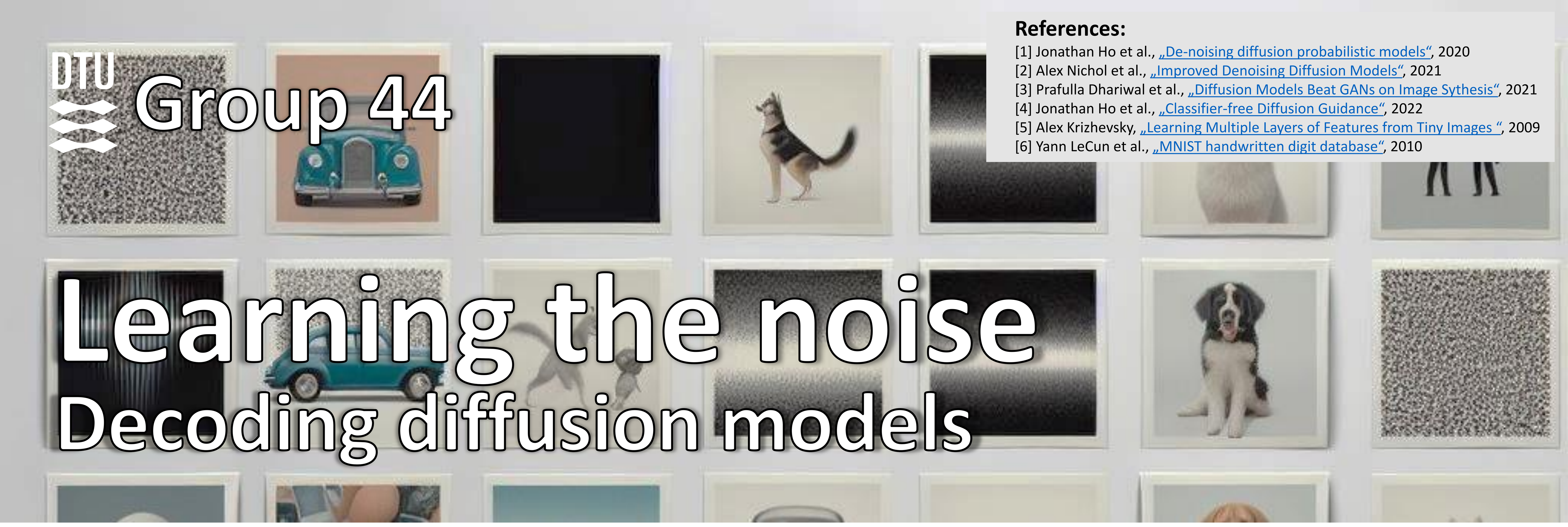


Learning the noise

Decoding diffusion models

References:

[1] Jonathan Ho et al., „De-noising diffusion probabilistic models", 2020
[2] Alex Nichol et al., „Improved Denoising Diffusion Models", 2021
[3] Prafulla Dhariwal et al., „Diffusion Models Beat GANs on Image Sythesis", 2021
[4] Jonathan Ho et al., „Classifier-free Diffusion Guidance", 2022
[5] Alex Krizhevsky, „Learning Multiple Layers of Features from Tiny Images ", 2009
[6] Yann LeCun et al., „MNIST handwritten digit database", 2010

Group 44: Zeljko Antunovic (s233025), Alex Belai (s233423), Lukas Samuel Czekalla (s233561), Nándor Takács (s232458)

Repository

Theory

Forward process:

- Iteratively transform image into $\mathcal{N}(0, I)$

$$q(x_{1:T}|x_0) \coloneqq \prod_{t=1}^{T} q(x_t|x_{t-1})$$

$$q(x_t|x_{t-1}) \coloneqq \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Reverse process:

- Predict image iteratively from $\mathcal{N}(0, I)$
- CNN predicts noise reduction

$$p_\theta(x_{0:T}) \coloneqq p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

$$p_\theta(x_{t-1}|x_t) \coloneqq \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

- Use reparametrization trick to predict ϵ_θ

Training & Sampling:

Algorithm 1 Training

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

Algorithm 2 Sampling

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

Algorithms taken from [1]

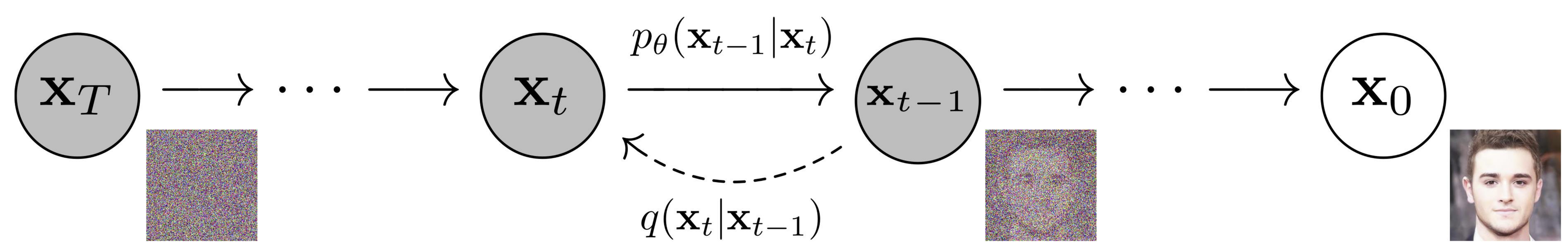


Image taken from [1]

Model Architecture

Architecture:

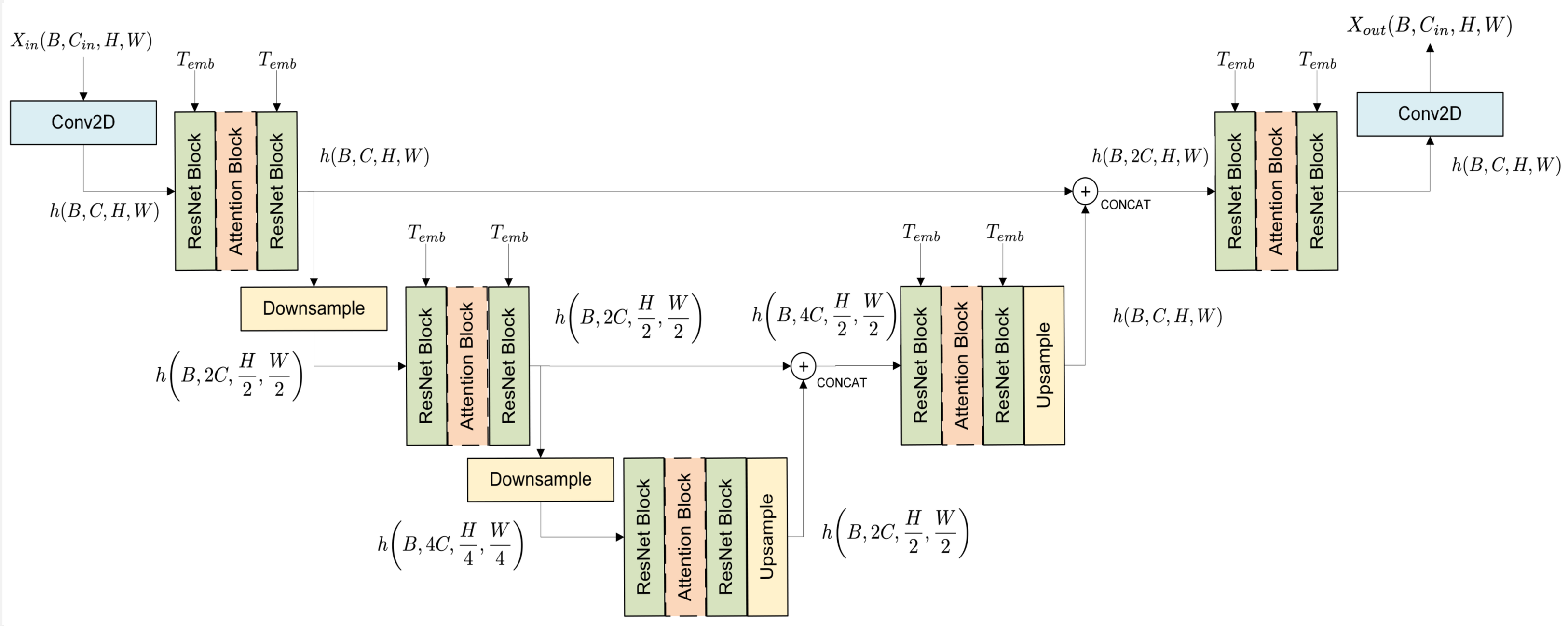
- UNet Architecture
  - 3 encoder layers
  - 2 ResNet Blocks per layer
  - Optionally 1 attention block per layer
  - Input & Output 2D convolution to match feature dimensions

Attention Blocks:

- Self-attention between feature map pixels

ResNet-Blocks:

- Group Normalisation
- Sinusoidal temporal embeddings passed through linear layer and added
- Label embeddings are optional, used for classifier-free guidance



Training

Data:

- Training on MNIST & CIFAR-10
- Images normalized to $[-1,1]$
- No resizing & data augmentation

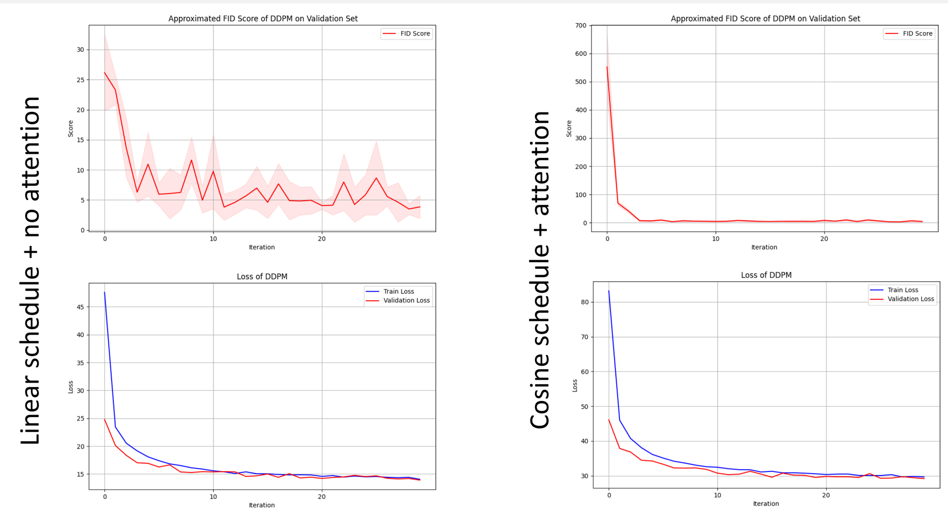
Hyperparameters:

- Linear schedule: $\beta_0 = 10^{-4}$ and $\beta_T = 0.02$
- Cosine schedule: $s = 0.008$
- Diffusion steps: $T = 1000$

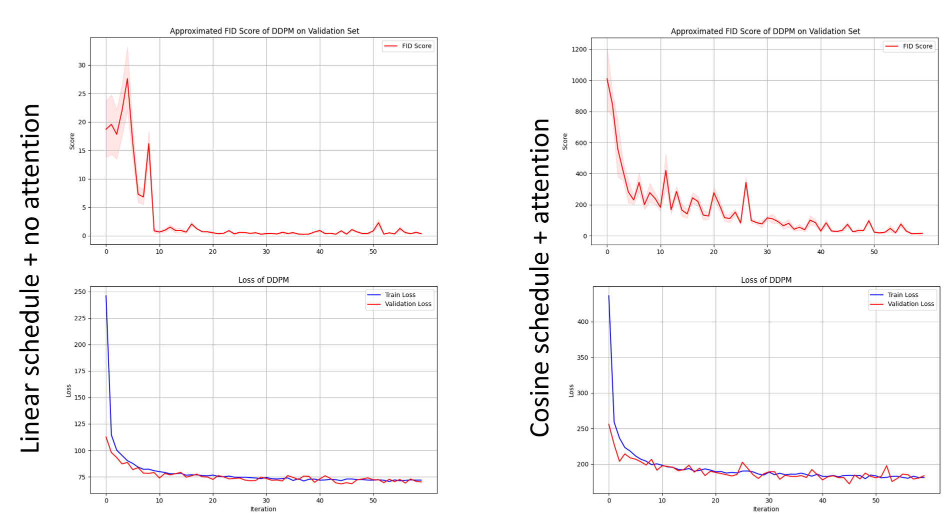
Training:

- Optimizer is $Adam$ with $l_r = 10^{-4}$
- Training for 30 epochs on MNIST & 60 epochs on CIFAR-10
- Validation-loss on validation set & FID score on 5 minibatches of validation set
- FID for final model on training & test set for 8192 generated samples
- Training with different ablations
  - Linear & cosine schedule
  - Attention & no-attention layers in network

MNIST



CIFAR-10

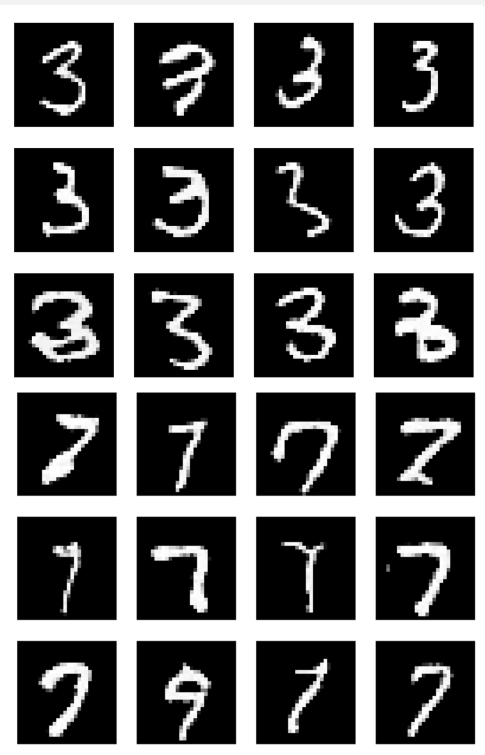


Guided Sampling

Classifier-free guided sampling:

- label embeddings added to temporal embeddings
- Trained for 30 epochs, $Adam$ optimizer with $l_r = 10^{-4}$

MNIST



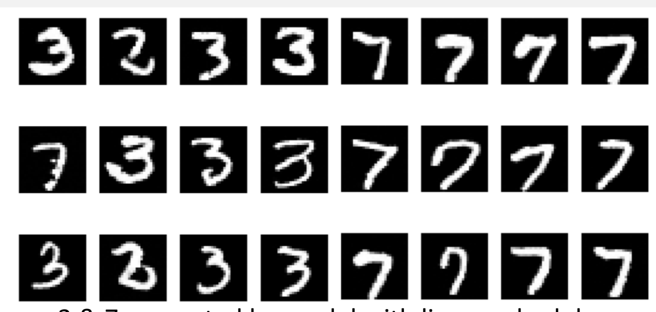
3 & 7 generated by model with linear schedule + attention

Classifier guided sampling:

- Subtracting gradient w.r.t. input of classifier at each sampling step

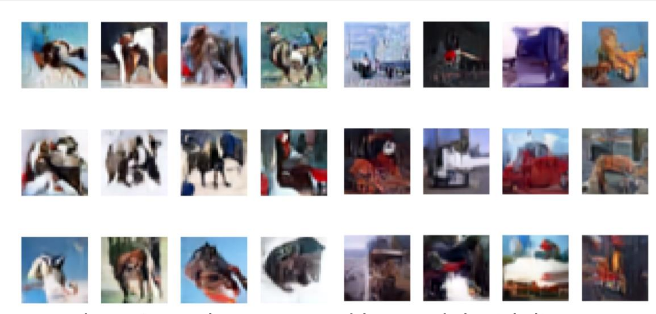
$$\epsilon_\theta = \epsilon_{\theta'} - \lambda \nabla \ln(p_c(y|x_t))$$

MNIST



3 & 7 generated by model with linear schedule + no attention & $\lambda = 200$

CIFAR-10

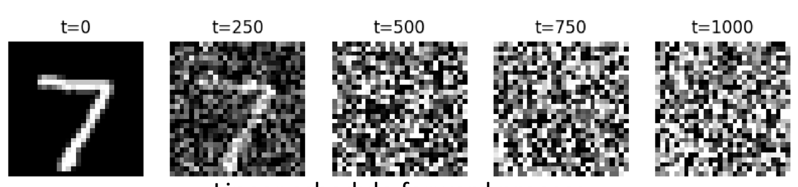


dogs & trucks generated by model with linear schedule + no attention & $\lambda = 200$

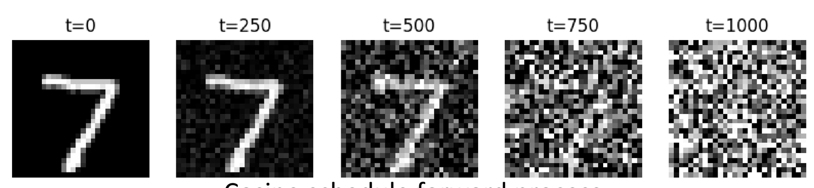
Sampling & Evaluation

MNIST

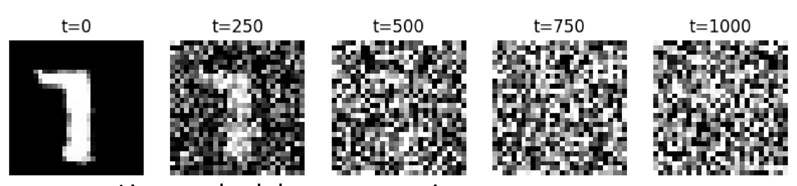
8192 samples On MNIST classifier	Linear schedule + no attention	Linear schedule + attention	Cosine schedule + no attention	Cosine schedule + attention
$Train - FID$	2.286	2.849	2.105	1.766
$Test - FID$	2.803	3.490	2.724	2.385



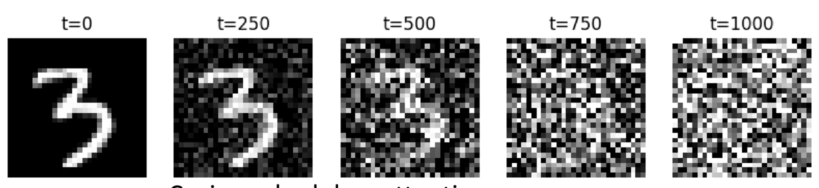
Linear schedule forward process



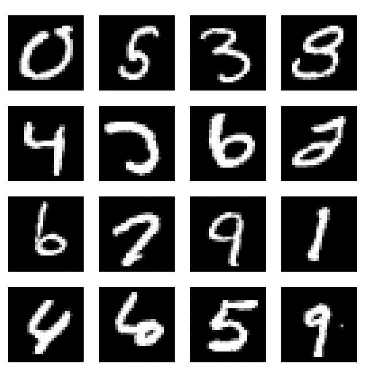
Cosine schedule forward process



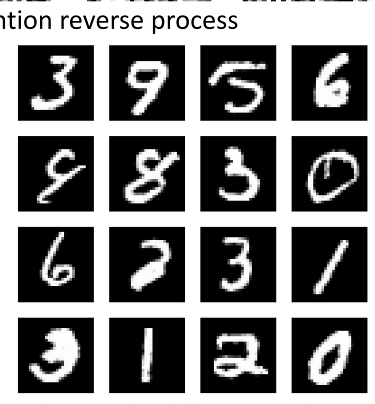
Linear schedule + no attention reverse process



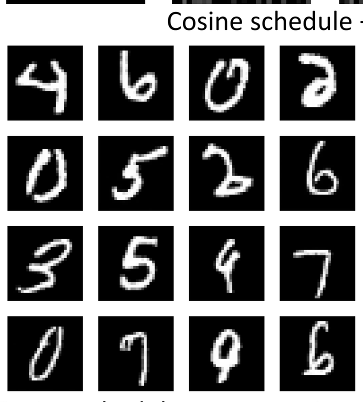
Cosine schedule + attention reverse process



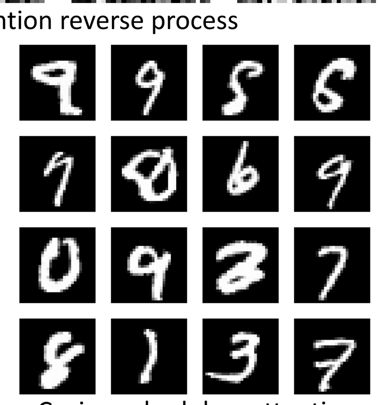
Linear schedule + no attention



Linear schedule + attention



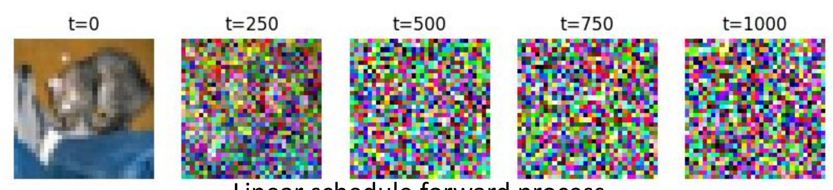
Cosine schedule + no attention



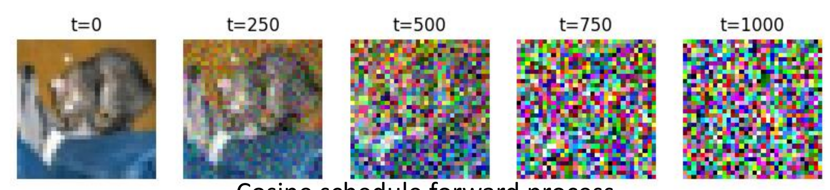
Cosine schedule + attention

CIFAR-10

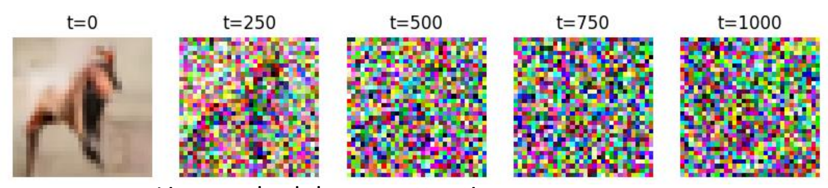
8192 samples On MNIST classifier	Linear schedule + no attention	Linear schedule + attention	Cosine schedule + no attention	Cosine schedule + attention
$Train - FID$	0.114	0.165	49.393	14.565
$Test - FID$	0.128	0.153	49.058	14.508



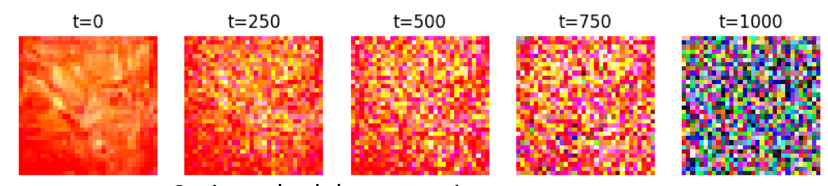
Linear schedule forward process



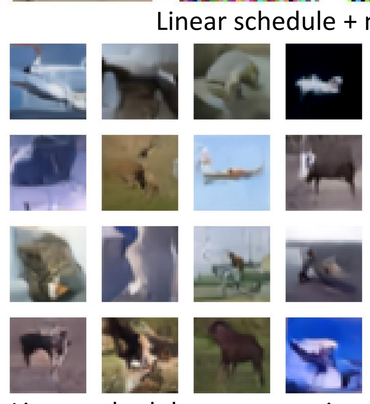
Cosine schedule forward process



Linear schedule + no attention reverse process

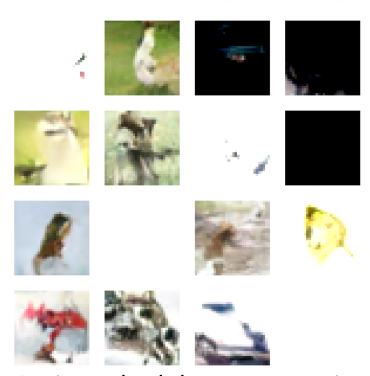


Cosine schedule + attention reverse process

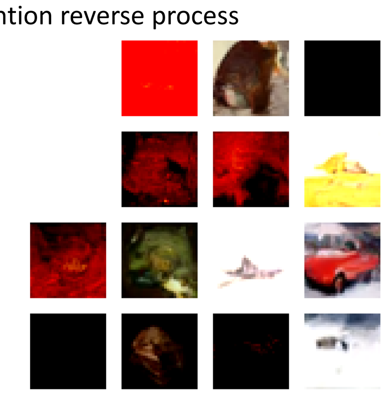


Linear schedule + no attention

Linear schedule + attention



Cosine schedule + no attention



Cosine schedule + attention