## Tuesday, April 28

Exam:

57 / 62.5

31 $\xleftrightarrow[22]{}$ 98

Grades: All B- or above

Hypothetical

(details on Discord and Canvas)

## Plan

Proposals due

↳ I'll message w/ feedback

Diffusion today

Muon next

## Talks

Ahmet 4-5pm Wednesday @ Davidson

Feyza 4:15-5:15pm Friday @ Estella 1021

# Generative AI

So far: Use VAEs

1. Sample latent from Gaussian
2. Decode

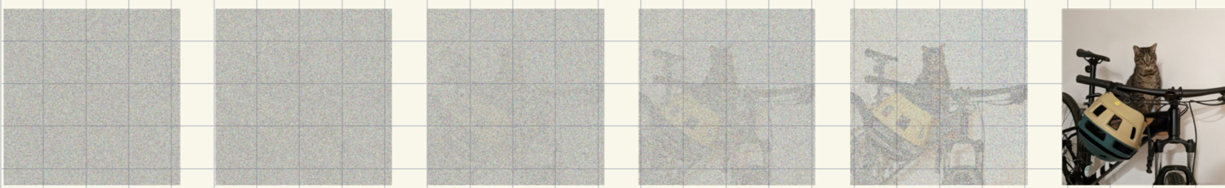
Goal: Scale inference time compute for better outputs (e.g., reasoning)

## Diffusion

$x_T \sim \mathcal{N}(0, I)$

training

$x_0 \sim p_{data}$



inference

# Adding Noise

$T$ steps $\quad t \in \{1, \dots, T\}$ $\quad \beta_t \in (0,1)$ is noise rate

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \epsilon_t \quad \text{where} \quad \epsilon_t \sim \mathcal{N}(0, I)$$

$$\mathbb{E}\, \|x_t\|_2^2 = \mathbb{E}\left[ 1-\beta_t \|x_{t-1}\|_2^2 + \sqrt{1-\beta_t}\sqrt{\beta_t} < x_{t-1}, \epsilon_t > + \beta_t \|\epsilon_t\|_2^2 \right]$$

$$= (1-\beta_t)\cdot 1 + 0 + \beta_t \cdot \mathbb{E}\|\epsilon_t\|_2^2 = 1$$

Fact: $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2 I)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2 I)$

$$Y_1 + Y_2 \sim \mathcal{N}(\mu_1 + \mu_2, (\sigma_1^2 + \sigma_2^2) I)$$

$$x_t \sim \mathcal{N}(\sqrt{1-\beta_t}\, x_{t-1}, \beta_t I)$$

When $T$ is large $x_T$ is indistinguishable from pure noise

# Direct Jump Property

Let $\alpha_t = 1 - \beta_t$.

$$x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \epsilon_t$$

$$= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_{t-1}}\, \epsilon_{t-1}\right) + \sqrt{1-\alpha_t}\, \epsilon_t$$

$$\vdots$$

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

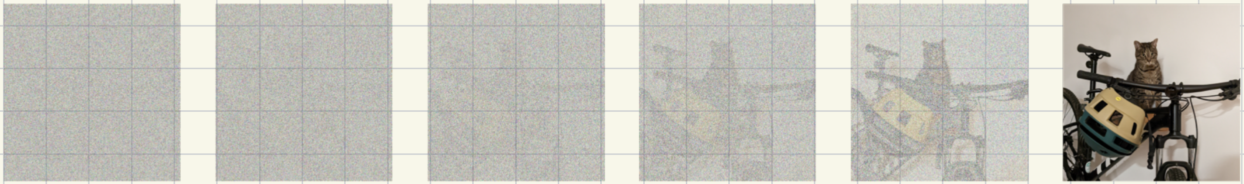
$$x_t \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\, x_0, \sqrt{1-\bar{\alpha}_t}\, I\right) \qquad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$$

# Thursday, April 30

- Final time slot Wed. May 13 7-10pm
- Project Proposals
- Thank you for coming to Ahmet's talk
- Feyza's talk tomorrow at 4:15pm in Estella 1021

# Diffusion

$x_T \sim \mathcal{N}(0, I)$ training $x_0 \sim p_{data}$



inference

$\beta_t \in (0,1) \qquad \alpha_t = 1-\beta_t$

$$x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \epsilon_t \quad \text{where} \quad \epsilon_t, \epsilon \sim \mathcal{N}(0, I)$$

$$= \sqrt{\alpha_t \alpha_{t-1}}\, x_{t-2} + \sqrt{1-\alpha_t \alpha_{t-1}}\, \epsilon$$

$$\vdots$$

$$= \sqrt{\underbrace{\prod_{s=1}^{t} \alpha_s}_{\bar{\alpha}_t}}\, x_0 + \sqrt{1-\prod_{s=1}^{t} \alpha_s}\, \epsilon \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\, I\right)$$

# Removing Noise

Q: what is the distribution to remove noise?

← Must condition on initial image $x_0$

$$q(x_{t-1} | x_t, x_0) = \mathcal{N}(\tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$$

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

Hint: Use Bayes Rule to write $q(x_{t-1} | x_t, x_0)$ in terms of $q(x_t | x_{t-1}, x_0)$, $q(x_{t-1} | x_0)$, $q(x_t, x_0)$

... But how do we know $x_0$?

Idea: Train model to learn $x_0 \sim p_\theta(x_{t-1} | x_t)$

$$= \mathcal{N}(\mu_\theta(x_t, t), \tilde{\beta}_t I)$$

$$\mathcal{L}oss = \mathbb{E}_{t, x_0, \epsilon} \; KL\left(q(x_{t-1} | x_t, x_0) \,||\, p_\theta(x_{t-1} | x_t)\right)$$

$$= \mathbb{E}_{t, x_0, \epsilon} \; \frac{1}{2\tilde{\beta}_t} \| \tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t) \|_2^2$$

Instead of predicting $x_0$, let's predict noise.

$$\approx \mathbb{E}_{t, x_0, \epsilon} \; \| \epsilon - \epsilon_\theta(x_t, t) \|_2^2$$

## Inference

Sample $x_T \sim \mathcal{N}(0,1)$

For $t = T, \dots, 1$ :

current pred.

$$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)\right) \leftarrow$$

rearranged

$$x_t = \sqrt{\bar{\alpha}_t}\,x_0 + \sqrt{1-\bar{\alpha}}\,\epsilon$$

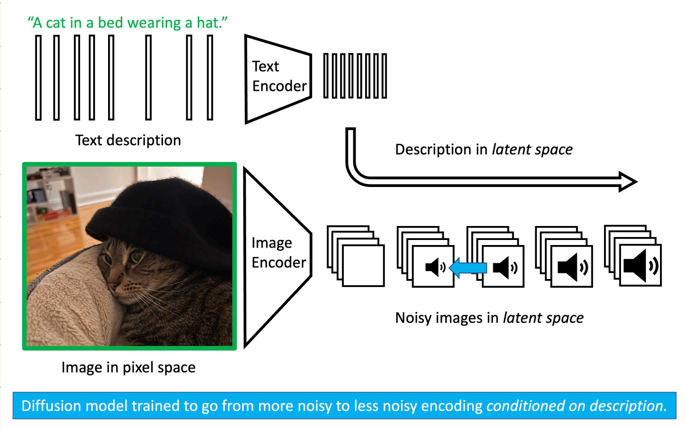
renoised

$$\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\,x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\,\beta_t\,\hat{x}_0$$

$$x_{t-1} = \mu_\theta(x, t) + \tilde{\beta}_t z$$

Return $x_0$

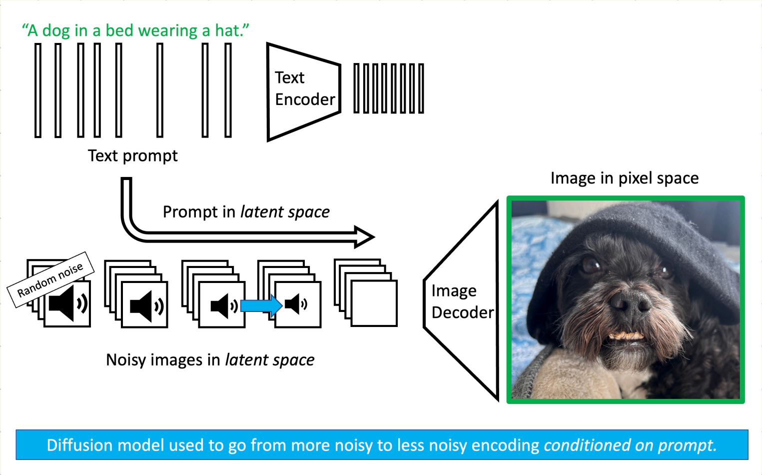
# Conditional Training




Condition on caption

$$\epsilon_\theta(x_t, t, c)$$

# Conditional Inference



# Classifier-free Guidance

Move **toward** caption

1. Prediction with text prompt $\epsilon_\theta(x_t, t, c)$

2. Prediction without text prompt $\epsilon_\theta(x_t, t, \emptyset)$

$$\tilde{\epsilon}_\theta = (1+w)\,\epsilon_\theta(x_t, t, c) - w\,\epsilon_\theta(x_t, t, \theta)$$

where $w$ is guidance scale