb{E}_{t>1,q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\varepsilon}_t\|^2]$$

Diffusion Matching Term



In practice, we generally use the $\boldsymbol{\varepsilon}_t$ predictor $\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)$ since

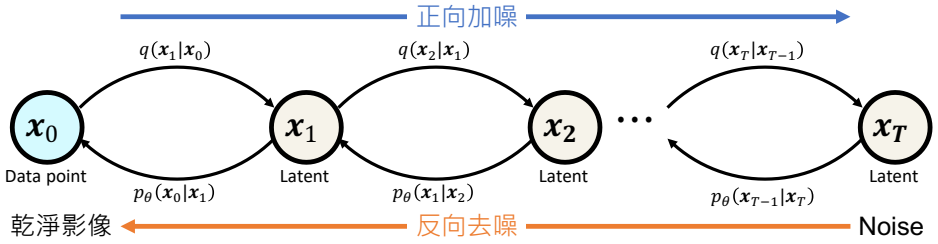
$\boldsymbol{\varepsilon}_t$ are standard normal samples

➔well normalized and scaled

➔which makes it easier to train a neural network.

數值穩定：噪音的範圍固定在 $\mathcal{N}(\mathbf{0}, \mathbf{I})$
效果最好：實驗證明這種參數化表現最佳

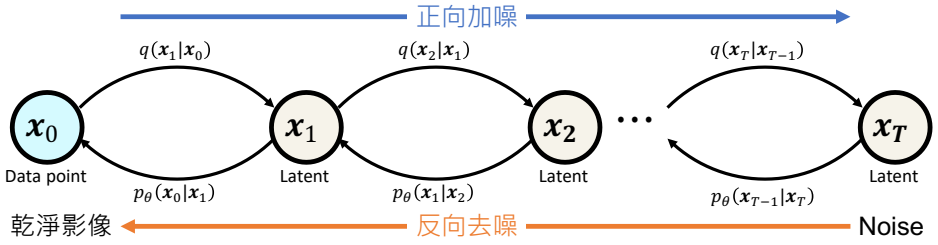
Training



How to train the $\boldsymbol{\varepsilon}_t$ predictor $\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)$ with the loss function:

$$\mathbb{E}_{t>1, q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\varepsilon}_t\|^2]$$

Training



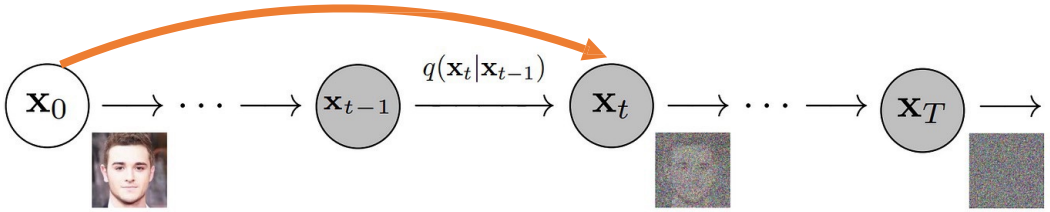
$$\mathbb{E}_{t>1,q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{\varepsilon}_t\|^2]$$

Repeat:

1. Take a random $\boldsymbol{x}_0$.
2. Sample $t\sim\mathcal{U}(\{1,\dots,T\})$.
3. Sample $\boldsymbol{\varepsilon}_t\sim\mathcal{N}(\mathbf{0},\mathbf{I})$.
4. Compute $\boldsymbol{x}_t=\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t$.
5. Take gradient decent step on $\nabla_\theta\|\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{\varepsilon}_t\|^2$.

Same as sampling

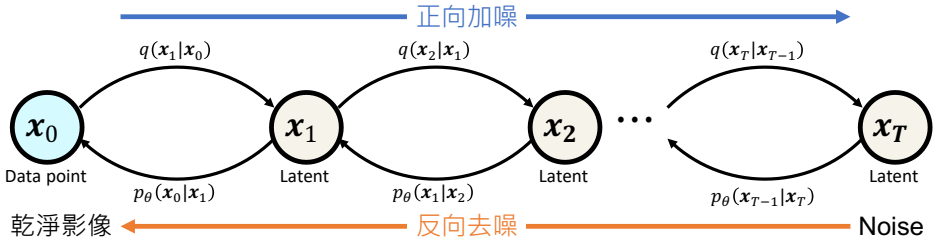
$\boldsymbol{x}_t\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)=\mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0,\sqrt{1-\bar{\alpha}_t}\mathbf{I}\right)$



Training data: 用噪音去淹沒圖片內容

Network 需要學會正確預測噪音

Training



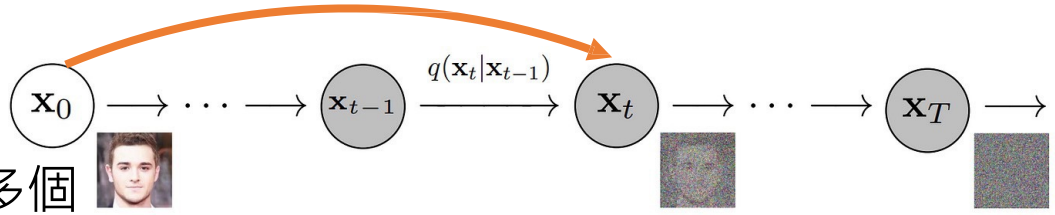
$$\mathbb{E}_{t>1,q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{\varepsilon}_t\|^2]$$

Repeat:

1. Take a random $\boldsymbol{x}_0$.
2. Sample $t\sim\mathcal{U}(\{1,\dots,T\})$.
3. Sample $\boldsymbol{\varepsilon}_t\sim\mathcal{N}(\mathbf{0},\mathbf{I})$.
4. Compute $\boldsymbol{x}_t=\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t$.
5. Take gradient decent step on $\nabla_\theta\|\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{\varepsilon}_t\|^2$.

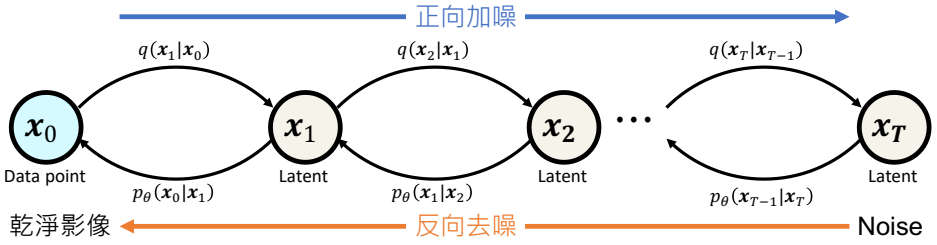
Same as sampling

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



並不是讓 network 去死背 noise，因為可能有很多個 $\boldsymbol{x}_0$ 在正向加噪跳躍會變成一樣的 $\boldsymbol{x}_t$，網絡學習的是條件期望的統計規律（最可能的去噪方向），而不是簡單的記憶一對一映射

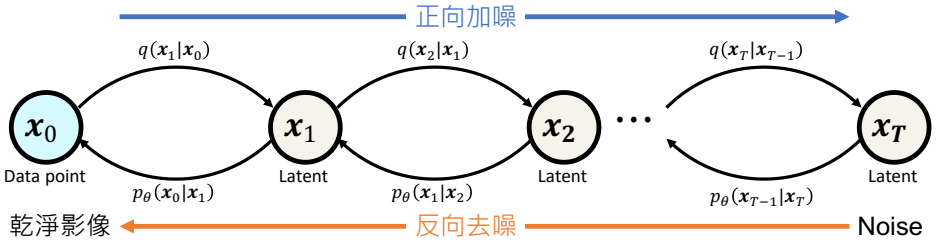
Reverse Process (Generation)



How to proceed the reverse process with the learned reverse transitional distribution:

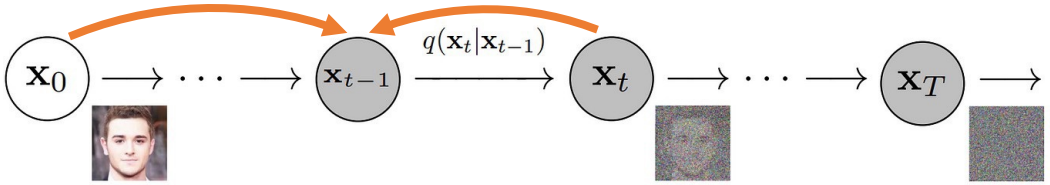
$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)\right), \tilde{\sigma}_t^2\mathbf{I}\right)$$

Reverse Process (Generation)

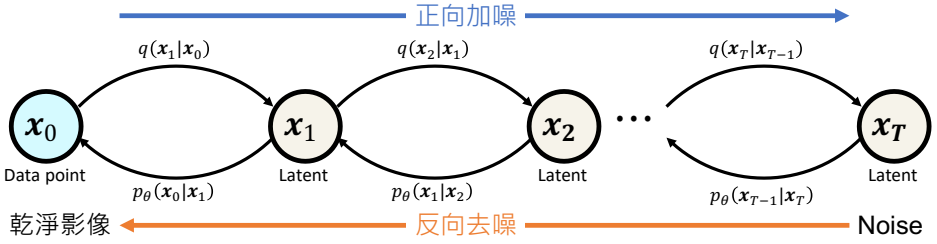


1. Sample $\boldsymbol{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
2. For $t = T, \ldots, 1$, repeat:
   1. Compute $\tilde{\mu} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)\right)$.
   2. Sample $\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
   3. Compute $\boldsymbol{x}_{t-1} = \tilde{\mu} + \tilde{\sigma}\boldsymbol{z}_t$.

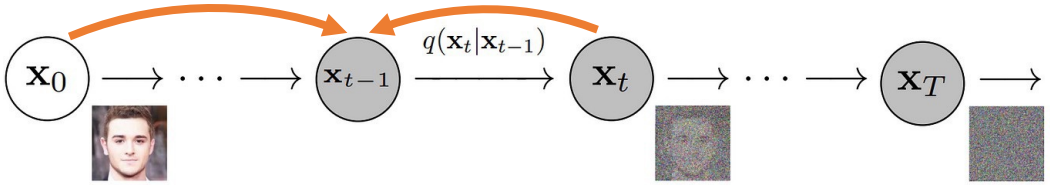
Same as sampling $\boldsymbol{x}_t \sim p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$.



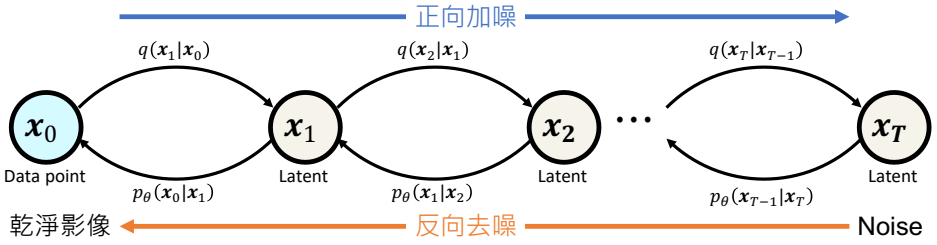
Reverse Process (Generation)



1. Sample $\boldsymbol{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
2. For $t = T, \ldots, 1$, repeat:
    1. Compute $\tilde{\mu} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)\right)$.
    2. Sample $\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
    3. Compute $\boldsymbol{x}_{t-1} = \tilde{\mu} + \tilde{\sigma}\boldsymbol{z}_t$.



Reverse Process (Generation)

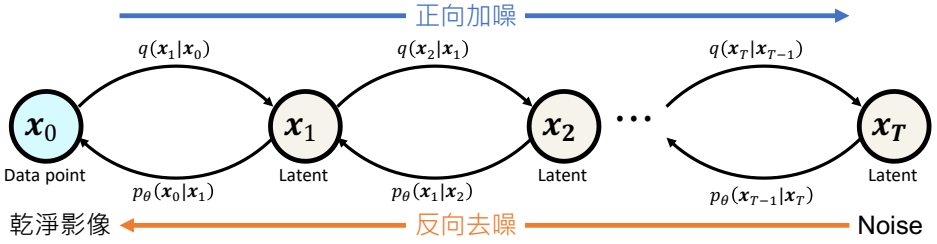


At each timestep t, given $\boldsymbol{x}_t$,

- $\boldsymbol{\varepsilon}_t$ is predicted.
- The prediction of $\boldsymbol{x}_0$ can be computed from $\boldsymbol{x}_t$ and $\boldsymbol{\varepsilon}_t$.

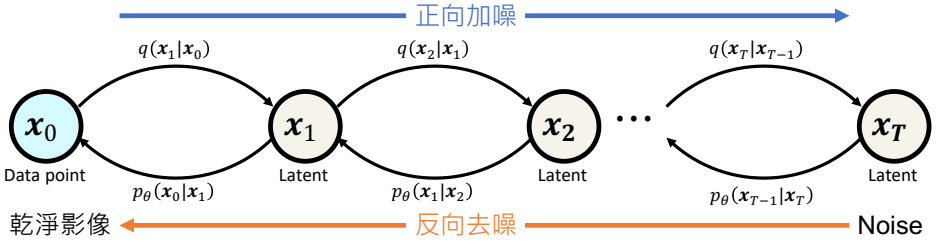
The denoising (reverse) process can also be viewed as a refinement process!

Reverse Process (Generation)



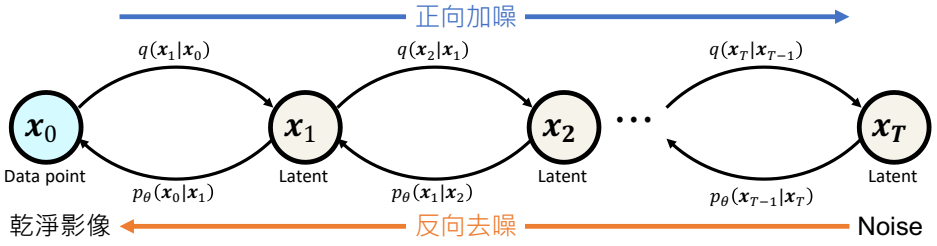
Q. Given $\boldsymbol{x}_t$ and the noise prediction $\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)$, what is the prediction of $\boldsymbol{x}_0$ (which will be denoted as $\boldsymbol{x}_{0|t}$)?

Reverse Process (Generation)



$$\textbf{A. } \boldsymbol{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)\right)$$

Reverse Process (Generation)



Reverse Process with ε_t Predictor

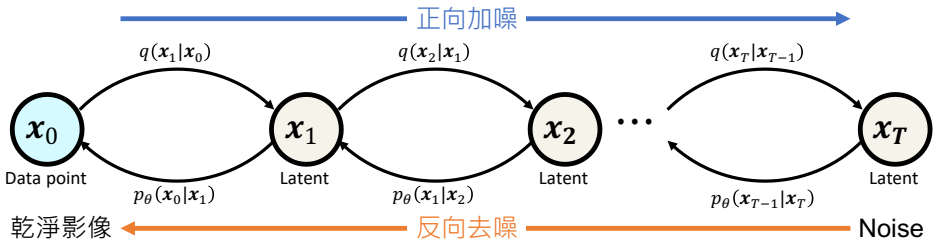
1. Sample $\boldsymbol{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
2. For $t = T, \ldots, 1$, repeat:
   1. Compute $\boldsymbol{x}_{0|t}$ from $\boldsymbol{x}_t$ and $\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t)$.
   2. Compute $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_{0|t})$
   3. Sample $\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
   4. Compute $\boldsymbol{x}_{t-1} = \tilde{\mu} + \tilde{\sigma}\boldsymbol{z}_t$.

跟前面的 reverse process 不太一樣
得到 noise prediction $\hat{\boldsymbol{\varepsilon}}_\theta$ 之後，不是直接扣掉
而是 (1) 先用 $\boldsymbol{x}_t$ 跟 $\hat{\boldsymbol{\varepsilon}}_\theta$ 外推出 $\boldsymbol{x}_{0|t}$
再 (2) 計算出 mean of the likelihood distribution $\tilde{\mu}$

Same as sampling
$\boldsymbol{x}_t \sim p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$.

這就是他的 mean!
不要被下標 $t, t-1$ 混淆了

Recap of All the Steps



ELBO

Decompose the negative ELBO in a different way:

$-\log p(\boldsymbol{x}_0) = -\log \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[-\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$

$= \cdots =$

$-\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$ — Reconstruction term $\mathcal{L}_0$

$+D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$ — New prior matching term $\mathcal{L}_T$

$+\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$ — Denoising matching term $\mathcal{L}_{t-1}$

Denoising Matching Term $\mathcal{L}_{t-1}$

$$\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

The variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ should be close to $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ for each t.

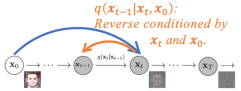
$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$

- The mean $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0$ is a function of both $\boldsymbol{x}_t$ and $\boldsymbol{x}_0$.

  $\boldsymbol{x}_{t-1}$ 的 mean $\tilde{\mu}$ 是 $\boldsymbol{x}_t$ 與 $\boldsymbol{x}_0$ 的線性內差

- The covariance $\tilde{\sigma}_t^2\mathbf{I} = \left(\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\right)\mathbf{I}$ is predefined from the user-defined $\{\beta_t\}_{t=1}^T$.

$\boldsymbol{x}_{t-1}$ 的 variance $\tilde{\sigma}_t^2$ 與 $\boldsymbol{x}_t$ 或 $\boldsymbol{x}_0$ 無關，僅由 time step t 與 β_t 預先定義



Denoising Diffusion Probabilistic Models (DDPM)

1. $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})$
   where $\alpha_t = 1 - \beta_t$.
   正向加噪過程，是我們精心定義的
2. $q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$
   where $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$.
   正向跳躍，串起來推導可得
3. $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2\mathbf{I})$
   where $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0$ and $\tilde{\sigma}_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

$\boldsymbol{x}_{t-1}$ 是 Gaussian, mean 是 $\boldsymbol{x}_t$ 與 $\boldsymbol{x}_0$ 的線性內差，variance 由 time step t 與 β_t 預先定義

Denoising Diffusion Probabilistic Models (DDPM)

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\sigma}_t^2\mathbf{I})$$

1. $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0$ or
   mean 可以由 $\boldsymbol{x}_t$ 與 $\boldsymbol{x}_0$ 線性內差得到
2. $\tilde{\mu}(\boldsymbol{x}_t, \boldsymbol{\varepsilon}_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_t\right)$

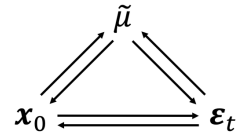
   where $\boldsymbol{\varepsilon}_t = \frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{x}_0$

   $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_t$

mean 也可以由 $\boldsymbol{x}_t$ 與 noise $\boldsymbol{\varepsilon}_t$ 得到

Denoising Diffusion Probabilistic Models (DDPM)

Given $\boldsymbol{x}_t$, each of $\tilde{\mu}$, $\boldsymbol{x}_0$, and $\boldsymbol{\varepsilon}_t$ can be computed from any of the others.



Direct Sampling vs. Generative Models

Q. Why don't we directly sample from the training data instead of training a generative model?

Direct Sampling vs. Generative Models

- **Efficiency and Scalability:**
  Generative models can efficiently encode the data distribution into a neural network, eliminating the need to store the entire dataset.

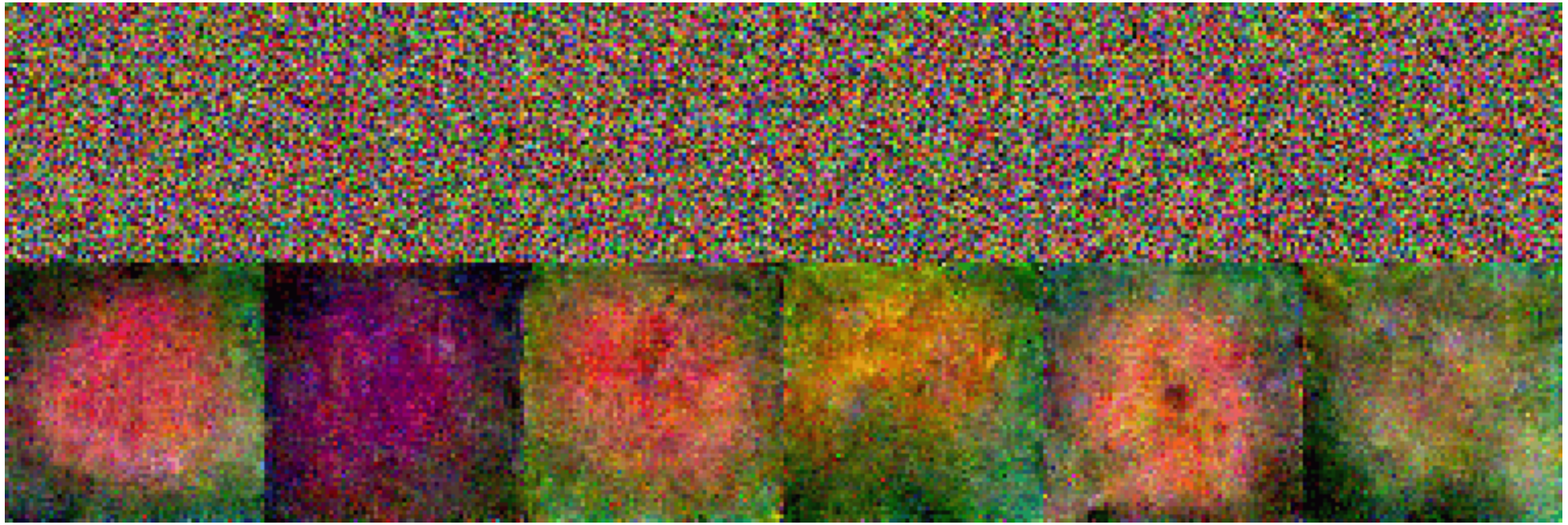
- **Novelty and Diversity / Privacy and Security Concerns:**
  Generative models can generate new samples that are not identical to any specific training data but still appear plausible.

- **Adaptability to New Conditions:**
  For conditional generation, generative models can be more efficient at generating data based on the given conditions compared to retrieval methods.

Reverse Process (Generation)

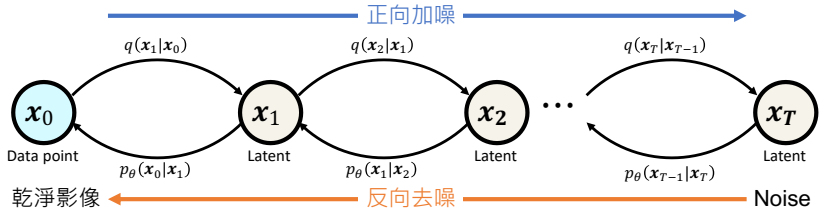
$\boldsymbol{x}_t$

$\boldsymbol{x}_0$ Prediction



在每個 time step t，network 都在嘗試直接預測 $\boldsymbol{x}_0$
不是盲目地去噪，而是基於對最終目標的理解這種「**全局視野**」，每一步都在糾正對原始圖像的估計，讓模型能更好地保持圖像的結構

Denoising Matching Term $\mathcal{L}_{t-1}$



$$\sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

$$= \mathbb{E}_{t>1, q(\boldsymbol{x}_t|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

How to model the variational distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$?

1st Assignment