DDPM sur MNIST

Arthur Hamard, Arthur Guelennoc, Léo Sambrook, Etienne Senigout,
Pierre Braud, Paul Pazart

Introduction

Le Projet en un Clin d'œil

- **Maîtriser le cycle :** Destruction ↔ Reconstruction
- **Technique :** Denoising Diffusion Probabilistic Models (DDPM)
- **Dataset :** MNIST (60k chiffres)

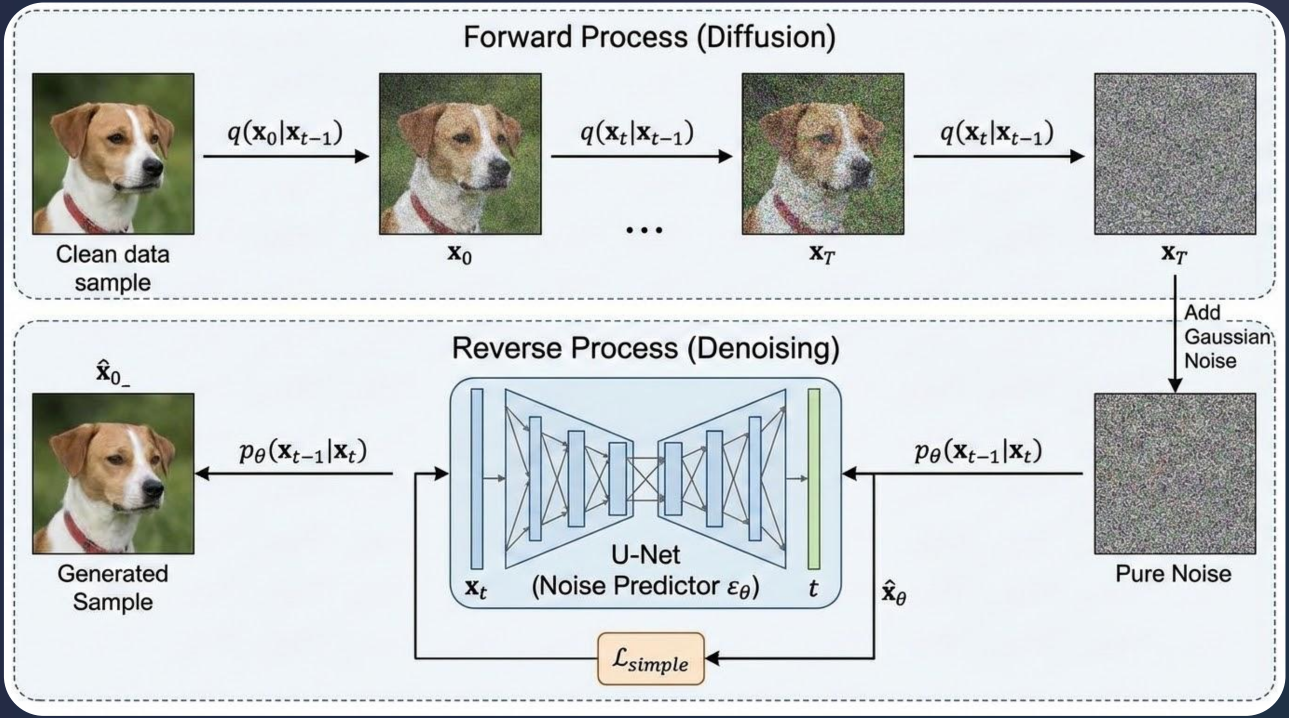
Partie 1 : Théorie - Le Principe

Forward vs Reverse

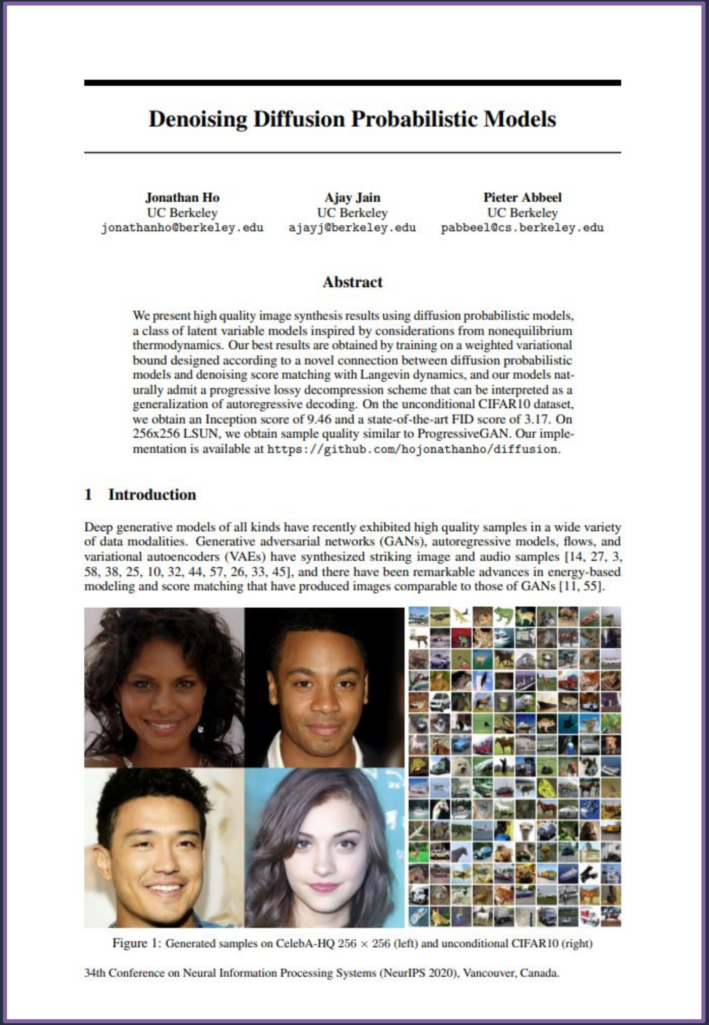
L'idée centrale est de détruire l'information progressivement pour ensuite apprendre à la reconstruire.

- **Processus Forward (q) :**
  Une chaîne de Markov fixe qui ajoute du bruit gaussien. À T, l'image est un bruit pur N(0, 1).

- **Processus Reverse (pθ) :**
  Un réseau de neurones apprend à inverser ce processus, étape par étape.



Formulation Mathématique : Intro (accrochez vous!)



Denoising Diffusion Probabilistic Models

Jonathan Ho UC Berkeley jonathanho@berkeley.edu

Ajay Jain UC Berkeley ajayj@berkeley.edu

Pieter Abbeel UC Berkeley pabbeel@cs.berkeley.edu

Abstract

We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at https://github.com/hojonathanho/diffusion.

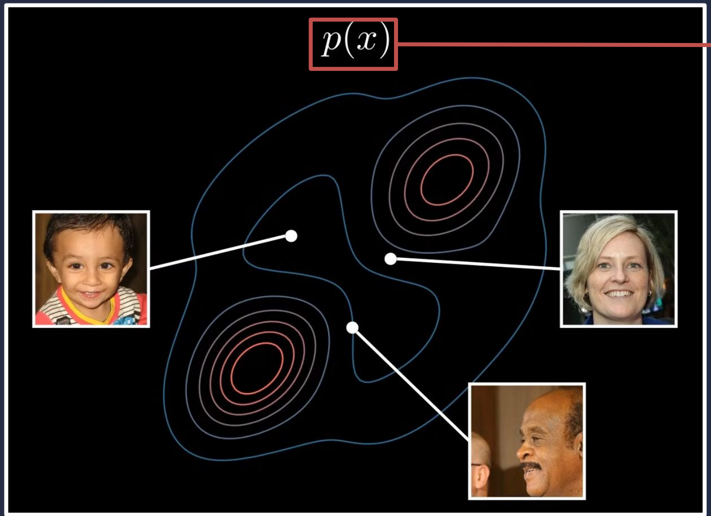
1 Introduction

Deep generative models of all kinds have recently exhibited high quality samples in a wide variety of data modalities. Generative adversarial networks (GANs), autoregressive models, flows, and variational autoencoders (VAEs) have synthesized striking image and audio samples [14, 27, 3, 58, 38, 25, 10, 32, 44, 57, 26, 33, 45], and there have been remarkable advances in energy-based modeling and score matching that have produced images comparable to those of GANs [11, 55].

Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

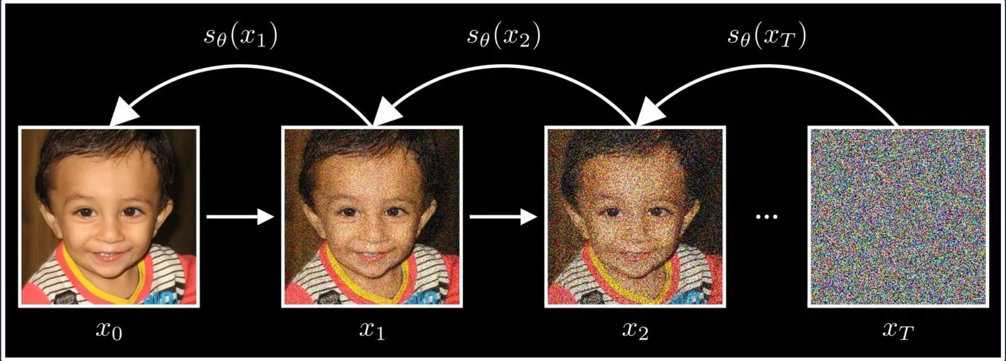
34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

- a “introduit” la diffusion
- 20K+ citations



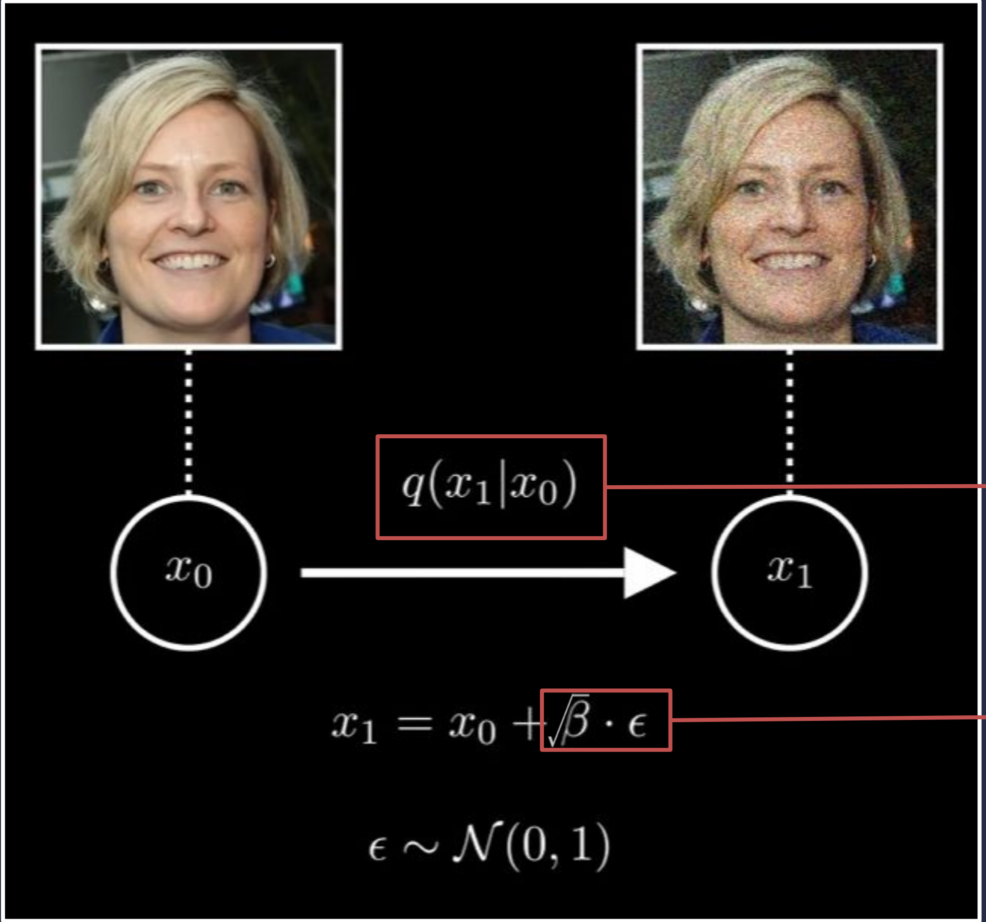
distribution des données

- dataset d’images
- *p* inconnu
- *p*(x) ~ plausibilité de l’image
- image du dataset ⇒ *p*(x) élevé
- image absurde ⇒ *p*(x) petit
- générer images similaires



- modélisation impossible
- corruption graduelle
- bruit pur
- x0, x1, ... nos images
- reverse
- forward = +, backward = -

Formulation Mathématique : Forward



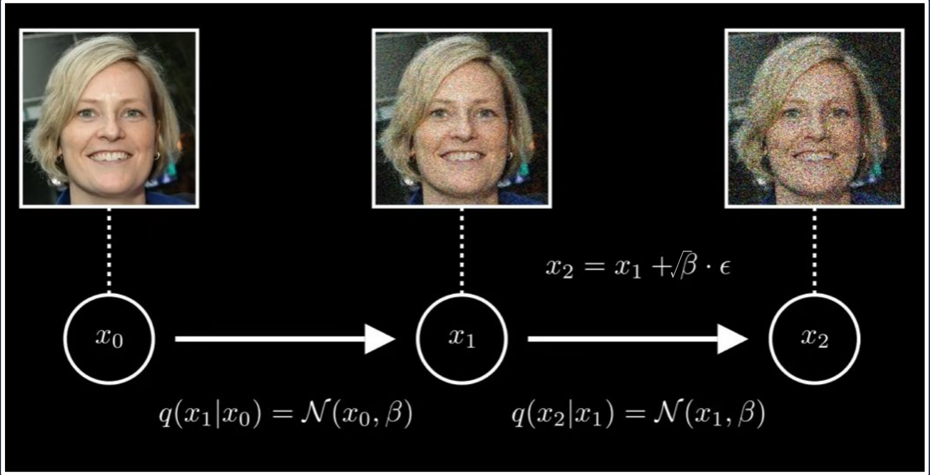
distribution conditionnelle

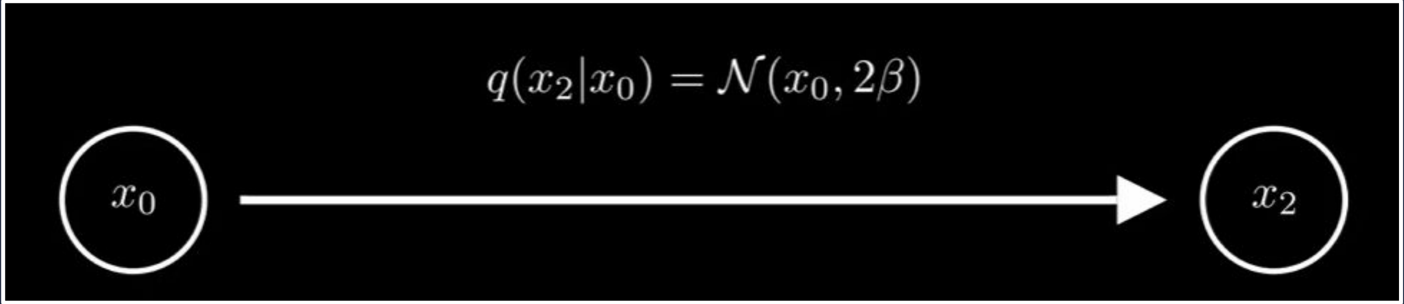
bruit Gaussien contrôlé par un scalaire β

$$q(x_1|x_0) = \mathcal{N}(x_0, \beta)$$

distrib de *x1* = Gaussienne centrée sur *x0*, bruit de variance β

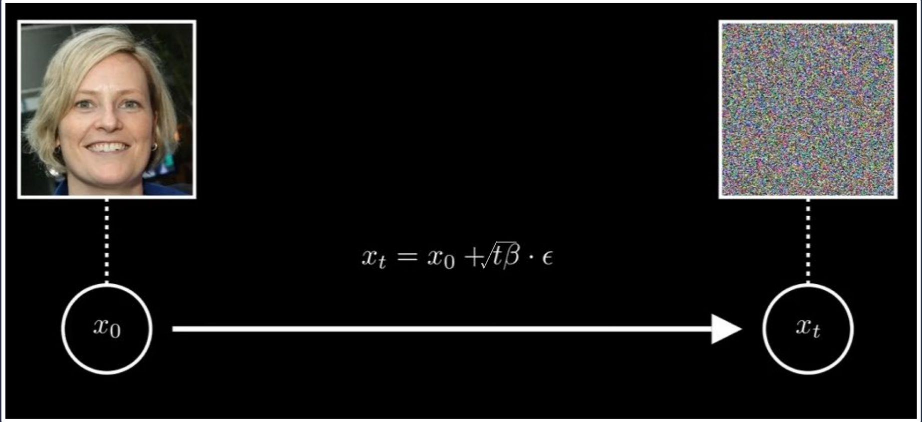
Formulation Mathématique : Forward





- pareil que sampler la Gaussienne de variance 2β car $x_2 = x_0 + \sqrt{\beta}\,\varepsilon_1 + \sqrt{\beta}\,\varepsilon_2$ $\quad x_2 = x_0 + \sqrt{\beta}\,(\varepsilon_1 + \varepsilon_2)$ or $\varepsilon_1 + \varepsilon_2 \sim \mathcal{N}(0, 2)$
  $x_2 = x_0 + \sqrt{2\beta}\,\varepsilon$

- on répète l'opération



- $\sqrt{(t*\beta)}$ de bruit
- aka $q(x_t \mid x_0) = \mathcal{N}(x_0, t\beta)$ à t,

- $q(xt \mid x0)$ =
  distribution Gaussienne centrée sur $x0$, de variance $t*\beta$

Formulation Mathématique : Forward (2)

$$q(x_t \mid x_0) = \mathcal{N}(x_0, t\beta)$$

$$q(x_t|x_0) \xrightarrow[t \to \infty]{} \mathcal{N}(0,1)$$

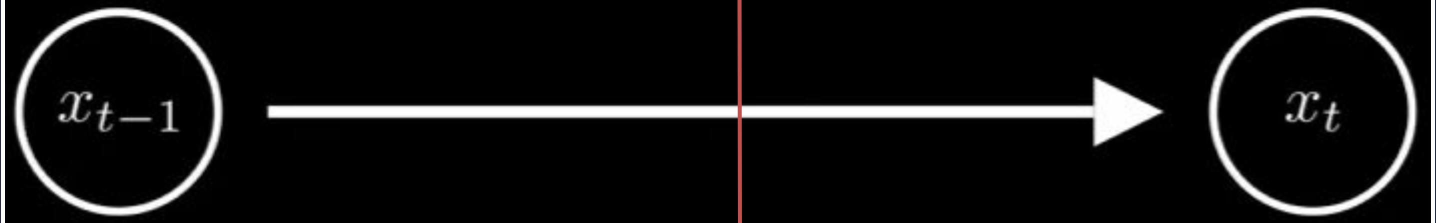
Explosion de la variance de diffusion

Simplification des calculs

- Forward
- Backward
- Base de génération

Formulation Mathématique : Forward (2)

$$q(x_t|x_{t-1}) = \boxed{\sqrt{1-\beta}}\, x_{t-1} + \sqrt{\beta} \cdot \epsilon$$



coefficient pour diminuer x

pourquoi ?

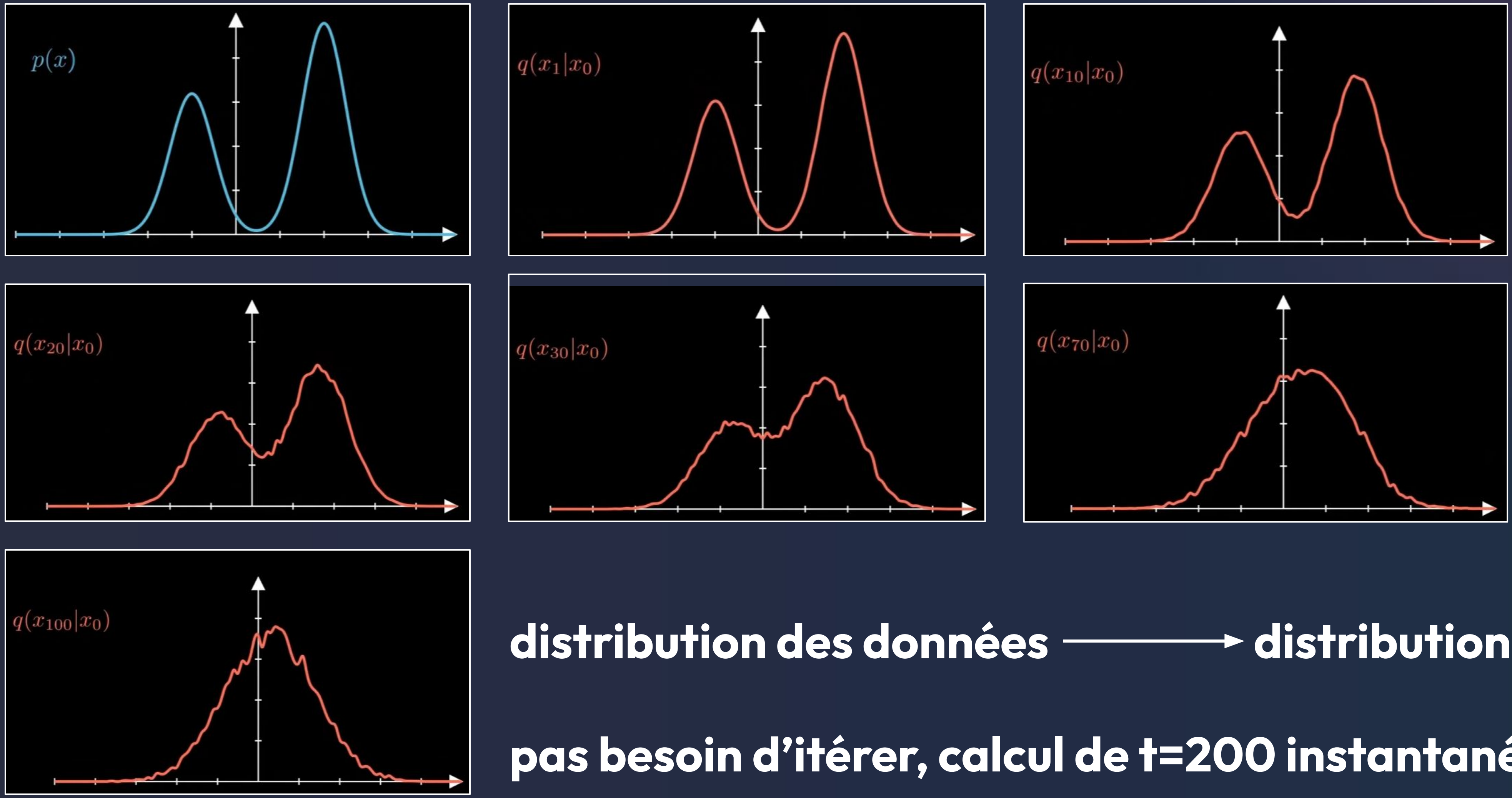
- auteurs de DDPM
- non-expliqué
- simple

$$q(x_t|x_0) = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{(1-\bar{\alpha}_t)} \cdot \epsilon$$

$$\boxed{\bar{\alpha}_t = (1-\beta)^t}$$

moyenne → **0**
variance → **1**
temps → **∞**

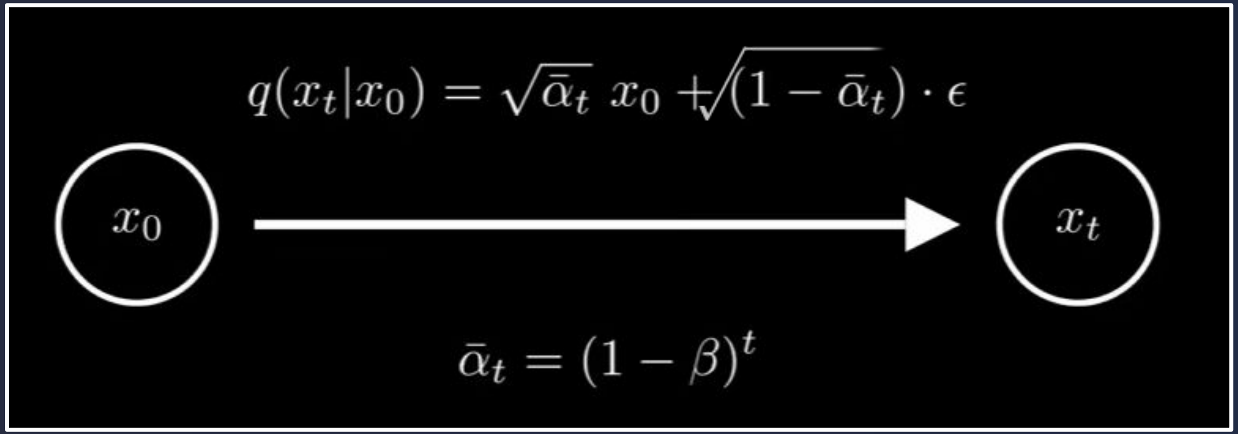
Formulation Mathématique : Forward (2)

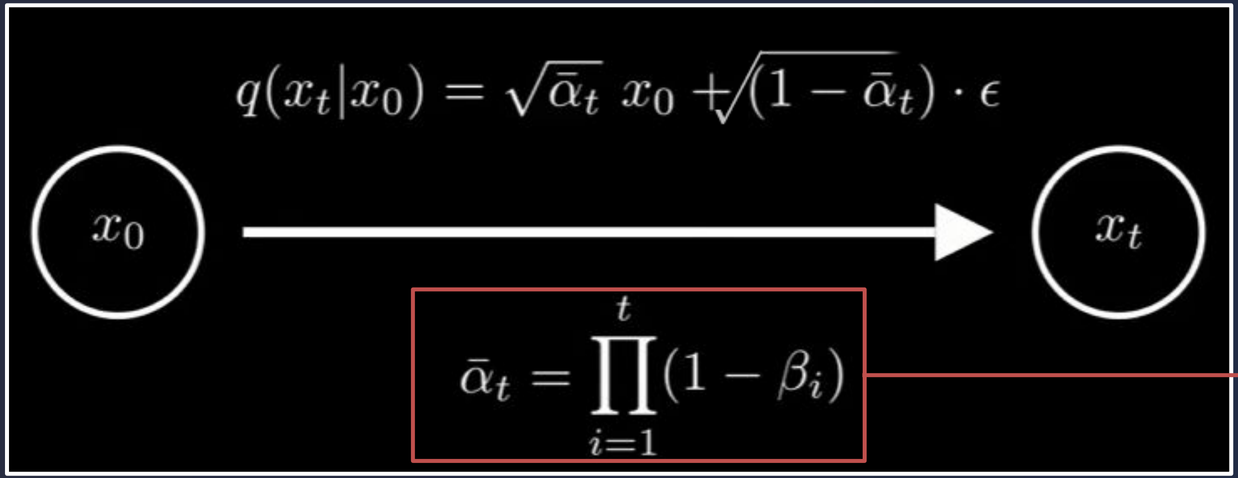


distribution des données $\longrightarrow$ **distribution centrée réduite**

pas besoin d'itérer, calcul de t=200 instantané

Formulation Mathématique : Forward (2)



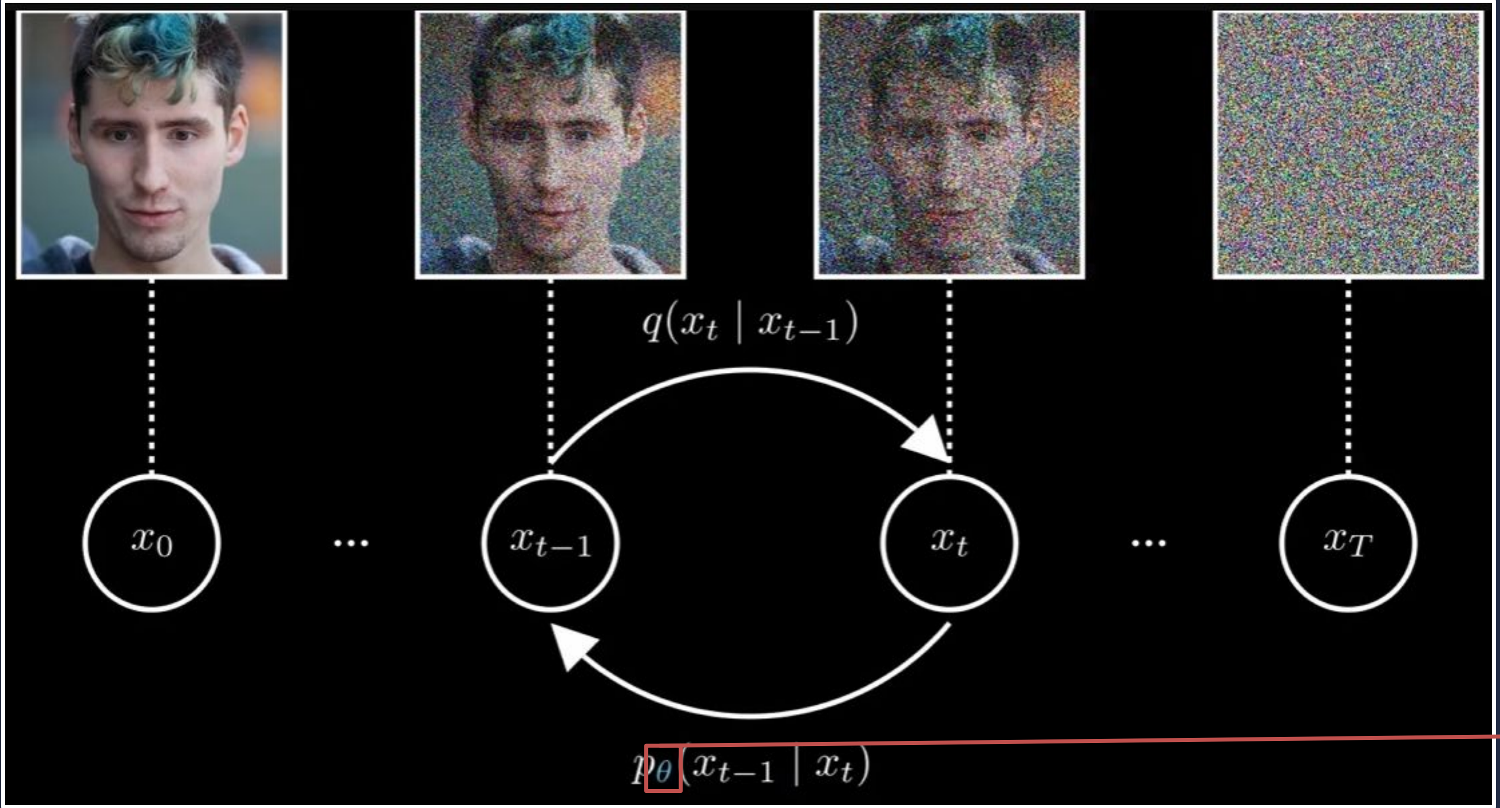


- β n'est pas fixe
- β scheduler
- 0 < β < 1

expression du papier

Formulation Mathématique : Backward

Comment inverser le procédé ?

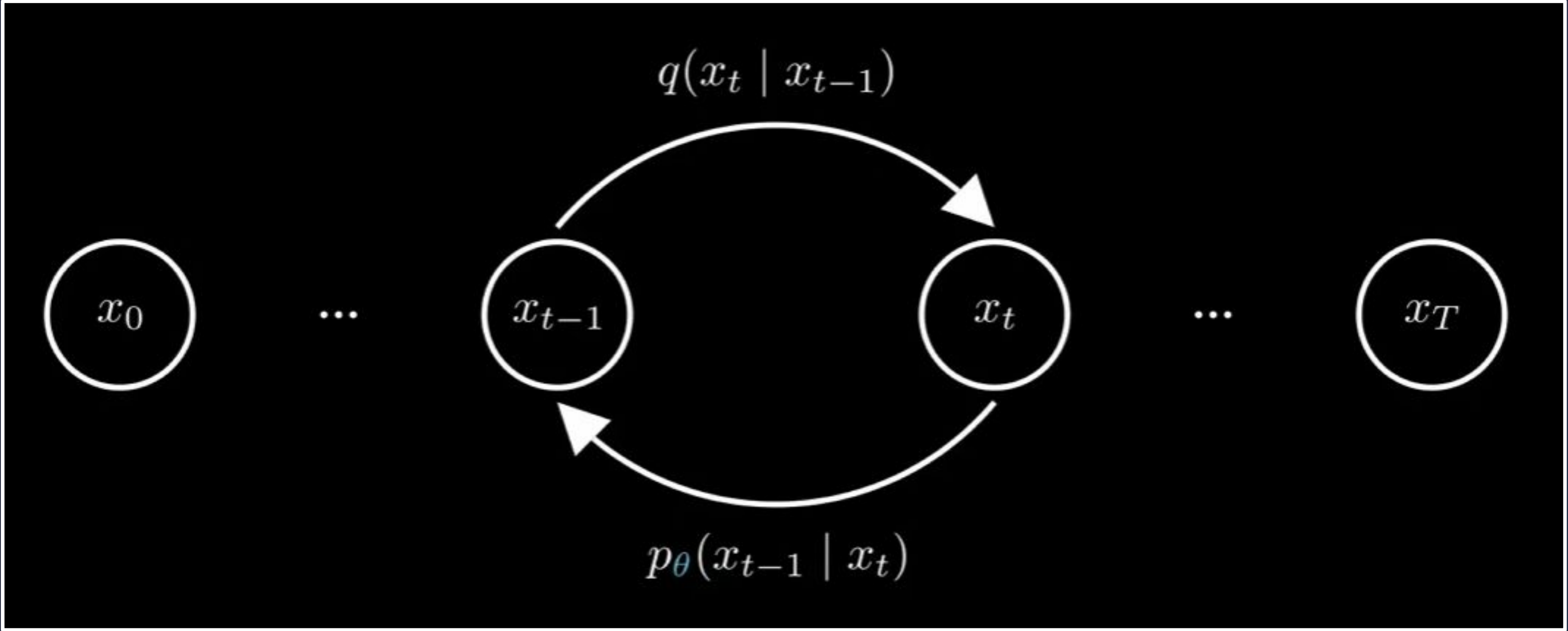


paramètres fixe en forward

mais en backward

on veut le meilleur θ
= poids du réseau

Formulation Mathématique : Backward (entraînement)



approche bayésienne

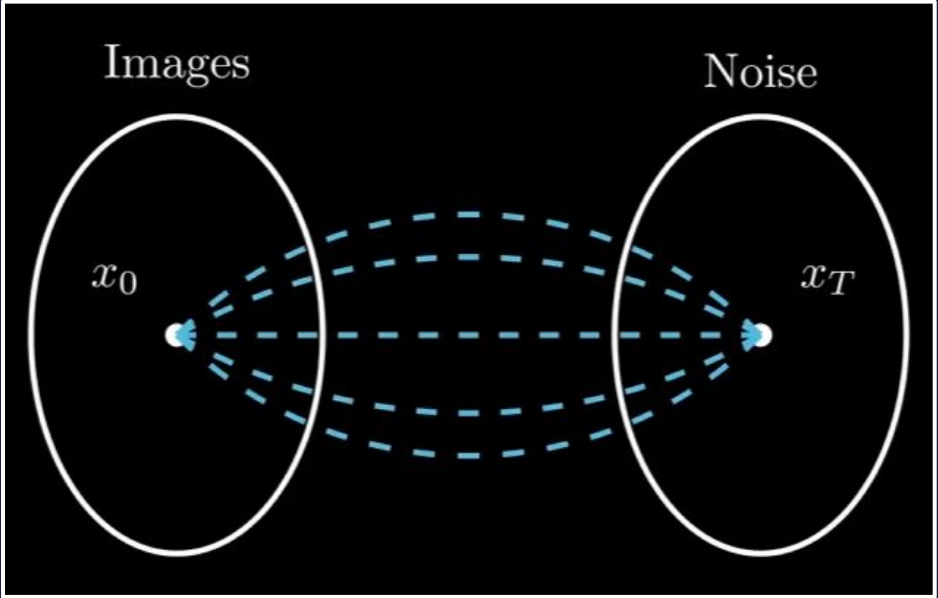
on veut minimiser $-\log p_\theta(x_0)$

negative log-likelihood

Formulation Mathématique : Loss

$$-\log p_\theta(x_0) = -\log \int p_\theta(x_{0:T}) dx_{1:T}$$

intégrale de la probabilité conjointe sur toutes les variables



“chemins” qui connectent image à leur bruit respectif

negative log likelihood nécessite addition de tous ces chemins possibles

beaucoup trop de chemins possibles

Formulation Mathématique : Loss (2)

$$-\log p_\theta(x_0) = -\log \int p_\theta(x_{0:T}) dx_{1:T}$$

multiplication + division par la densité
du procédé forward

$$-\log p_\theta(x_0) = -\log \int q(x_{1:T} \mid x_0) \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)} dx_{1:T}$$

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

inégalité de Jensen
negative log convexe

$$-\log p_\theta(x_0) \leq -\mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)} \right]$$

ELBO
Evidence Lower Bound

$$-\mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)} \right]$$

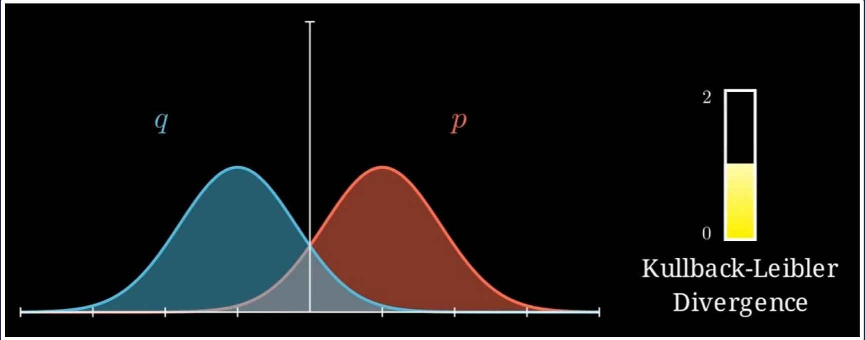
Formulation Mathématique : Loss (3)

$$-\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)}\right]$$

calculs expliqués en annexe du papier

$$-\mathbb{E}_q\left[D_{\mathrm{KL}}(q(x_T \mid x_0) \,||\, p(x_T)) + \sum_{t>1} \boxed{D_{\mathrm{KL}}}(q(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)) - \log p_\theta(x_0 \mid x_1)\right]$$

Kullback-Leibler divergence



distance entre
distributions de probabilité

quand 0 ~ pareil

Formulation Mathématique : Loss (4)

$$-\mathbb{E}_q\left[\overbrace{\boxed{D_{\mathrm{KL}}(q(x_T \mid x_0) \,||\, p(x_T))}}^{\text{ne dépends pas des paramètres}} + \sum_{t>1} D_{\mathrm{KL}}(q(x_{t-1} \mid x_t, x_0) \,||\, p_\theta(x_{t-1} \mid x_t)) - \overbrace{\boxed{\log p_\theta(x_0 \mid x_1)}}^{\text{Négligeable}}\right]$$

Formulation Mathématique : Loss (5)

$$-\mathbb{E}_q\left[\sum_{t>1} D_{\mathrm{KL}}\left(\boxed{q(x_{t-1} \mid x_t, x_0)} \,\|\, \boxed{p_\theta(x_{t-1} \mid x_t)}\right)\right]$$

vrai posterior

$$\begin{aligned} q(x_{t-1} \mid x_t, x_0) &\propto q(x_t \mid x_{t-1})\, q(x_{t-1} \mid x_0) \\ &= \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I\right) \end{aligned}$$

posterior approximé

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta, \sigma_t)$$

$$\tilde{\beta}_t = \sigma_t$$

$$\begin{aligned} D_{\mathrm{KL}}\left(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\right) &= D_{\mathrm{KL}}\left(\mathcal{N}(x_{t-1}; \tilde{\mu}_t, \tilde{\beta}_t I) \,\|\, \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \tilde{\beta}_t I)\right) \\ &= \frac{1}{2\tilde{\beta}_t}\left\|\tilde{\mu}_t - \mu_\theta\right\|^2 \end{aligned}$$

Formulation Mathématique : Loss (6)

$$L = \mathbb{E}_{x_0,\epsilon}\left[\sum_{t=1}^{T} \frac{1}{2\tilde{\beta}_t} \left\| \tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t) \right\|^2\right]$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

$$\tilde{\mu}_t(x_t, \epsilon) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon\right)$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(x_t, t)\right)$$

Formulation Mathématique : Recap

$$-\log p_\theta(x_0) = -\log \int p_\theta(x_{0:T})dx_{1:T}$$

$$-\log p_\theta(x_0) \leq -\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)}\right]$$

$$L = \mathbb{E}_{x_0,\epsilon}\left[D_{KL}(q(x_T \mid x_0)\|p(x_T)) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1} \mid x_t, x_0)\|p_\theta(x_{t-1} \mid x_t)) - \log p_\theta(x_0 \mid x_1)\right]$$

$$L = \mathbb{E}_{x_0,\epsilon}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\Big(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t)\Big)\right]$$

$$L = \mathbb{E}_{x_0,\epsilon}\left[\sum_{t=1}^{T} \frac{1}{2\tilde{\beta}_t}\left\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\right\|^2\right]$$

Formulation Mathématique : Recap

$$L = \mathbb{E}_{x_0,\epsilon}\Big[\sum_{t=1}^{T} \frac{1}{2\tilde{\beta}_t} \Big\| \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\Big) - \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\Big)\Big\|^2\Big]$$

$$L = \mathbb{E}_{x_0,\epsilon}\Big[\sum_{t=1}^{T} \frac{1}{2\tilde{\beta}_t\alpha_t} \Big\| \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\left(\epsilon - \epsilon_\theta(x_t,t)\right)\Big\|^2\Big]$$

$$L = \mathbb{E}_{x_0,\epsilon}\Big[\sum_{t=1}^{T} \frac{\beta_t^2}{2\tilde{\beta}_t\alpha_t(1-\bar{\alpha}_t)} \Big\|\epsilon - \epsilon_\theta(x_t,t)\Big\|^2\Big]$$

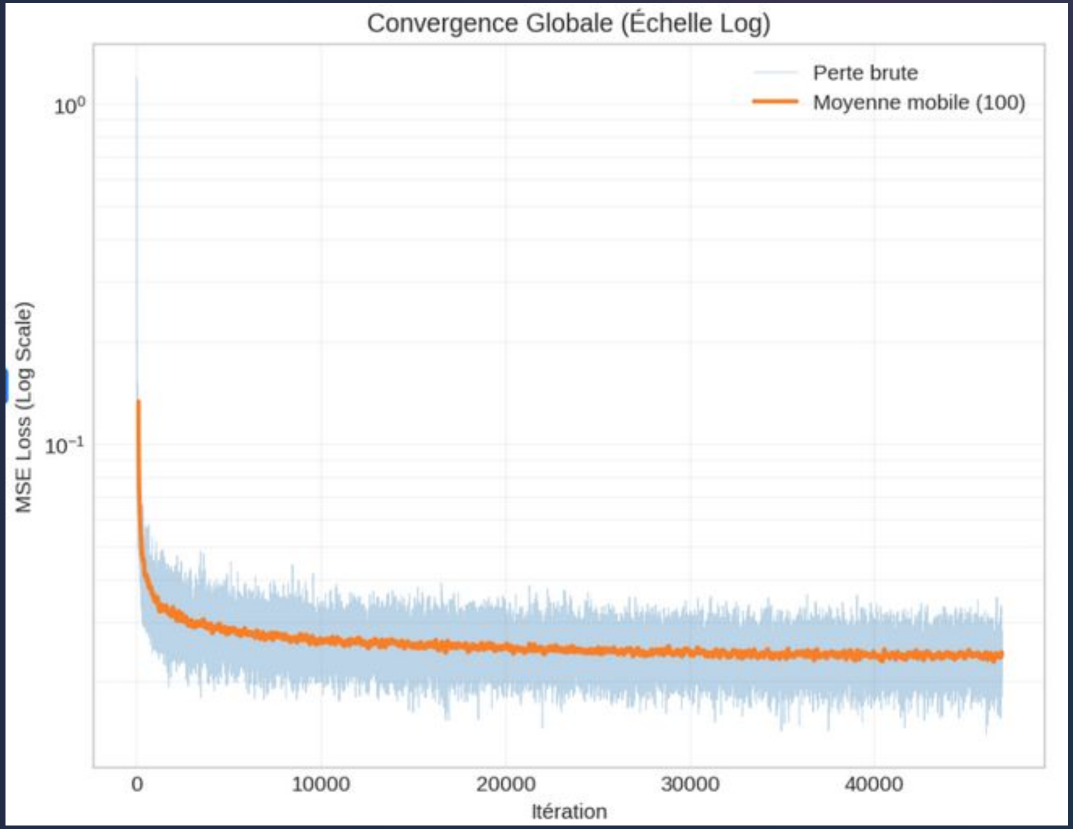
$$L \propto \mathbb{E}_{x_0,\epsilon}\Big[\sum_{t=1}^{T} \Big\|\epsilon - \epsilon_\theta(x_t,t)\Big\|^2\Big]$$

$$L = \mathbb{E}_{x_0,\epsilon,t}\Big[\|\epsilon - \epsilon_\theta(x_t,t)\|^2\Big], \quad t \sim \text{Uniform}\{1,\dots,T\}$$

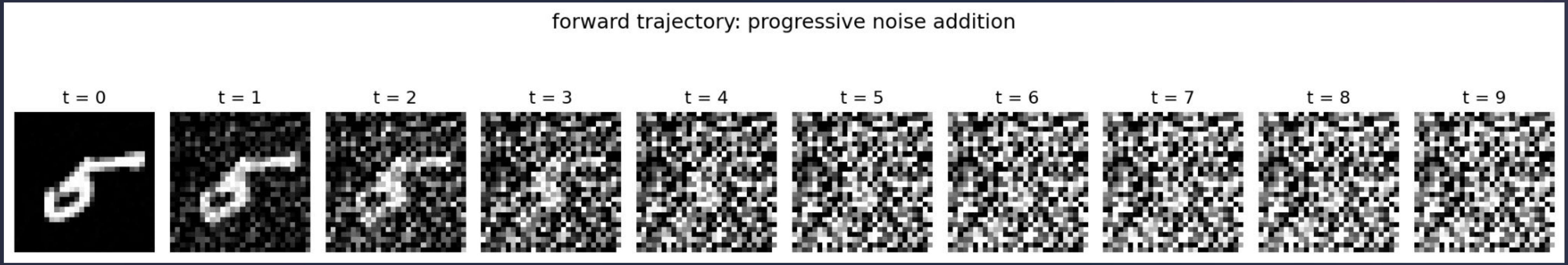
L'Entraînement (Context & Loss)

Convergence et Stabilité

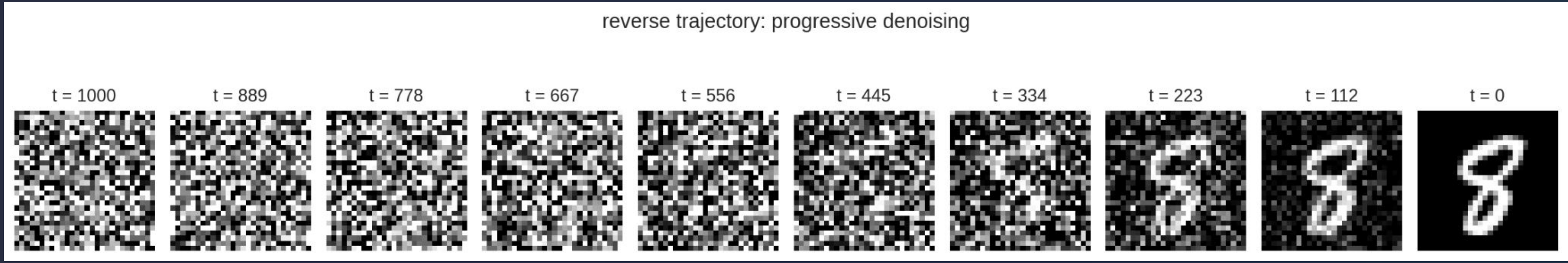
- **Architecture :** U-Net Light (~1.7M params)
- **Dataset :** MNIST (28x28)
- **Steps de Diffusion :** T=1000



Le Cycle de Diffusion : Destruction & Reconstruction



Forward Process (q) : Ajout de bruit



Reverse Process (pθ) : Génération

Preuve de Qualité (Métriques)

Évaluation réalisée sur **1000 échantillons** générés.

22.58

FID SCORE

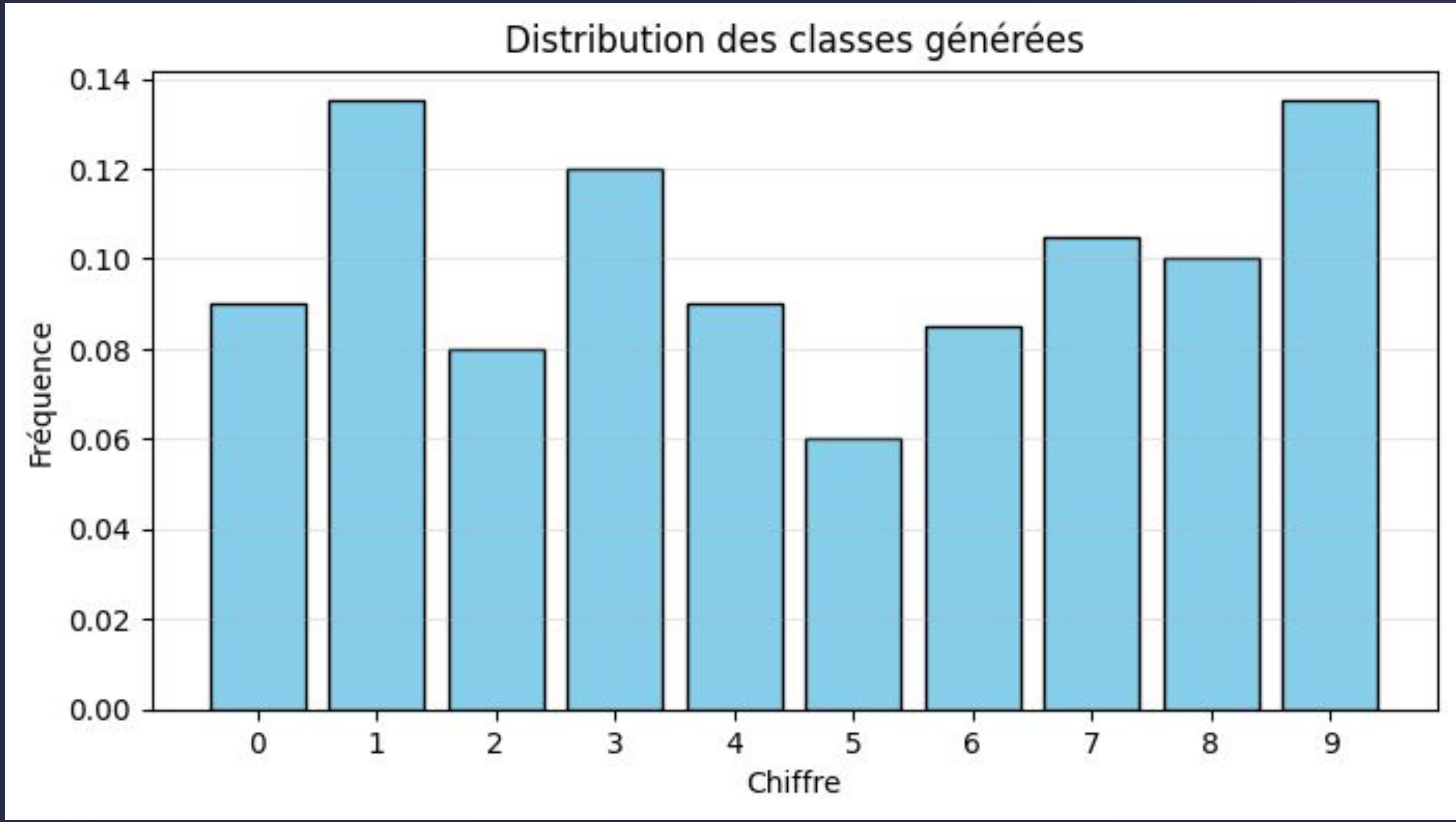
Similarité perceptuelle avec MNIST.

93.05

CONFIANCE MOYENNE

Les chiffres sont clairs et non ambigus.

Preuve de Qualité (Métriques)



Pas de "Mode Collapse" majeur. Toutes les classes sont générées.

DEMO

Conclusion

- **Objectif atteint :** Implémentation fonctionnelle.
- **Performance :** Entraînement stable, FID ~22.
- **Futur :** Explorer DDIM et Latent diffusion.