

Score-based Denoising Diffusion Models

- a tutorial

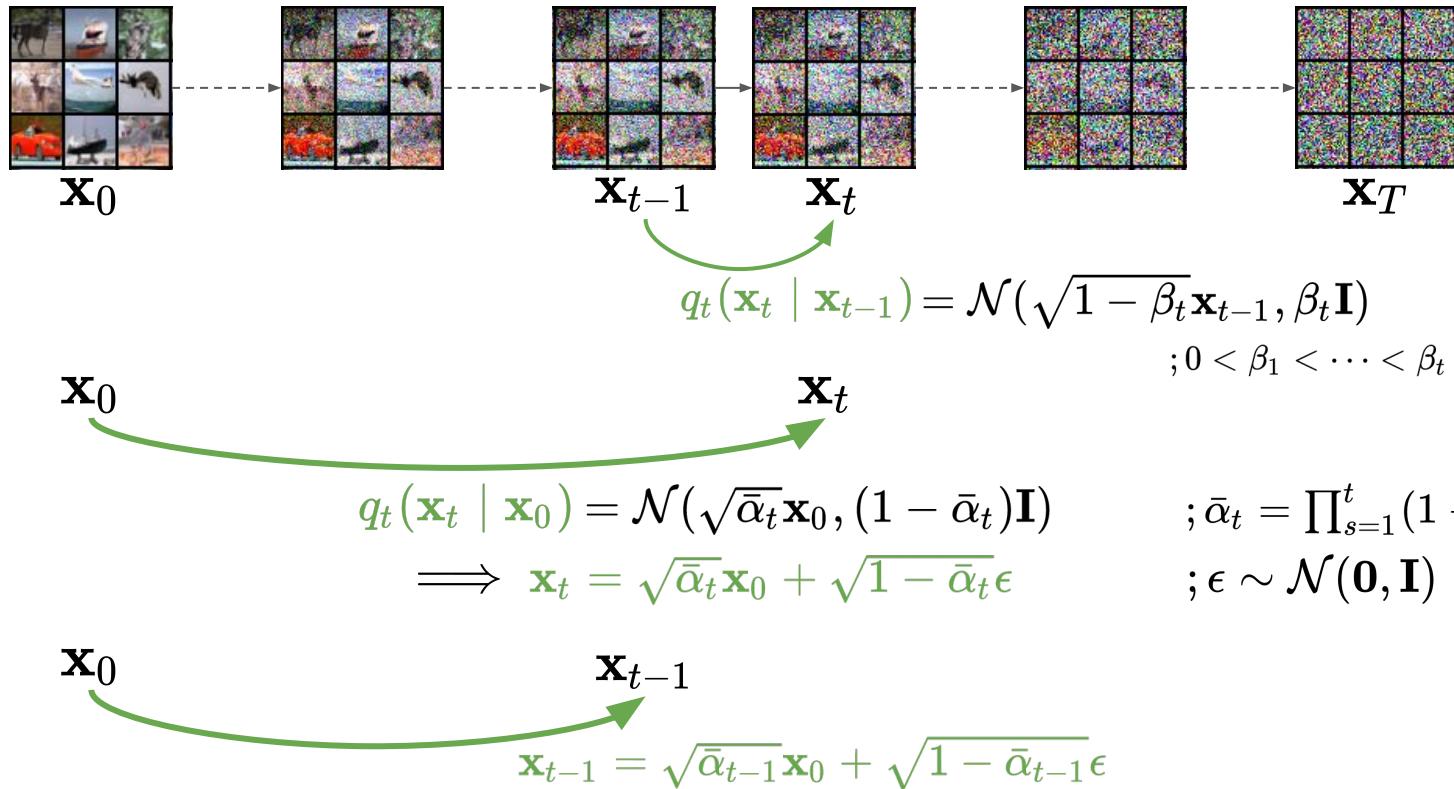
Vikram Voleti

PhD candidate, Mila, University of Montreal

Supervisor: Prof. Christopher Pal

1. Denoising Diffusion Probabilistic Model (DDPM)
2. Score-based Generative Model (SGM)
3. MCVD: Masked Conditional Video Diffusion

DDPM : Forward process

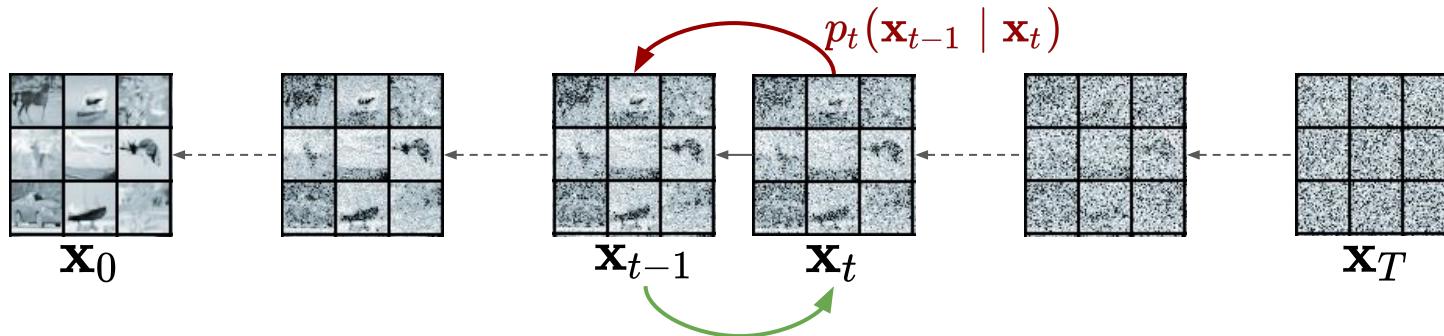


$$(1) \quad \mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

$$(2) \quad \mathbf{x}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon) / \sqrt{\bar{\alpha}_t}$$

$$(3) \quad \epsilon = \frac{1}{\sqrt{1 - \bar{\alpha}_t}} (\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)$$

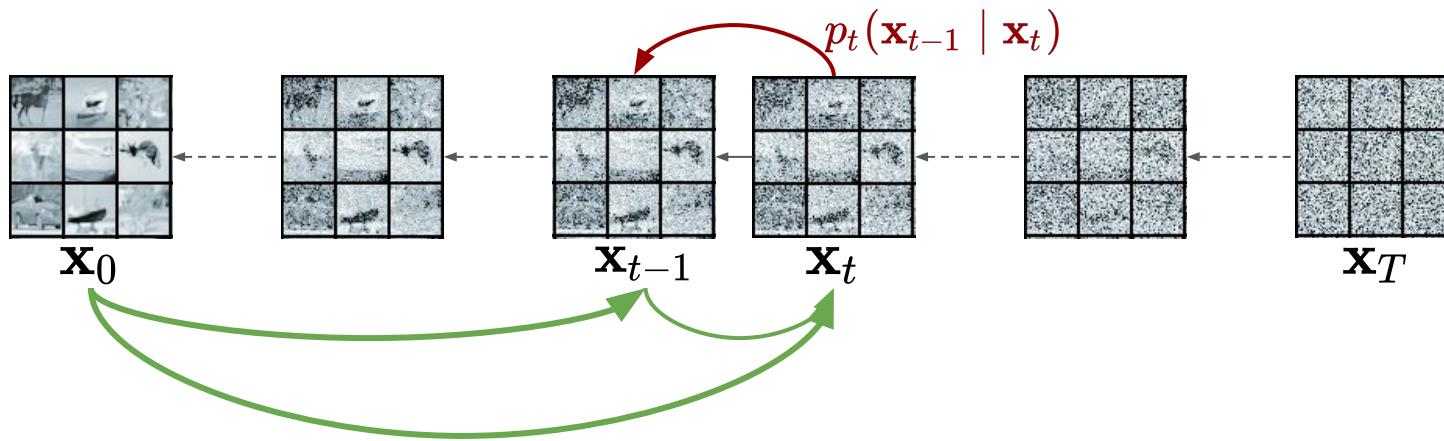
DDPM : Reverse process



$$p_t(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \frac{q_t(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \checkmark \quad p_t(\mathbf{x}_{t-1})?}{p_t(\mathbf{x}_t)?}$$

(Bayes' theorem)

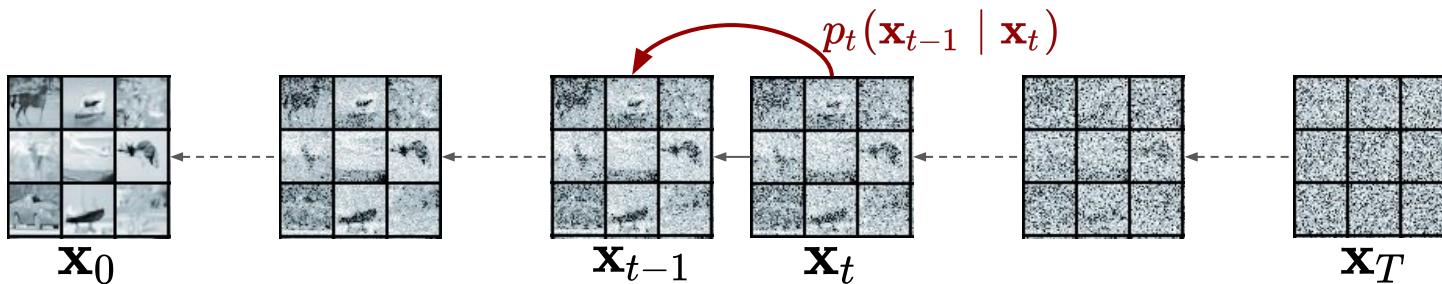
DDPM : Reverse process



$$p_t(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \boxed{\mathbf{x}_0}) = \frac{q_t(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \checkmark q_t(\mathbf{x}_{t-1} \mid \boxed{\mathbf{x}_0}) \checkmark}{q_t(\mathbf{x}_t \mid \boxed{\mathbf{x}_0}) \checkmark}$$

(Condition on \mathbf{x}_0)

DDPM : Reverse process



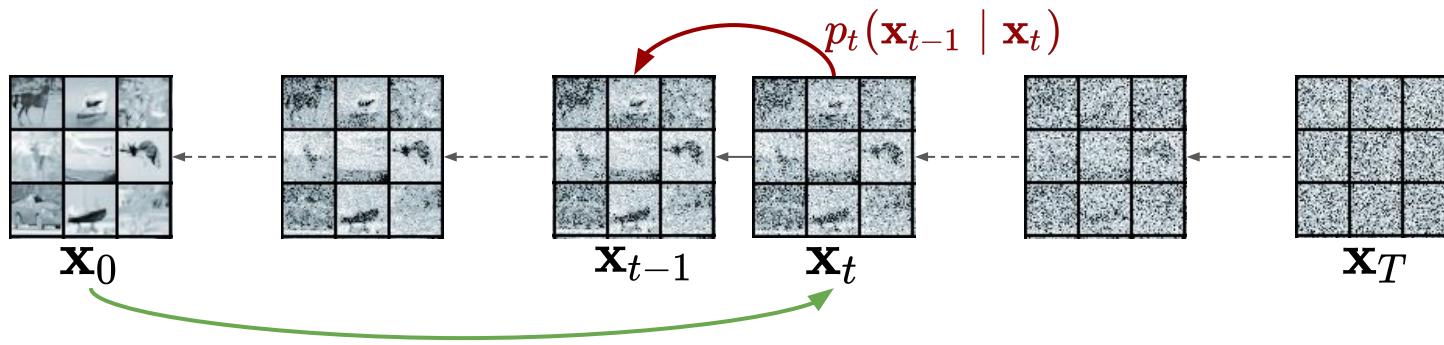
$$p_t(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$$

$$\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t ; \quad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

$p(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \mathbf{m}\mathbf{u}, \mathbf{\Lambda}^{-1}),$
 $p(\mathbf{v} \mid \mathbf{u}) = \mathcal{N}(\mathbf{v} \mid \mathbf{A}\mathbf{u} + \mathbf{b}, \mathbf{L}^{-1})$
 $\Rightarrow p(\mathbf{v}) = \mathcal{N}(\mathbf{v} \mid \mathbf{A}\mu + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T),$
 $\Rightarrow p(\mathbf{u} \mid \mathbf{v}) = \mathcal{N}(\mathbf{u} \mid \mathbf{C}(\mathbf{A}^T\mathbf{L}(\mathbf{v} - \mathbf{b}) + \mathbf{\Lambda}\mathbf{m}\mathbf{u}), \mathbf{C})$
 $[\mathbf{C} = (\mathbf{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}]$
 where $\mathbf{u} = \mathbf{x}_{t-1} \mid \mathbf{x}_0$, $\mathbf{v} = \mathbf{x}_t$

arxiv.org/abs/2006.11239

DDPM : Reverse process

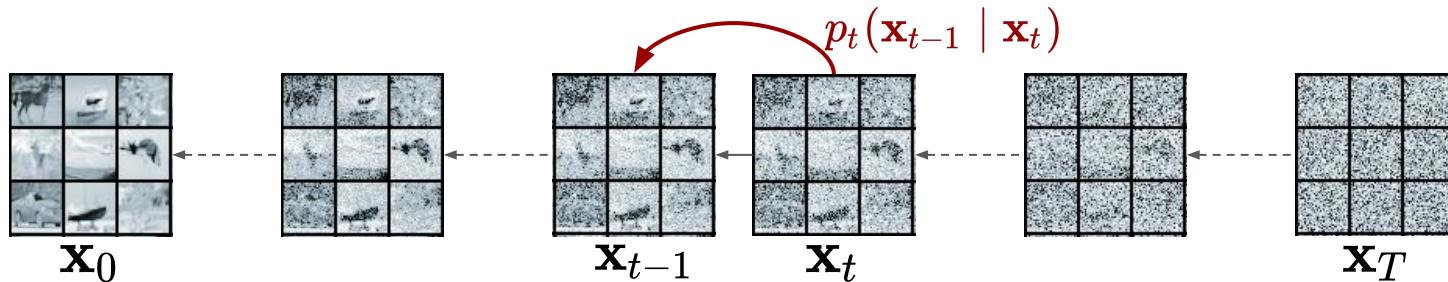


1 $\mathbf{x}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \quad \epsilon \quad) / \sqrt{\bar{\alpha}_t}$

2 $p_t(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I})$

From (2):
 $q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})$
 $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$
 $(2) \mathbf{x}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon) / \sqrt{\bar{\alpha}_t}$

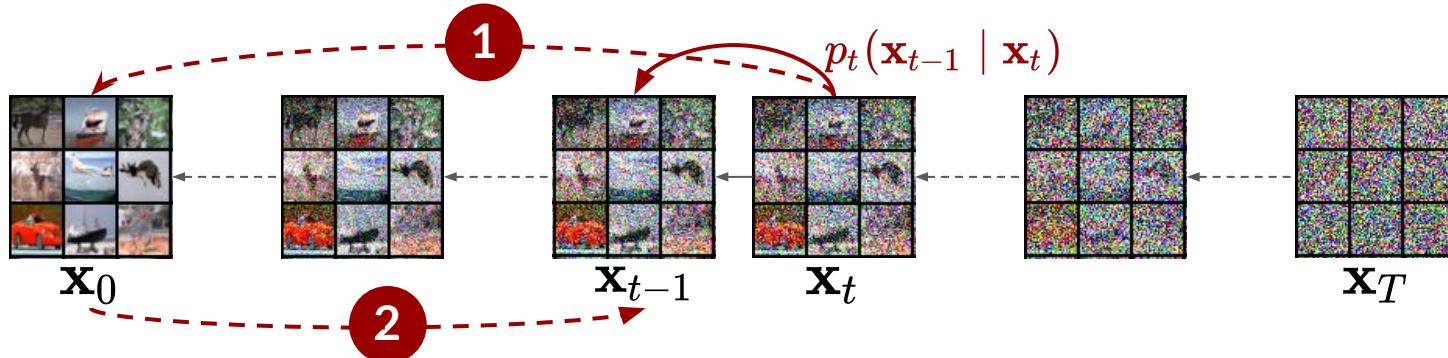
DDPM : Reverse process



Deep neural network!

1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$

2 $p_t(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \hat{\mathbf{x}}_0) = \mathcal{N}(\tilde{\mu}_t(\mathbf{x}_t, \hat{\mathbf{x}}_0), \tilde{\beta}_t \mathbf{I})$



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

DDPM

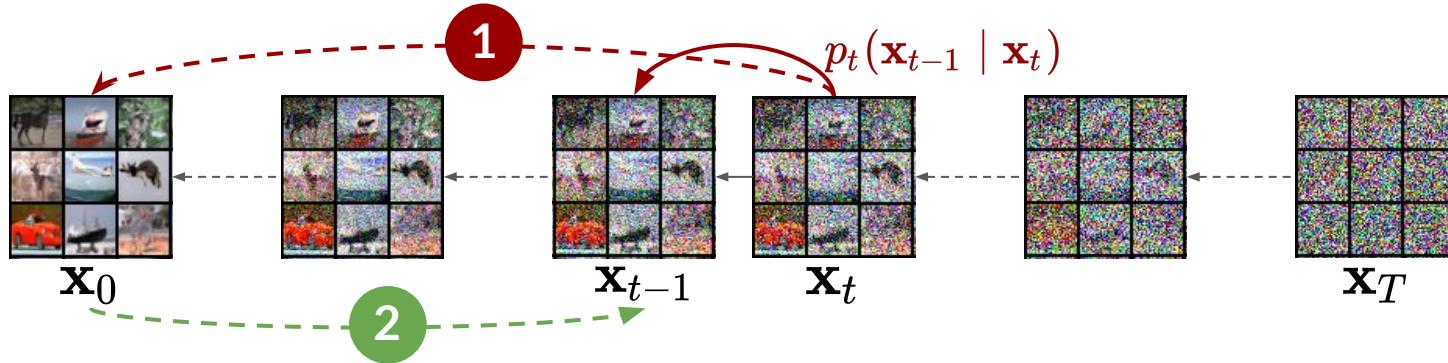
for $t = T \rightarrow 0 :$

1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$

2 $\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \hat{\mathbf{x}}_0 + \frac{\sqrt{1 - \beta_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \mathbf{z}_t$

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

DDIM : Generation



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

DDIM

for $t = T \rightarrow 0 :$

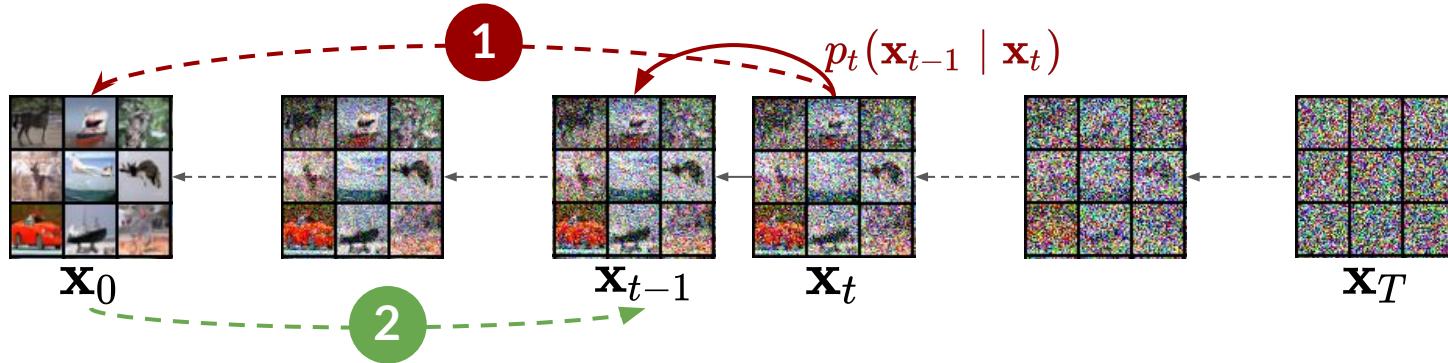
1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$

2 $\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{x}}_0 + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(\mathbf{x}_t, t)$

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})$$

$$(1) \quad \mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$

Deterministic!



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

DDIM

for $t = T \rightarrow 0 :$

1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$

2 $\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{x}}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(\mathbf{x}_t, t) + \sigma_t \mathbf{z}_t$

Deterministic $\Rightarrow \sigma_t = 0$

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

DDPM

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})$$

$$p_t(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \mathbf{x}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \mathbf{x}_t ; \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t \mathbf{I}\right)$$

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

for $t = T \rightarrow 0$:

1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$

2 $\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \hat{\mathbf{x}}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \mathbf{x}_t + \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t \mathbf{z}_t$

DDIM

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})$$

$$p_t(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon, \sigma_t^2 \mathbf{I}\right)$$

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

for $t = T \rightarrow 0$:

1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$

2 $\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{x}}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(\mathbf{x}_t, t) + \sigma_t \mathbf{z}_t$

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2 \right]$$

$(\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon)$

1. Noise matching

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2$$

$(\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon)$

2. Generation

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

DDPM

for $t = T \rightarrow 0$:

1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$

2 $\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \hat{\mathbf{x}}_0 + \frac{\sqrt{1 - \beta_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \mathbf{z}_t$

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

1. Training

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2 \right]$$

$(\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon)$

2. Generation

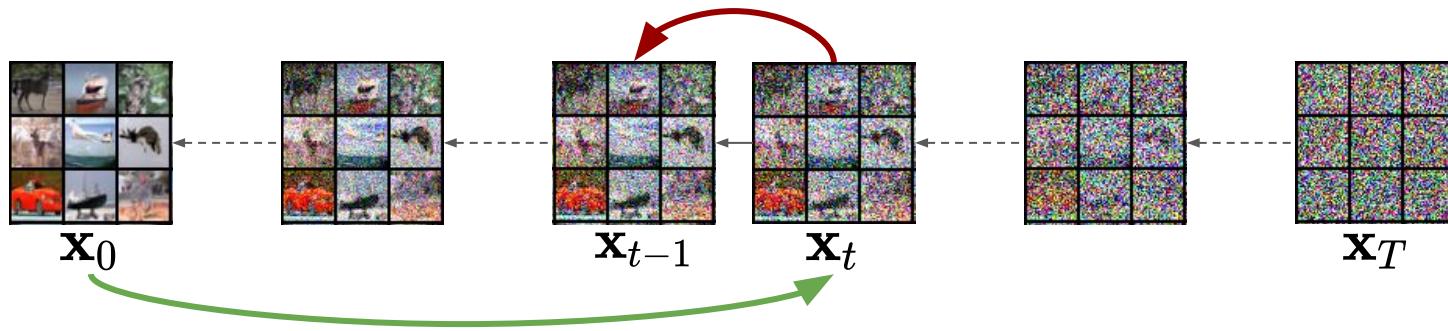
1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$

2 $\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \hat{\mathbf{x}}_0 + \frac{\sqrt{1 - \beta_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \mathbf{z}_t$

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

arxiv.org/abs/2006.11239

1. Denoising Diffusion Probabilistic Model (DDPM)
- 2. Score-based Generative Model (SGM)**
3. MCVD: Masked Conditional Video Diffusion



$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma_t^2 \mathbf{I}) \quad ; \sigma_1 < \dots < \sigma_t < \dots < \sigma_T$$

$$\implies \mathbf{x}_t = \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon} \quad ; \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Langevin Dynamics:



$$\mathbf{x}_{t-1} = \mathbf{x}_t + \lambda_t \nabla_{\mathbf{x}} \log p(\mathbf{x}_t) + \sqrt{2\lambda_t} \mathbf{z}_t \quad ; \lambda_t = \lambda \sigma_t^2 / \sigma_1^2; \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

1. Score Matching:

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)\|_2^2 \right]$$

$(\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon)$

Denoising Score Matching:

$$\approx \mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t | \mathbf{x}_0)\|_2^2 \right]$$

$$J_{ESMq}(\theta) = \mathbb{E}_{q(\mathbf{x})} \left[\frac{1}{2} \left\| \psi(\mathbf{x}; \theta) - \frac{\partial \log q(\mathbf{x})}{\partial \mathbf{x}} \right\|^2 \right]$$

$$J_{DSMq_\sigma}(\theta) = \mathbb{E}_{q_\sigma(\mathbf{x}, \tilde{\mathbf{x}})} \left[\frac{1}{2} \left\| \psi(\tilde{\mathbf{x}}; \theta) - \frac{\partial \log q_\sigma(\tilde{\mathbf{x}} | \mathbf{x})}{\partial \tilde{\mathbf{x}}} \right\|^2 \right]$$

$$J_{ESMq_\sigma} \curvearrowleft J_{DSMq_\sigma}$$

Score (gradient of log density)

1. Score Matching:

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma_t^2 \mathbf{I})$$

$$= \frac{1}{\sqrt{(2\pi)^{|\mathbf{x}|} \sigma_t^2}} \exp(-\frac{1}{2\sigma_t^2} (\mathbf{x}_t - \mathbf{x}_0)^T (\mathbf{x}_t - \mathbf{x}_0))$$

$$\Rightarrow \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \text{const} - \frac{1}{2\sigma_t^2} (\mathbf{x}_t - \mathbf{x}_0)^T (\mathbf{x}_t - \mathbf{x}_0)$$

Denoising Score Matching:

$$\Rightarrow (\text{Score}) \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{1}{\sigma_t^2} (\mathbf{x}_t - \mathbf{x}_0) = \frac{\mathbf{x}_0 - \mathbf{x}_t}{\sigma_t^2}$$

$$\approx \mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0)\|_2^2$$

$$\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \frac{\mathbf{x}_0 - \mathbf{x}_t}{\sigma_t^2}\|_2^2$$

1. Score Matching:

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)\|_2^2 \right]$$

($\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon$)

Denoising Score Matching:

$$\approx \mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t | \mathbf{x}_0)\|_2^2 \right]$$

$$= \mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\mathbf{s}_\theta(\mathbf{x}_t, t) - \frac{\mathbf{x}_0 - \mathbf{x}_t}{\sigma_t^2}\|_2^2 \right]$$

($\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon$)

1. Denoising Score Matching:

$$\ell(\theta) := \mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\| \mathbf{s}_\theta(\mathbf{x}_t, t) - \frac{\mathbf{x}_0 - \mathbf{x}_t}{\sigma_t^2} \right\|_2^2 \right]$$

($\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon$)

Variance of score:

$$\mathbb{E} \left[\left\| \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) \right\|_2^2 \right] = \mathbb{E} \left[\left\| -\frac{\mathbf{x}_t - \mathbf{x}_0}{\sigma_t^2} \right\|_2^2 \right] = \mathbb{E} \left[\left\| \frac{\sigma_t \epsilon}{\sigma_t^2} \right\|_2^2 \right] = \frac{1}{\sigma_t^2} \mathbb{E} \left[\|\epsilon\|_2^2 \right] = \frac{1}{\sigma_t^2} \dim(\epsilon)$$

(Un)weighted objective function:

$$\mathcal{L}(\theta) = \sigma_t^2 \ell(\theta) := \mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\| \sigma_t \mathbf{s}_\theta(\mathbf{x}_t, t) - \frac{\mathbf{x}_0 - \mathbf{x}_t}{\sigma_t} \right\|_2^2 \right]$$

($\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon$)

1. Denoising Score Matching:

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\| \sigma_t \mathbf{s}_\theta(\mathbf{x}_t, t) - \frac{\mathbf{x}_0 - \mathbf{x}_t}{\sigma_t} \right\|_2^2$$

$(\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon)$

- $\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon \implies \frac{\mathbf{x}_0 - \mathbf{x}_t}{\sigma_t} = \epsilon$
- Make: $\sigma_t \mathbf{s}_\theta(\mathbf{x}_t, t) = -\epsilon_\theta(\mathbf{x}_t, t)$ ----- How?

$$\implies \mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\| \epsilon - \epsilon_\theta(\mathbf{x}_t, t) \right\|_2^2$$

$(\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon)$

Objective
of DDPM!

Score (gradient of log density)

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma_t^2 \mathbf{I})$$

$$= \frac{1}{\sqrt{(2\pi)^{|\mathbf{x}|} \sigma_t^2}} \exp(-\frac{1}{2\sigma_t^2} (\mathbf{x}_t - \mathbf{x}_0)^T (\mathbf{x}_t - \mathbf{x}_0))$$

$$\Rightarrow \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \text{const} - \frac{1}{2\sigma_t^2} (\mathbf{x}_t - \mathbf{x}_0)^T (\mathbf{x}_t - \mathbf{x}_0)$$

$$\Rightarrow (\text{Score}) \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{1}{\sigma_t^2} (\mathbf{x}_t - \mathbf{x}_0) = \frac{\mathbf{x}_0 - \mathbf{x}_t}{\sigma_t^2}$$

$$\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon \quad (3) \quad \epsilon = \frac{1}{\sigma_t} (\mathbf{x}_t - \mathbf{x}_0)$$

$$\Rightarrow (\text{Score}) \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{1}{\sigma_t} \epsilon$$

∴ Estimating ϵ
is equivalent to
estimating a
scaled version
of the **Score**!

1. Training

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left| \left| \mathbf{s}_\theta(\mathbf{x}_t, t) - \frac{\mathbf{x}_0 - \mathbf{x}_t}{\sigma_t^2} \right| \right|_2^2 \right]$$

$(\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon)$

2. Generation

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \lambda_t \mathbf{s}_\theta(\mathbf{x}_t, t) + \sqrt{2\lambda_t} \mathbf{z}_t \quad ; \lambda_t = \lambda \sigma_t^2 / \sigma_1^2; \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

1. Training

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2 \right] \\ (\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon)$$

2. Generation

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \lambda_t \frac{-1}{\sigma_t} \epsilon_\theta(\mathbf{x}_t, t) + \sqrt{2\lambda_t} \mathbf{z}_t \quad ; \lambda_t = \lambda \sigma_t^2 / \sigma_1^2; \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

1. Training

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2$$

$(\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon)$

2. Generation

1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$ $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

2 $\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \hat{\mathbf{x}}_0 + \frac{\sqrt{1 - \beta_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \mathbf{z}_t$

DDPM

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})$$

- 1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$
- 2 $\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \hat{\mathbf{x}}_0 + \frac{\sqrt{1 - \beta_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t + \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \mathbf{z}_t$

SGM

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma_t^2 \mathbf{I})$$

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \lambda_t \frac{-1}{\sigma_t} \epsilon_\theta(\mathbf{x}_t, t) + \sqrt{2\lambda_t} \mathbf{z}_t$$

Score (gradient of log density)

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})$$

$$= \frac{1}{\sqrt{(2\pi)^{|\mathbf{x}|}(1-\bar{\alpha}_t)}} \exp(-\frac{1}{2(1-\bar{\alpha}_t)}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)^T (\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0))$$

$$\Rightarrow \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \text{const} - \frac{1}{2(1-\bar{\alpha}_t)}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)^T (\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)$$

$$\Rightarrow (\text{Score}) \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{1}{1-\bar{\alpha}_t}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)$$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t} \epsilon \quad (3) \quad \epsilon = \frac{1}{\sqrt{1-\bar{\alpha}_t}} (\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)$$

$$\Rightarrow (\text{Score}) \nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \epsilon$$

∴ Estimating ϵ
is equivalent to
estimating a
scaled version
of the Score!

1. Denoising Diffusion Probabilistic Model (DDPM)
2. Score-based Generative Model (SGM)
3. **MCVD: Masked Conditional Video Diffusion**

MCVD: Masked Conditional Video Diffusion



p past frames: $\mathbf{p} = \{\mathbf{p}^i\}_{i=1}^p$

k current frames: $\mathbf{x}_0 = \{\mathbf{x}_0^i\}_{i=1}^k$

f future frames: $\mathbf{f} = \{\mathbf{f}^i\}_{i=1}^f$

$(\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon)$

- **Video Prediction:**

$$\mathbb{E}_{t, [\mathbf{p}, \mathbf{x}_0] \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} ||\epsilon - \epsilon_\theta(\mathbf{x}_t | \boxed{\mathbf{p}}, t)||_2^2$$

- **Video Generation:**

$$\mathbb{E}_{t, \mathbf{x}_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} ||\epsilon - \epsilon_\theta(\mathbf{x}_t | t)||_2^2$$

- **Video Interpolation:**

$$\mathbb{E}_{t, [\mathbf{p}, \mathbf{x}_0, \mathbf{f}] \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} ||\epsilon - \epsilon_\theta(\mathbf{x}_t | \boxed{\mathbf{p}, \mathbf{f}}, t)||_2^2$$

arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion



MCVD: Masked Conditional Video Diffusion



p past frames: $\mathbf{p} = \{\mathbf{p}^i\}_{i=1}^p$

k current frames: $\mathbf{x}_0 = \{\mathbf{x}_0^i\}_{i=1}^k$

f future frames: $\mathbf{f} = \{\mathbf{f}^i\}_{i=1}^f$

$(\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon)$

- **Video Prediction:**

$$\mathbb{E}_{t, [\mathbf{p}, \mathbf{x}_0] \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \|\epsilon - \epsilon_\theta(\mathbf{x}_t \mid \mathbf{p}, t)\|_2^2$$

Random masking!

- **Video Prediction + Generation:**

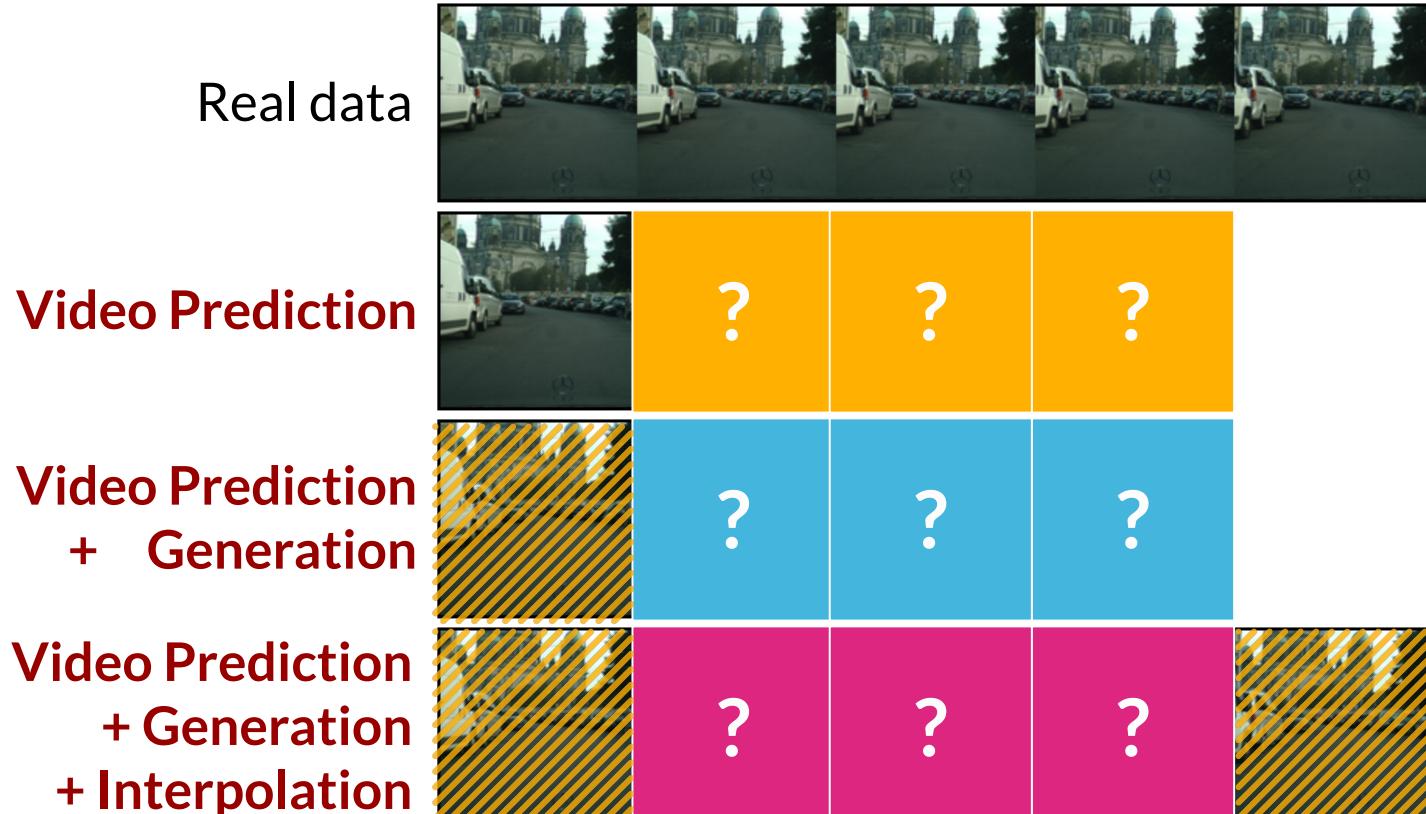
$$\mathbb{E}_{t, [\mathbf{p}, \mathbf{x}_0, \mathbf{f}] \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), m_p \sim \mathcal{B}(p_{\text{mask}})} \|\epsilon - \epsilon_\theta(\mathbf{x}_t \mid m_p \mathbf{p}, t)\|_2^2$$

- **Video Prediction + Generation + Interpolation:**

$$\mathbb{E}_{t, [\mathbf{p}, \mathbf{x}_0, \mathbf{f}] \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), (m_p, m_f) \sim \mathcal{B}(p_{\text{mask}})} \|\epsilon - \epsilon_\theta(\mathbf{x}_t \mid m_p \mathbf{p}, m_f \mathbf{f}, t)\|_2^2$$

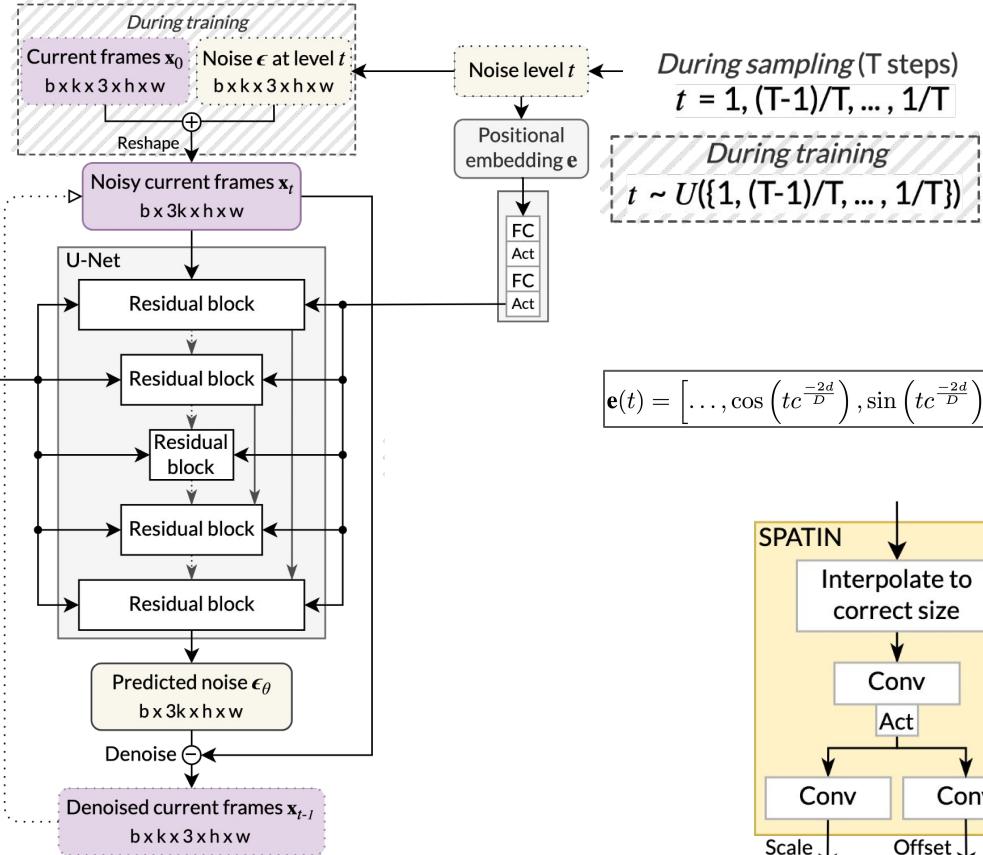
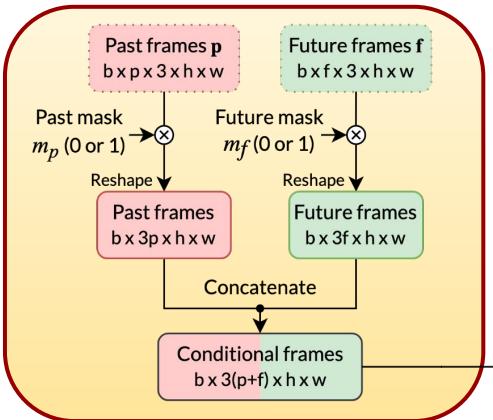
arxiv.org/abs/2205.09853

MCVD: Masked Conditional Video Diffusion

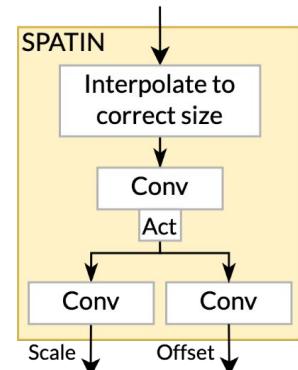


arxiv.org/abs/2205.09853

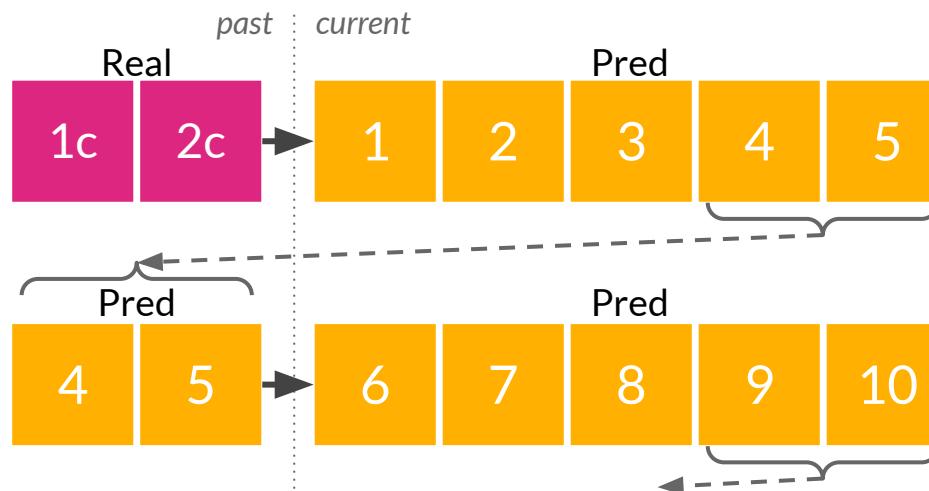
MCVD: Masked Conditional Video Diffusion



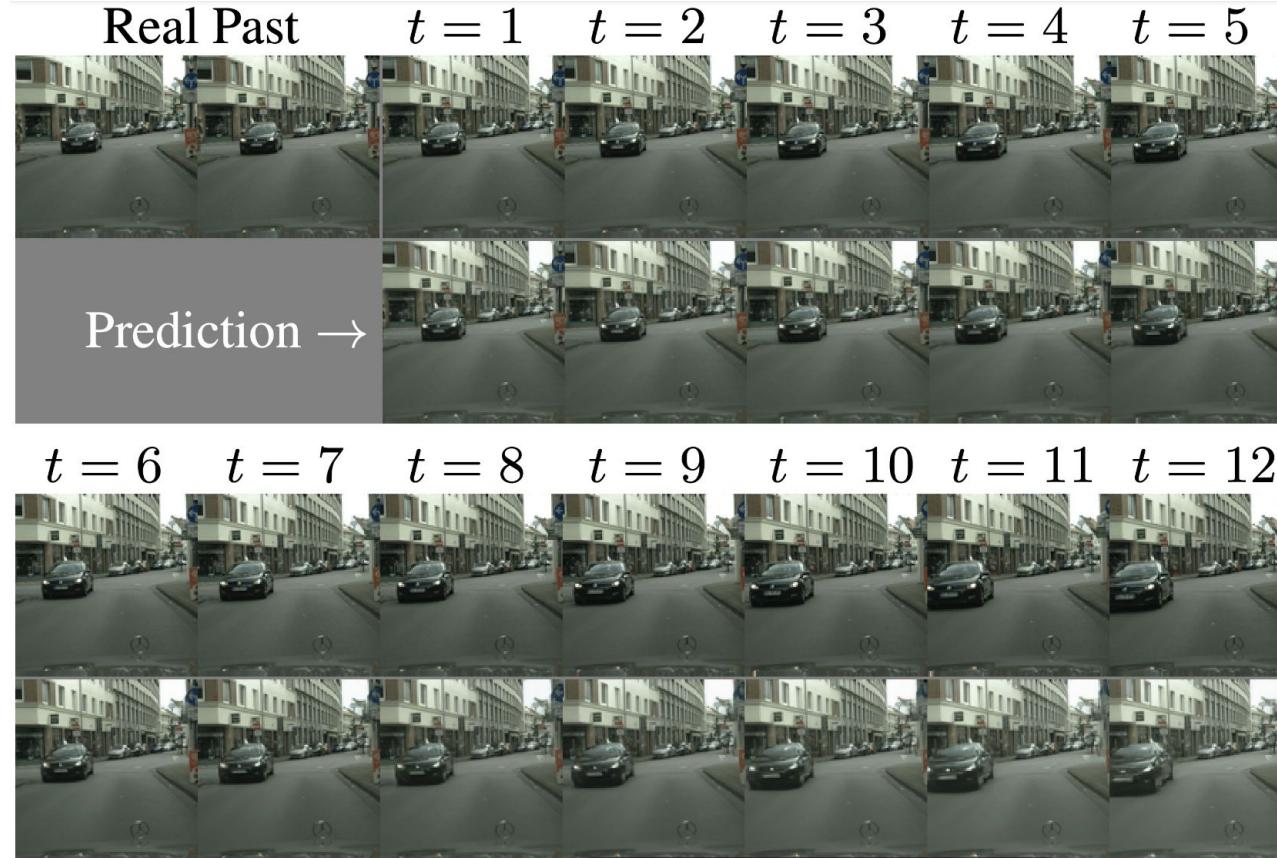
$$\mathbf{e}(t) = [\dots, \cos\left(tc^{\frac{-2d}{D}}\right), \sin\left(tc^{\frac{-2d}{D}}\right), \dots]^T$$



Block-autoregressive generation:



MCVD: Masked Conditional Video Diffusion



MCVD: Masked Conditional Video Diffusion



(128x128)

Cityscapes [2 → 28; trained on k]	k	FVD↓	LPIPS↓
SVG-LP Denton and Fergus [2018]	10	1300.26	0.549 ± 0.06
vRNN 1L Castrejón et al. [2019]	10	682.08	0.304 ± 0.10
Hier-vRNN Castrejón et al. [2019]	10	567.51	0.264 ± 0.07
GHVAE Wu et al. [2021]	10	418.00	0.193 ± 0.014
MCVD spatin (Ours)	5	184.81	0.121 ± 0.05
MCVD concat (Ours)	5	141.31	0.112 ± 0.05

(64x64)

BAIR [past $p \rightarrow pred$; trained on k]	p	k	$pred$	FVD↓	PSNR↑	SSIM↑
LVT [Rakhimov et al., 2020]	1	15	15	125.8	–	–
DVD-GAN-FP [Clark et al., 2019]	1	15	15	109.8	–	–
MCVD spatin (Ours)	1	5	15	103.8	18.8	0.826
TrIVD-GAN-FP [Luc et al., 2020]	1	15	15	103.3	–	–
VideoGPT [Yan et al., 2021]	1	15	15	103.3	–	–
CCVS [Le Moing et al., 2021]	1	15	15	99.0	–	–
MCVD concat (Ours)	1	5	15	98.8	18.8	0.829
MCVD spatin past-mask (Ours)	1	5	15	96.5	18.8	0.828
MCVD concat past-mask (Ours)	1	5	15	95.6	18.8	0.832
Video Transformer [Weissenborn et al., 2019]	1	15	15	94-96 ^a	–	–
FitVid [Babaeizadeh et al., 2021]	1	15	15	93.6	–	–
MCVD concat past-future-mask (Ours)	1	5	15	89.5	16.9	0.780
SAVP [Lee et al., 2018]	2	14	14	116.4	–	–
MCVD spatin (Ours)	2	5	14	94.1	19.1	0.836
MCVD spatin past-mask (Ours)	2	5	14	90.5	19.2	0.837
MCVD concat (Ours)	2	5	14	90.5	19.1	0.834
MCVD concat past-future-mask (Ours)	2	5	14	89.6	17.1	0.787
MCVD concat past-mask (Ours)	2	5	14	87.9	19.1	0.838
SVG-LP [Akan et al., 2021]	2	10	28	256.6	–	0.816
SLAMP [Akan et al., 2021]	2	10	28	245.0	19.7	0.818
SAVP [Lee et al., 2018]	2	10	28	143.4	–	0.795
Hier-vRNN Castrejón et al. [2019]	2	10	28	143.4	–	0.822
MCVD spatin (Ours)	2	5	28	132.1	17.5	0.779
MCVD spatin past-mask (Ours)	2	5	28	127.9	17.7	0.789
MCVD concat (Ours)	2	5	28	120.6	17.6	0.785
MCVD concat past-mask (Ours)	2	5	28	119.0	17.7	0.797
MCVD concat past-future-mask (Ours)	2	5	28	118.4	16.2	0.745

(64x64)

SMMNIST [5 → 10; trained on k]	k	FVD↓	SSIM↑
SVG [Denton and Fergus, 2018]	10	90.81	0.688
vRNN 1L [Castrejón et al., 2019]	10	63.81	0.763
Hier-vRNN [Castrejón et al., 2019]	10	57.17	0.760
MCVD concat (Ours)	5	25.63	0.786
MCVD spatin (Ours)	5	23.86	0.780
MCVD concat past-future-mask	5	20.77	0.753

(64x64)

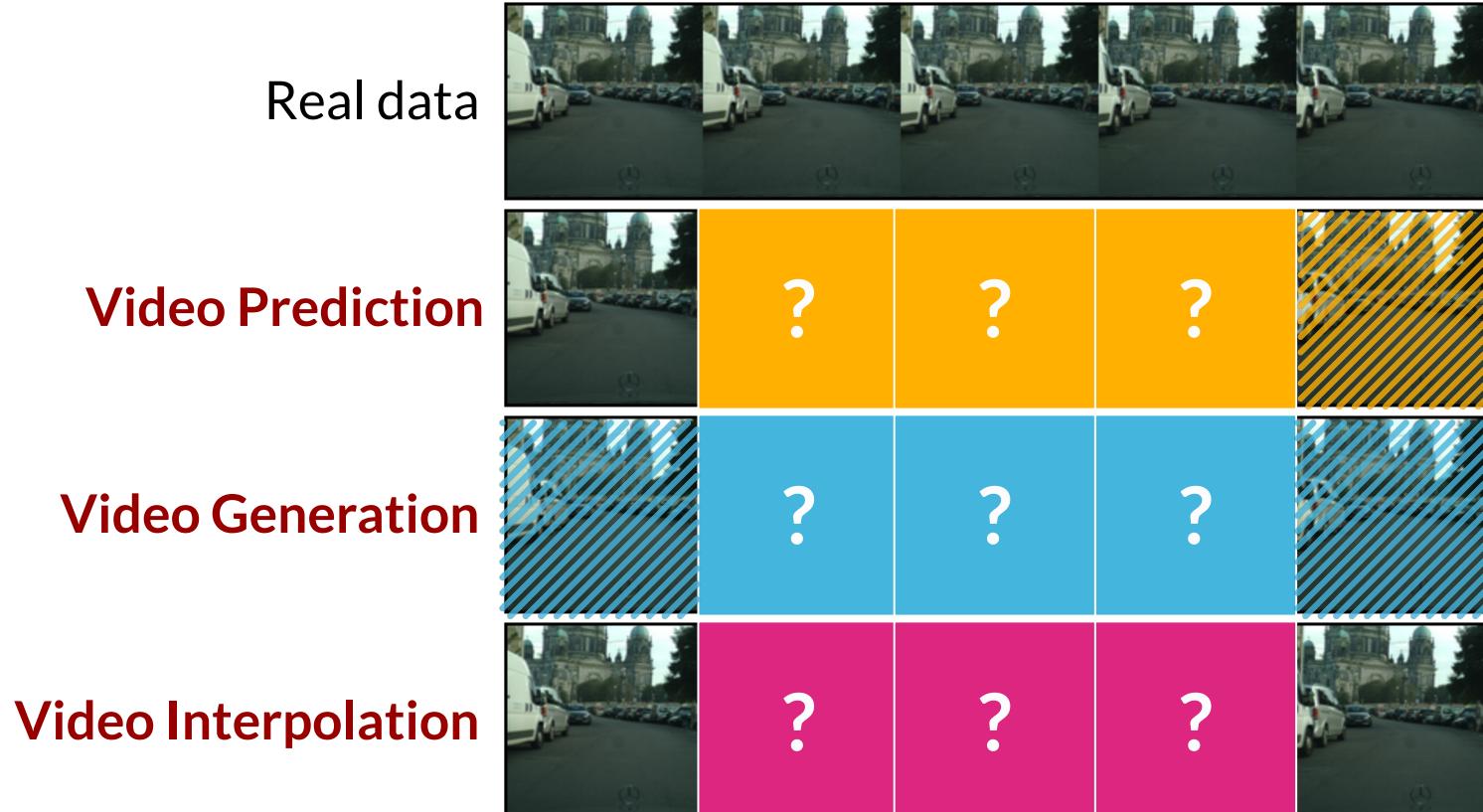
KTH [10 → pred; trained on 10]	$pred$	FVD↓	PSNR↑	SSIM↑
SAVP [Lee et al., 2018]	30	374	26.5	0.756
MCVD spatin (Ours)	30	323	27.5	0.835
MCVD concat past-future-mask (Ours)	30	295	24.3	0.746
SLAMP [Akan et al., 2021]	30	228	29.4	0.865
SRVP [Franceschi et al., 2020]	30	222	29.7	0.870

arxiv.org/abs/2205.09853

mask-cond-video-diffusion.github.io

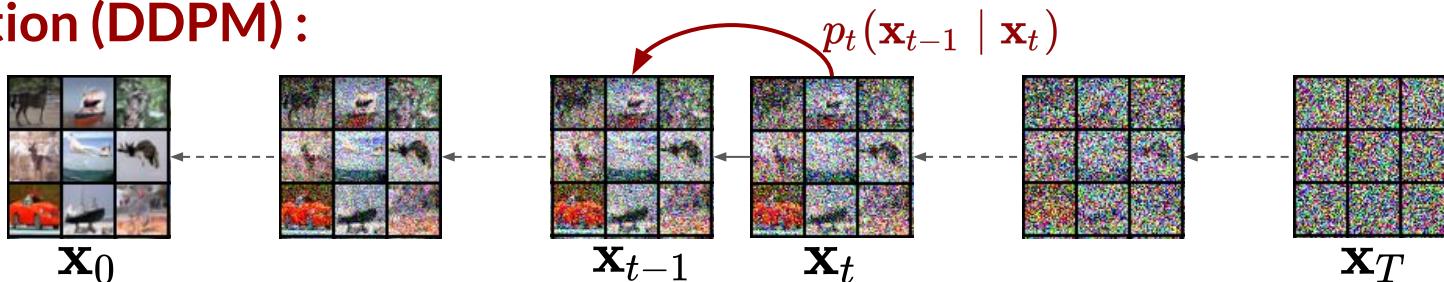


Thank you!



Score-based Denoising Diffusion Models

Generation (DDPM) :



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

for $t = T \rightarrow 0$:

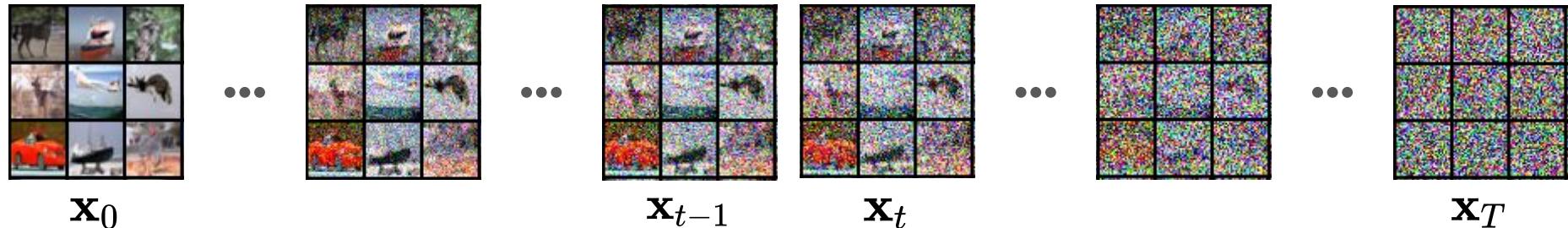
1 $\hat{\mathbf{x}}_0 = (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)) / \sqrt{\bar{\alpha}_t}$

2 $p_t(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \hat{\mathbf{x}}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}_t(\mathbf{x}_t, \hat{\mathbf{x}}_0), \tilde{\beta}_t \mathbf{I})$

$$\mathbf{x}_{t-1} \sim p_t(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \hat{\mathbf{x}}_0)$$

Score-based Denoising Diffusion Models

Forward Process (SMLD)

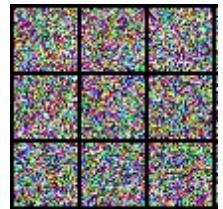
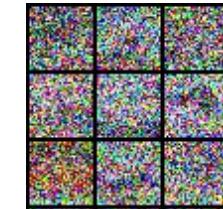


$$\mathbf{x}_0 \xrightarrow{\quad} \mathbf{x}_t$$

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0, \sigma_t^2 \mathbf{I})$$

$$\implies \mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon \quad ; \epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{I})$$

Score-based Denoising Diffusion Models

 \mathbf{x}_0  \mathbf{x}_{t-1}  \mathbf{x}_t 



MCVD: Masked Conditional Video Diffusion

