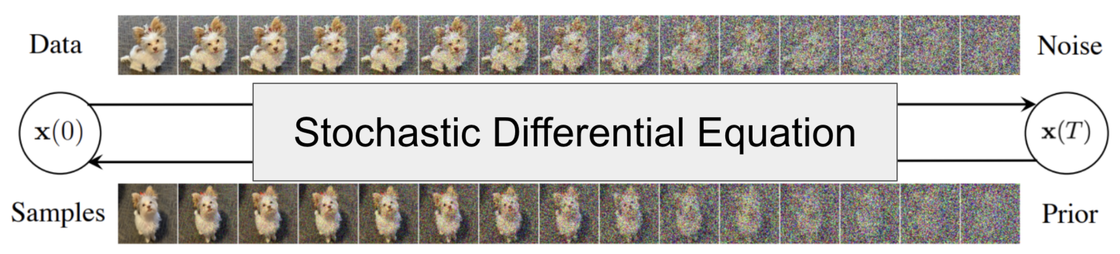
Optimal Stopping in Latent Diffusion Models

Yu-Han Wu(1),(2), Quentin Berthet(2), Gérard Biau(1), Claire Boyer(4), Romuald Elie(2) Pierre Marion(3)

(1) Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, Paris, France
(2) Google DeepMind
(3) EPFL, Institut de Mathematiques, Lausanne, Switzerland
(4) Université Paris-Saclay, CNRS, Laboratoire de Mathématiques d'Orsay, Orsay, France

SCAN ME

Diffusion Models [Ho et al., 2020]

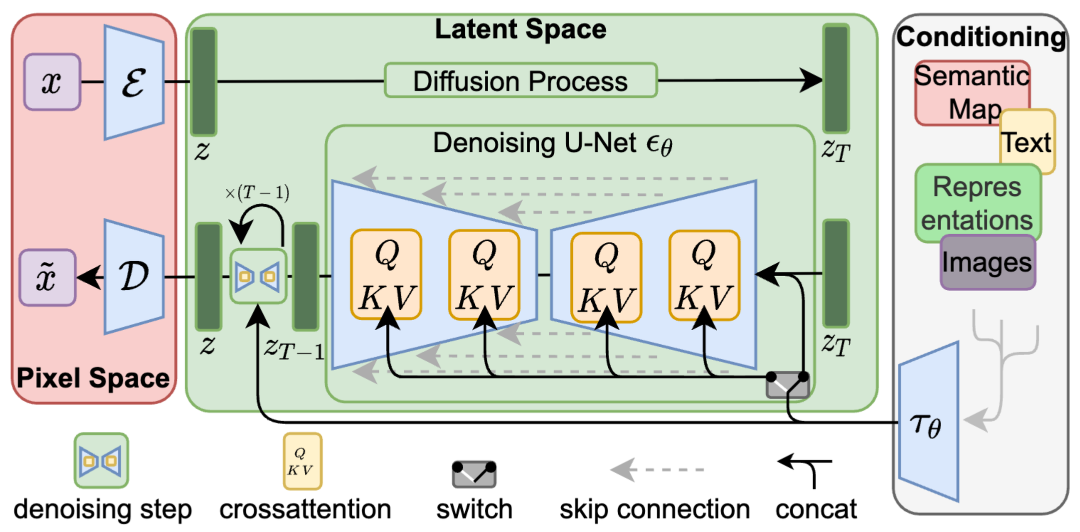


$$d\overleftarrow{X}_t = (\overleftarrow{X}_t + 2\underbrace{\nabla \log p_{T-t}(\overleftarrow{X}_t)}_{\text{score function}})dt + \sqrt{2}d\overleftarrow{B}_t, \quad \overleftarrow{X}_0 \sim p_T,$$

where $\overrightarrow{X}_t \sim p_t$ and $\overleftarrow{X}_{T-t} \overset{\mathcal{D}}{=} \overrightarrow{X}_t$.

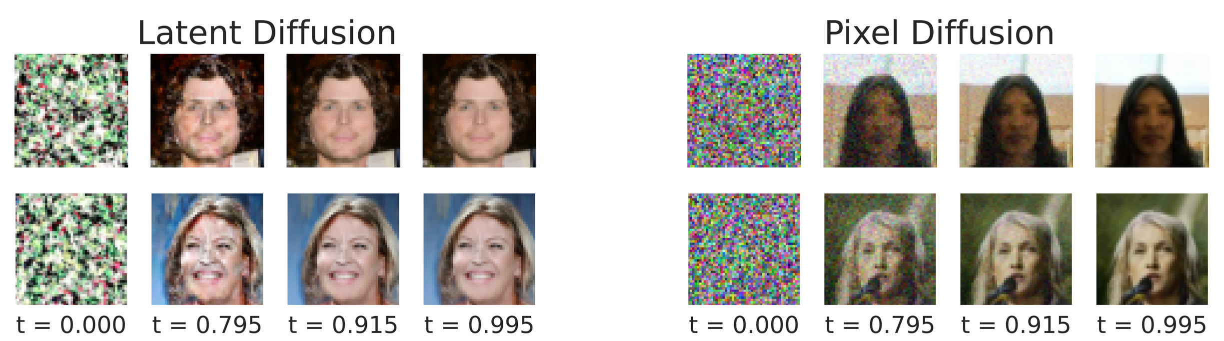
Latent Diffusion Model

We first train a pair of AE (E_θ, D_ϕ). E_θ maps the images in $\mathbb{R}^D$ to a smaller latent shape $\mathbb{R}^d$, and we train a diffusion model s with the encoded images in $\mathbb{R}^d$.



Observation

Diffusion on latent space usually causes the image quality to decrease in the last few steps of inference, which is not the case for pixel diffusion.



FID Score Comparison

Last steps in latent diffusion models actually degrade the image quality, which can be seen in the increasing of the FID score.

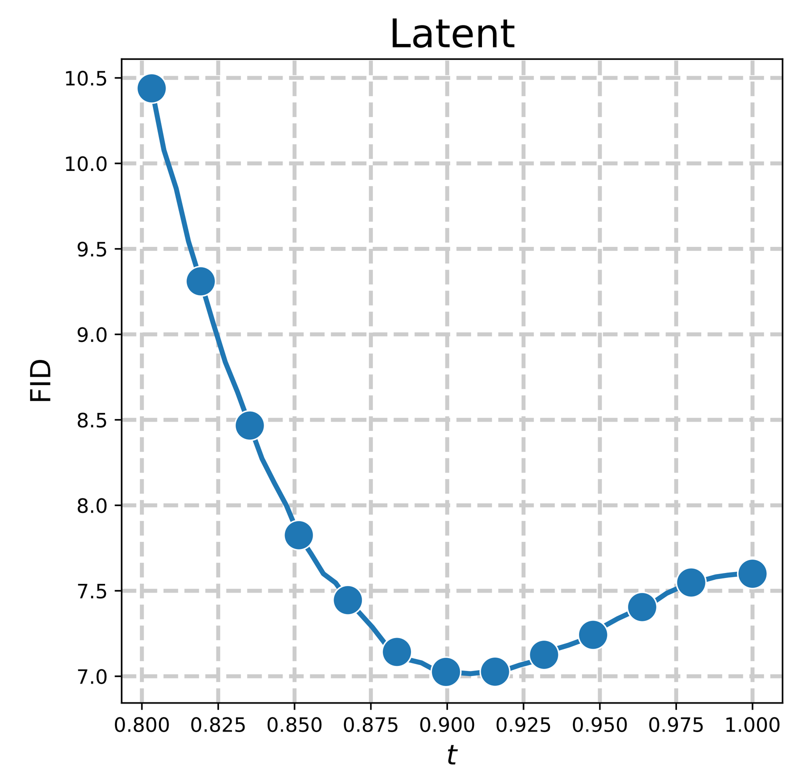


Figure 1:CelebA-HQ LDM

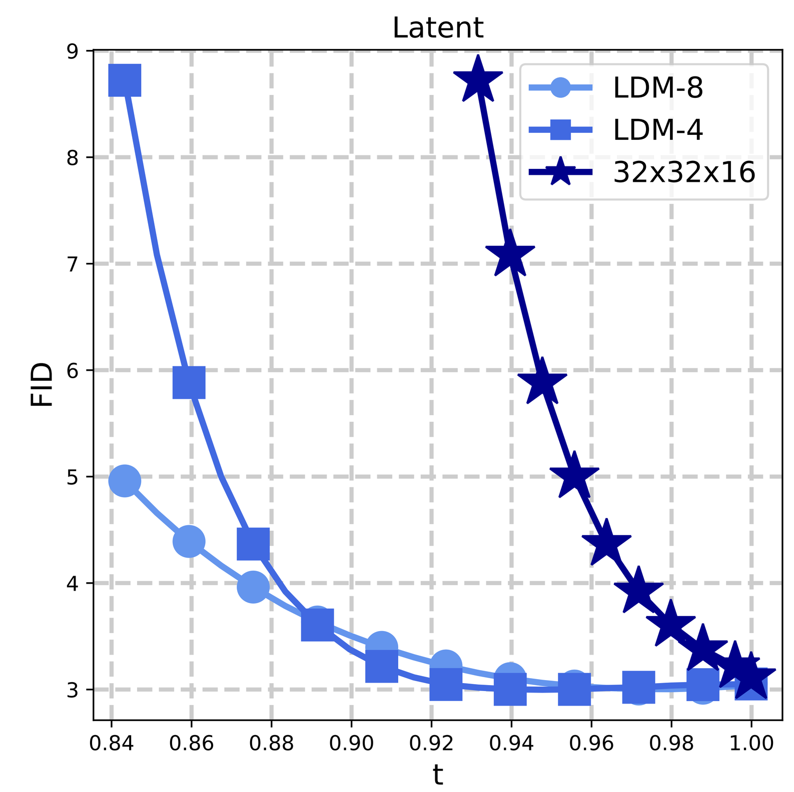


Figure 2:ImageNet-256 LDM

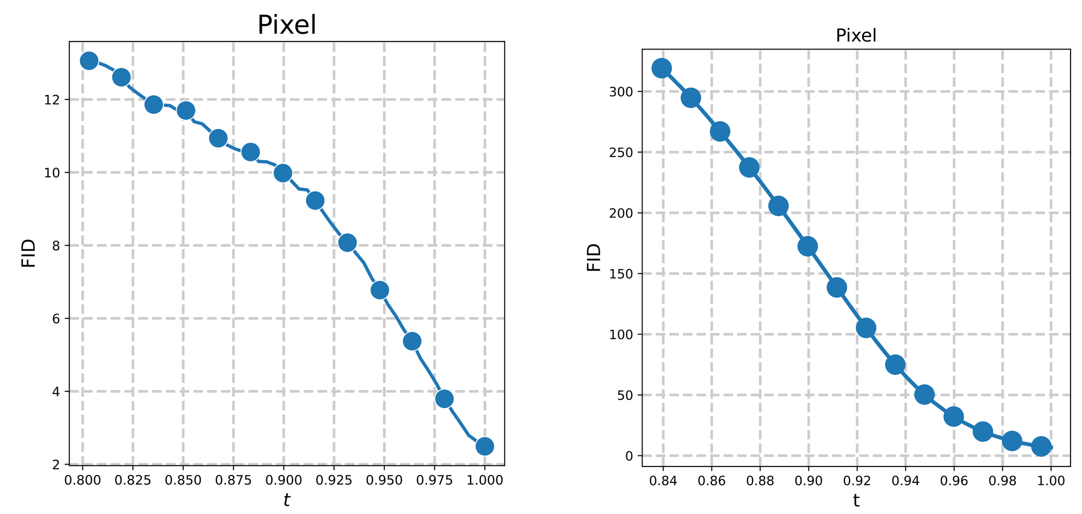


Figure 3:Pixel diffusion on CelebA and ImageNet-64

Theoretical Results

- **Gaussian Data.** Data distribution is centered D-dimensional Guassian with independent components, i.e. $p_{\text{data}} = \mathcal{N}(0, \text{diag}(\sigma_1^2, \ldots, \sigma_D^2))$, where $\sigma_1 > \ldots > \sigma_D$.
- **Projected diffusion process.** We consider orthogonal projection matrices P that map the diffusion process to a lower dimension as follows
$$dP\overrightarrow{X}_t = -w_t^2 P\overrightarrow{X}_t dt + \sqrt{2w_t^2}dP\overrightarrow{W}_t, \quad P\overrightarrow{X}_0 \sim P_{\#}p_0,$$
and can be reversed using the following backward diffusion process
$$dP\overleftarrow{X}_t = (w_{T-t}^2 P\overleftarrow{X}_t + 2w_{T-t}^2 s_P(P\overleftarrow{X}_t, T-t))dt + \sqrt{2w_{T-t}^2}dP\overleftarrow{W}_t.$$
- **Training.** In this simpler setup with Gaussian distribution, learning the score boils down to covariance matrix estimation, and we assume that the estimation is equal to $\hat{\Sigma} = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_D^2)$.

Non-Monotonicity

For $d \in \{1, \ldots, D\}$, the Fréchet distance $d_F(P_d^\top P_d \overleftarrow{X}_t, \overrightarrow{X}_0)$ is non-increasing with respect to t. On the other hand, $d_F(P_d^\top P_d \overleftarrow{\hat{X}}_t, \overrightarrow{X}_0)$ is non-increasing if and only if

$$\sum_{d'=1}^{d} (1 - \frac{\sigma_{d'}}{\hat{\sigma}_{d'}})(1 - \hat{\sigma}_{d'}^2) \geq 0.$$

Optimal projection and optimal stopping time

Assume that $\Sigma = \text{diag}(\sigma^2, \ldots, \sigma^2, 0, \ldots, 0)$ with the last $D - d_0$ entries equal to 0. Let $\varepsilon \in (0, 1)$. Then, there exists $\hat{\delta}_{d_0} \in [0, T]$ such that with probability $1 - 2d_0 e^{-\frac{n}{8}}$,

$$d_F(P_{d_0}^\top P_{d_0} \overleftarrow{\hat{X}}_{T-\hat{\delta}_{d_0}}, \overrightarrow{X}_0) = \min_{\substack{t \in [0,T] \\ d' \in \{1, \ldots, D\}}} d_F(P_{d'}^\top P_{d'} \overleftarrow{\hat{X}}_t, \overrightarrow{X}_0).$$

Theoretical Results (generalization)

- **General Gaussian data and estimation.** We now consider $p_{\text{data}} = \mathcal{N})(0, \Sigma)$ to be a general centered Gaussian distribution and $\hat{\Sigma}$ an estimation of Σ.
- **Time intervals.** We define time steps

$$\hat{T}_d(u) = T - \bar{a}^{-2}\left(\frac{\hat{\sigma}_d^2 - 4S(\Sigma)\varepsilon_u + 2\hat{\sigma}_d\sqrt{\hat{\sigma}_d^2 - 4S(\Sigma)\varepsilon_u}}{(1 - \hat{\sigma}_d^2)_+}\right),$$

$$\hat{t}_d(u) = T - \bar{a}^{-2}\left(\frac{\hat{\sigma}_d^2 + 4S(\Sigma)\varepsilon_u + 2\hat{\sigma}_d\sqrt{\hat{\sigma}_d^2 + 4S(\Sigma)\varepsilon_u}}{(1 - \hat{\sigma}_d^2)_+}\right),$$

where $\varepsilon_u = \frac{8C}{3}(\sqrt{\frac{D+u}{n}} + \frac{D+u}{n})$.

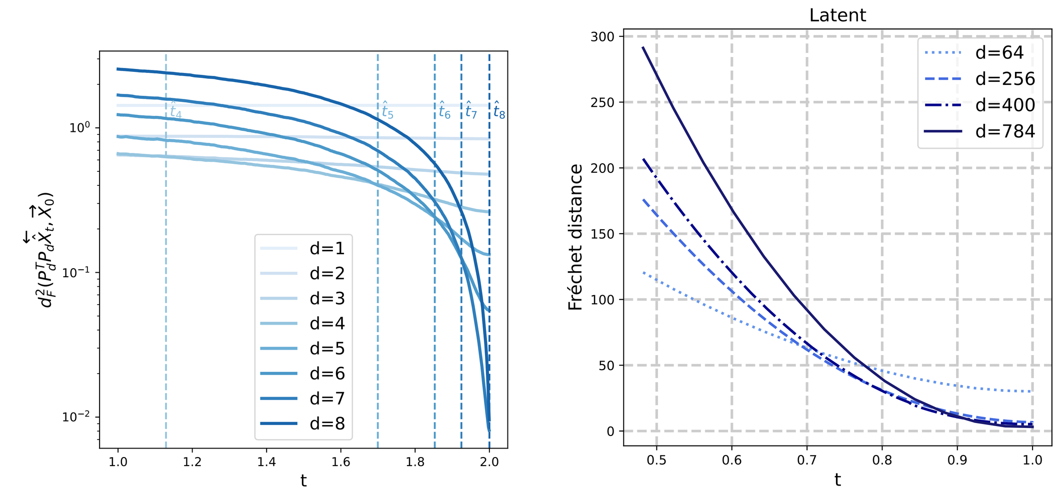
Optimal projection given t

For $d \in \{1, \ldots, D\}$ and any $t \in [\hat{T}_d(u), \hat{t}_{d+1}(u)]$, with probability $1 - 2e^{-u}$,

$$d \in \underset{d' \in \{1, \ldots, D\}}{\arg\min}\, d_F(\hat{O}P_{d'}^\top P_{d'}\hat{O}^\top \overleftarrow{X}_t, \overrightarrow{X}_0).$$

Numerical Verification

Experiment 1: Optimal projection given t.



Experiment 2: Optimal stopping time.