SC395:

Image Generative Models in Computer Vision

Viraj Shah

Lecture 5

Jan 16th, 2026

sc395.virajshah.com

ILLINOIS

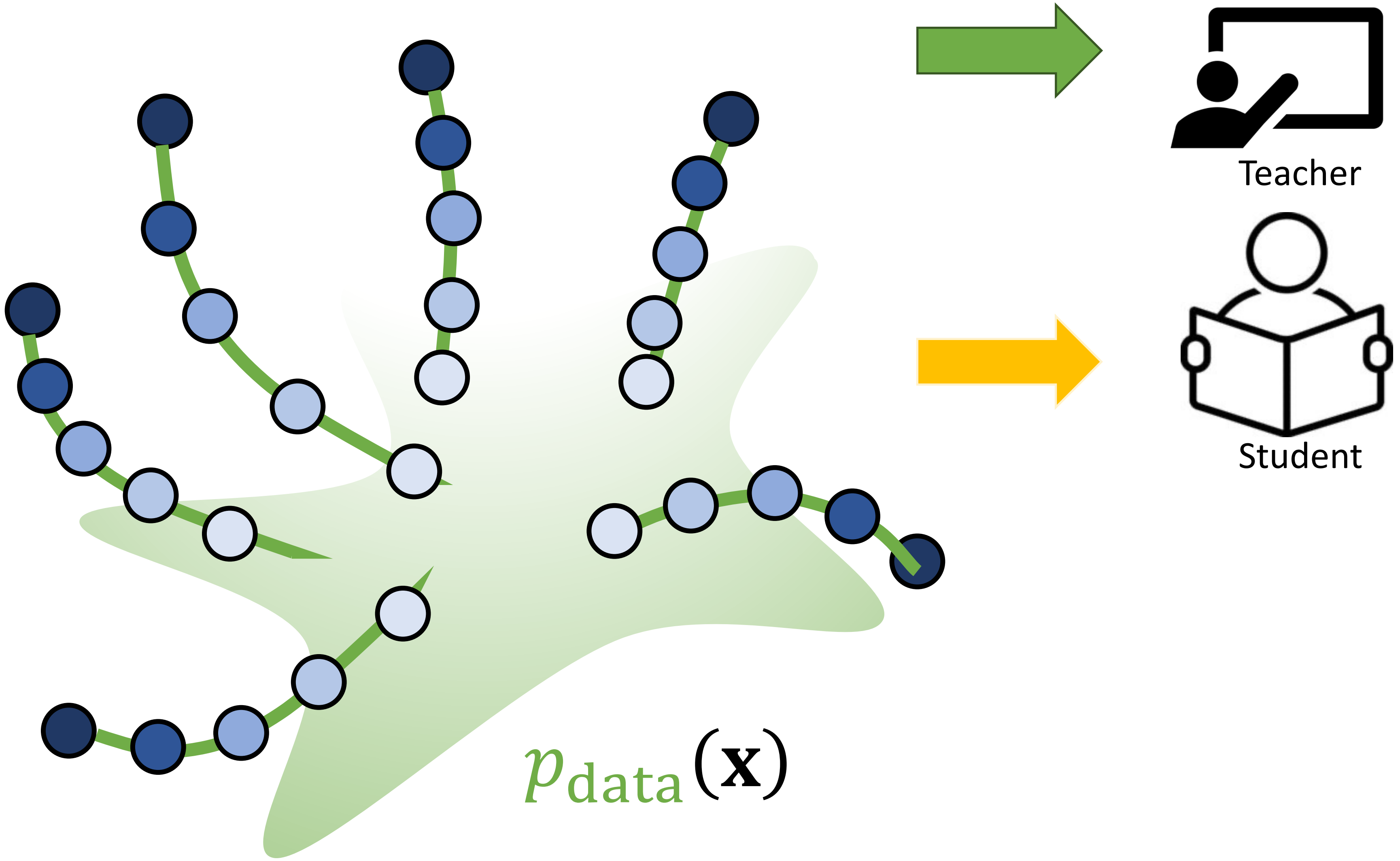
INDIAN INSTITUTE OF TECHNOLOGY GANDHINAGAR

Google

Diffusion Models

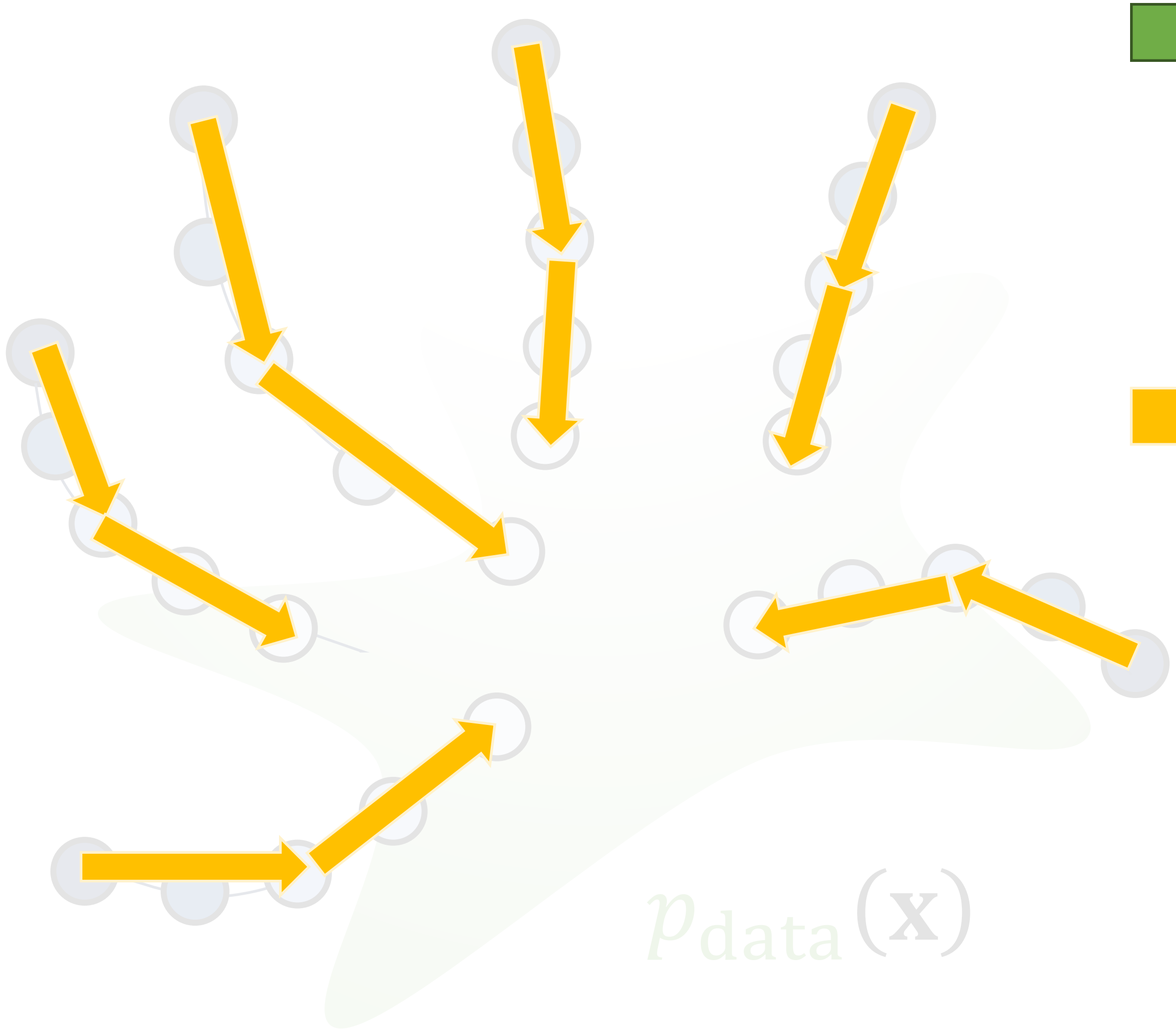
Controllable Image Synthesis Techniques, DiT architecture

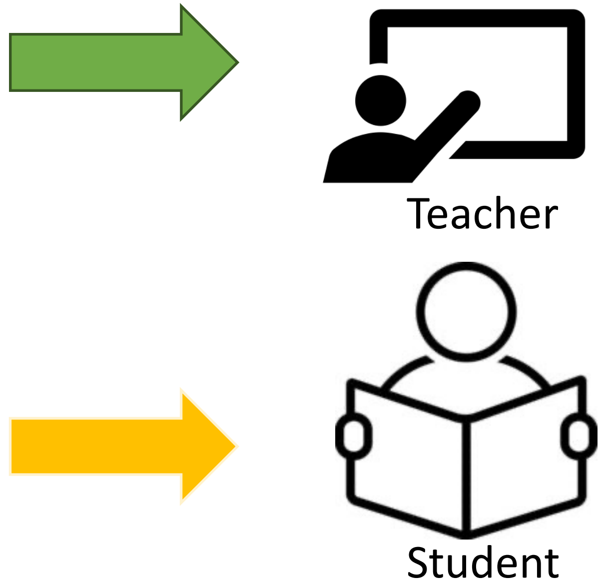
Progressive Distillation



Slide from Jia-Bin Huang

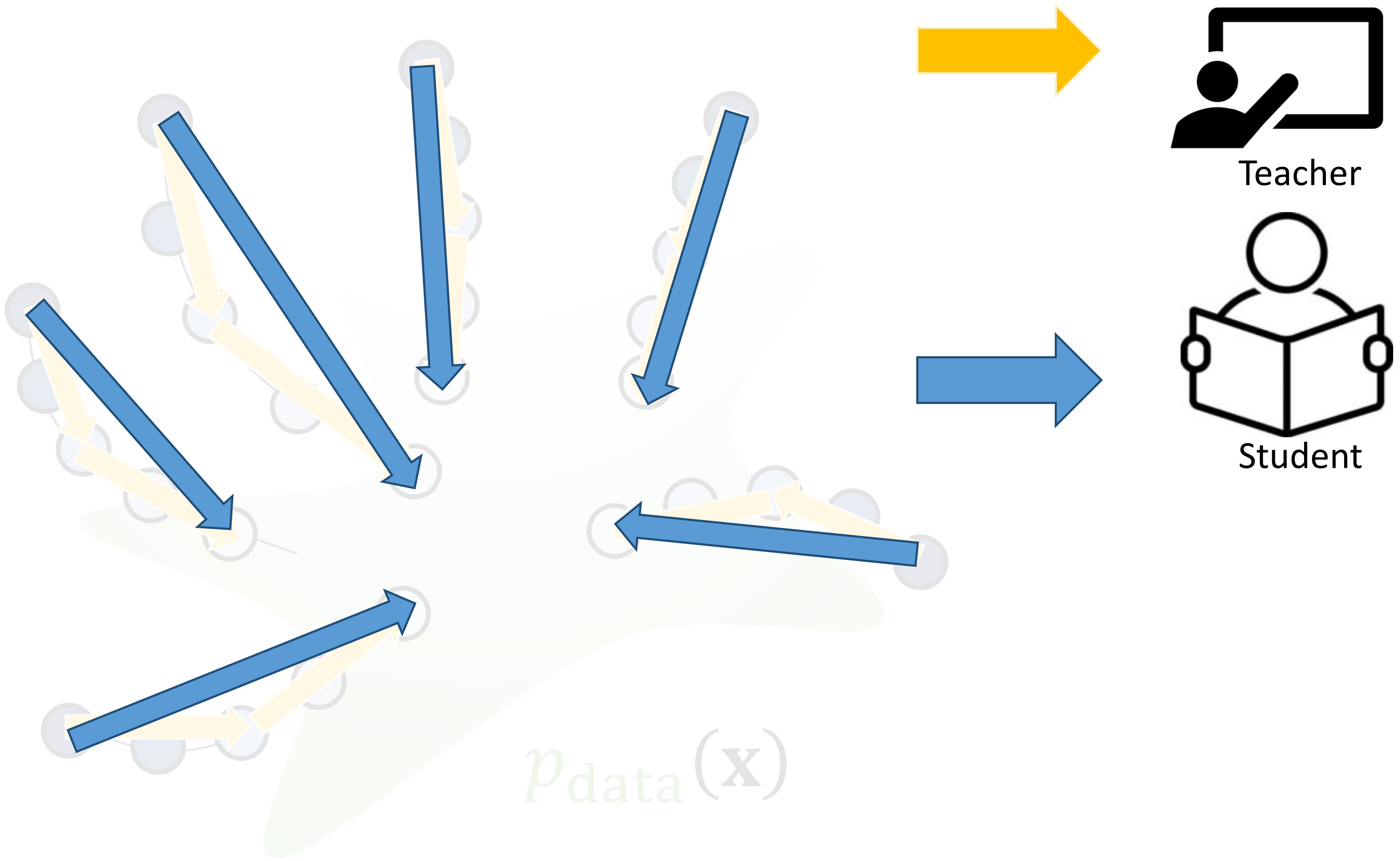
Progressive Distillation





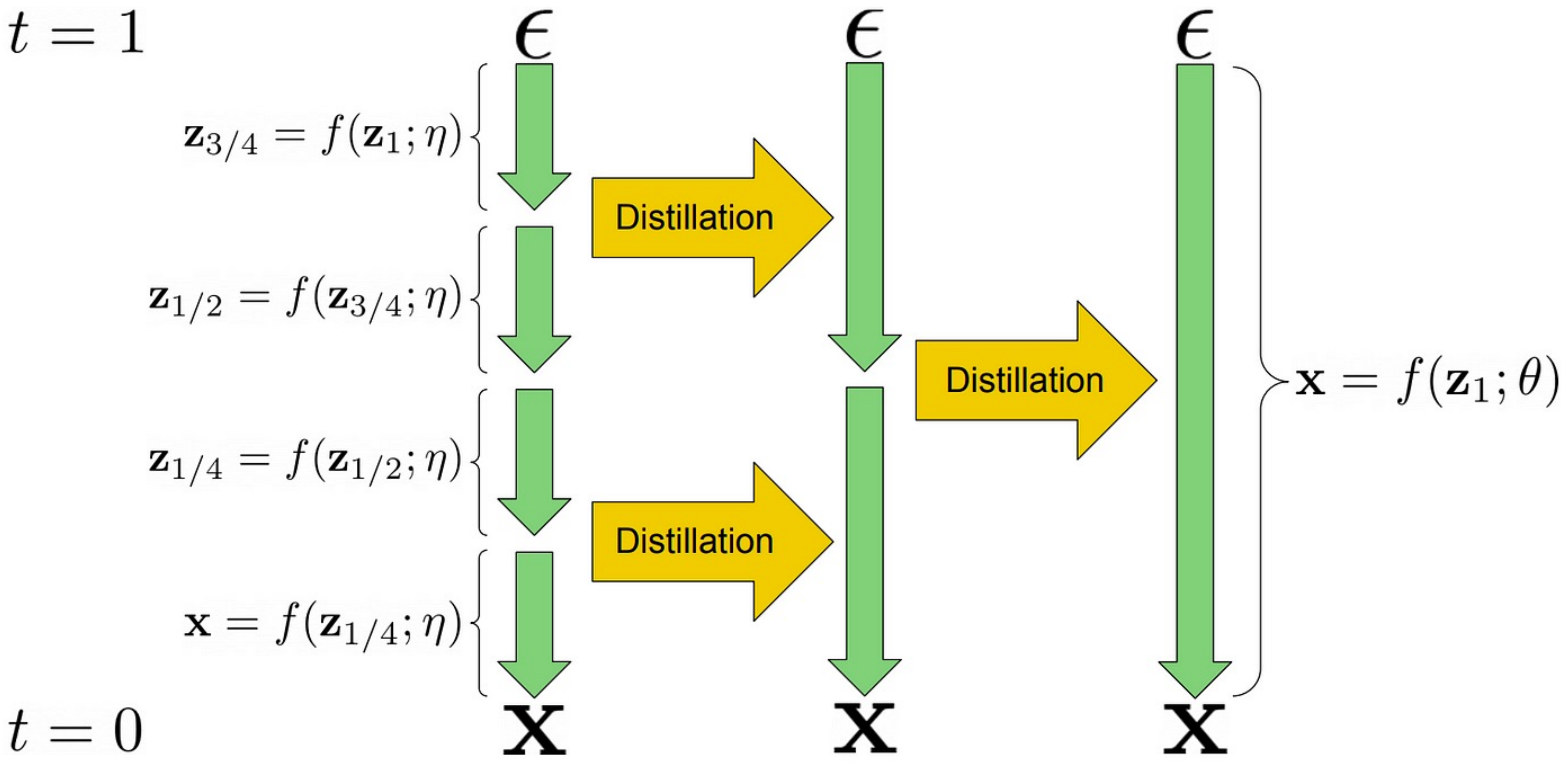
Slide from Jia-Bin Huang

Progressive Distillation



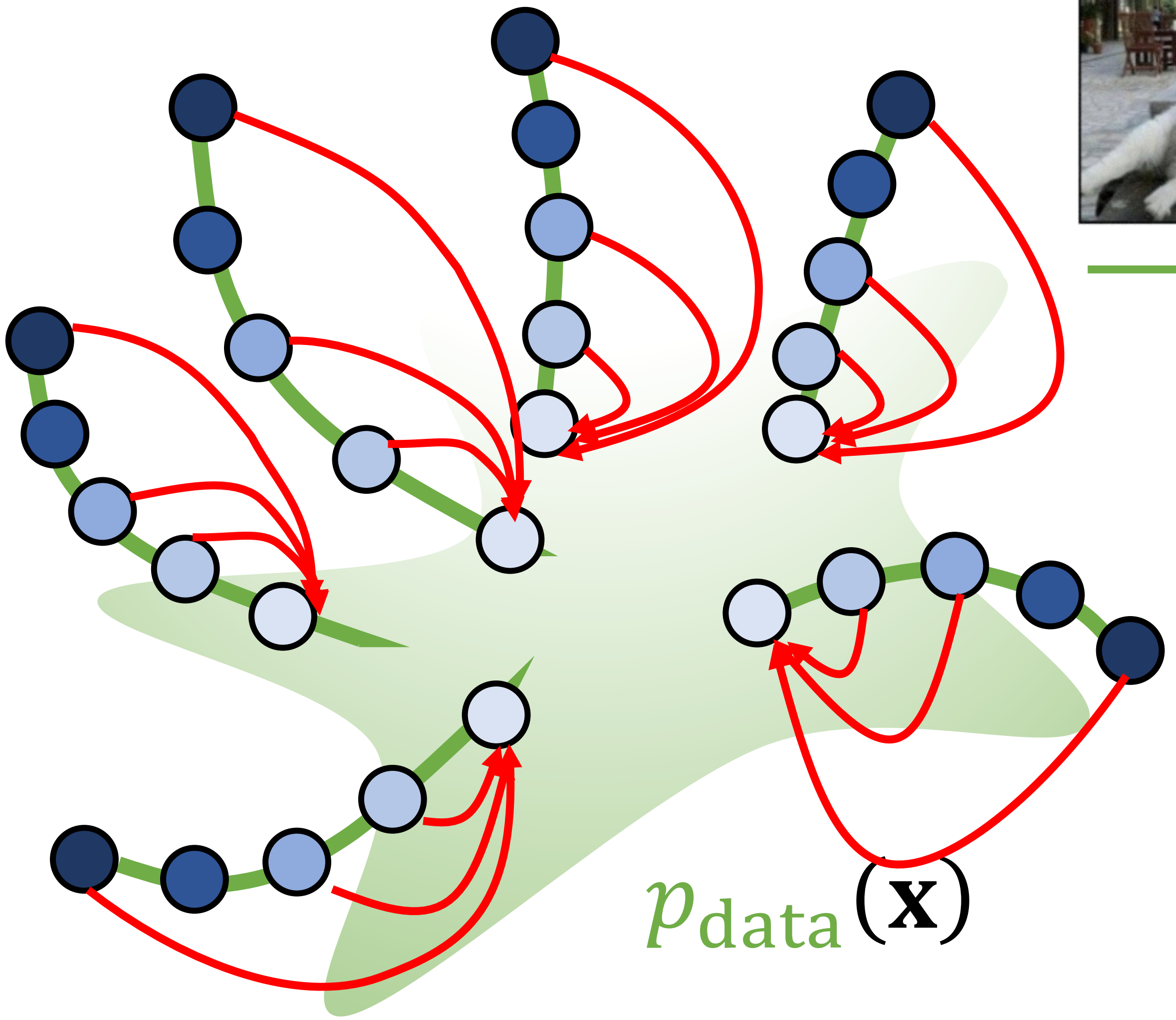
Slide from Jia-Bin Huang

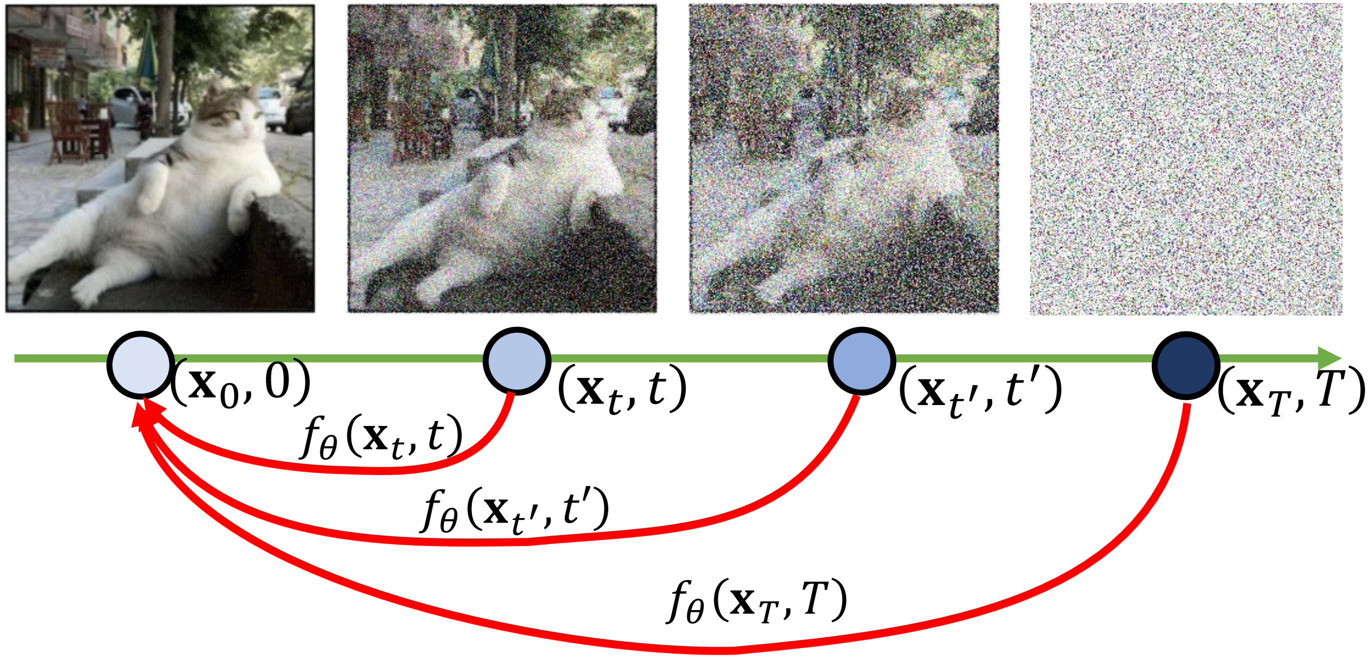
Progressive Distillation



Salimans et al, Progressive distillation for fast sampling of diffusion models

Consistency Models





$$\min_\theta d\Big(f_{EMA}(\mathbf{x}_t, t), f_\theta(\mathbf{x}_{t'}, t')\Big)$$

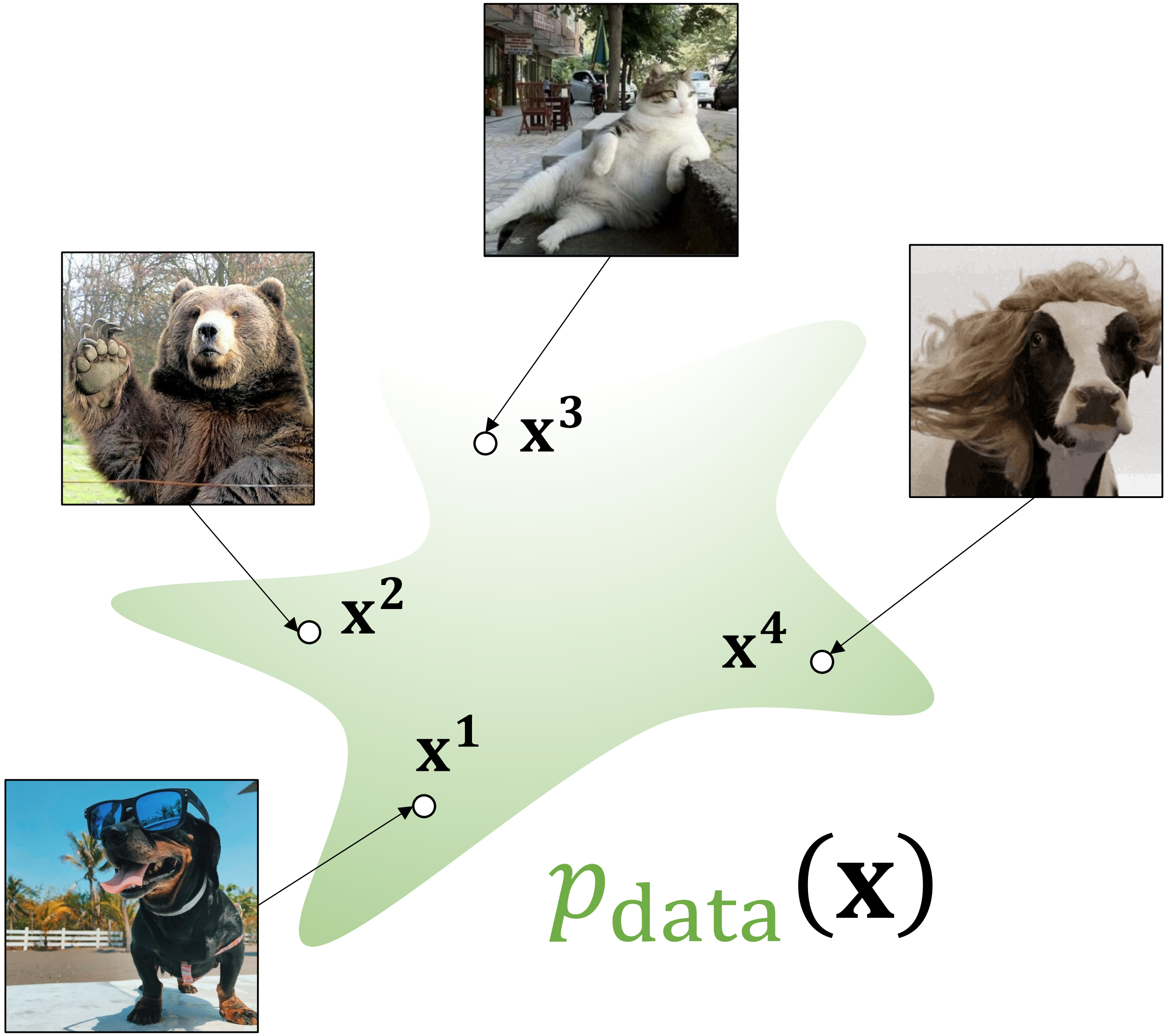
target network (teacher) online network (student)

Slide from Jia-Bin Huang

[Song et al. 2023]

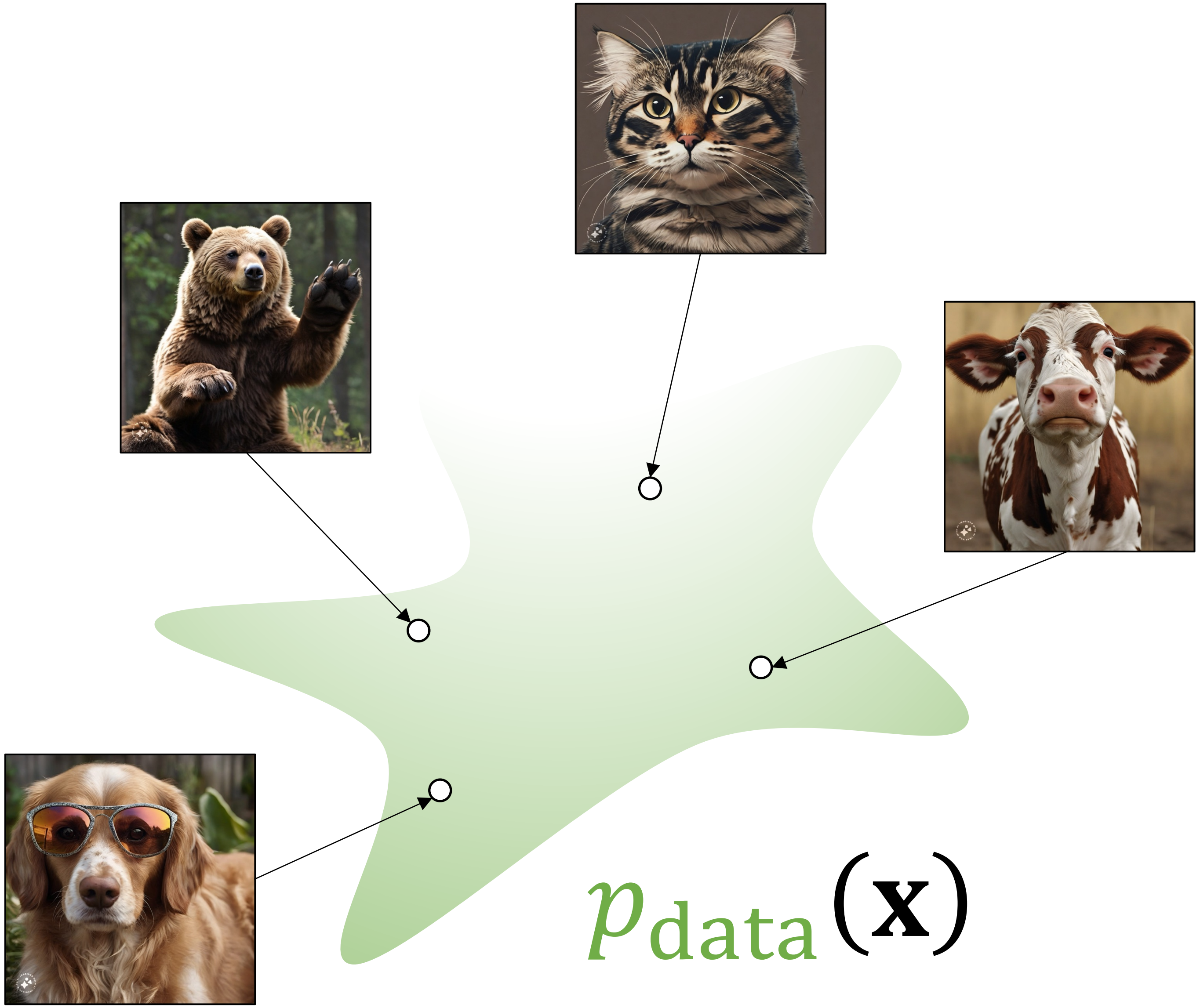
Flow Matching

Latest and the Best



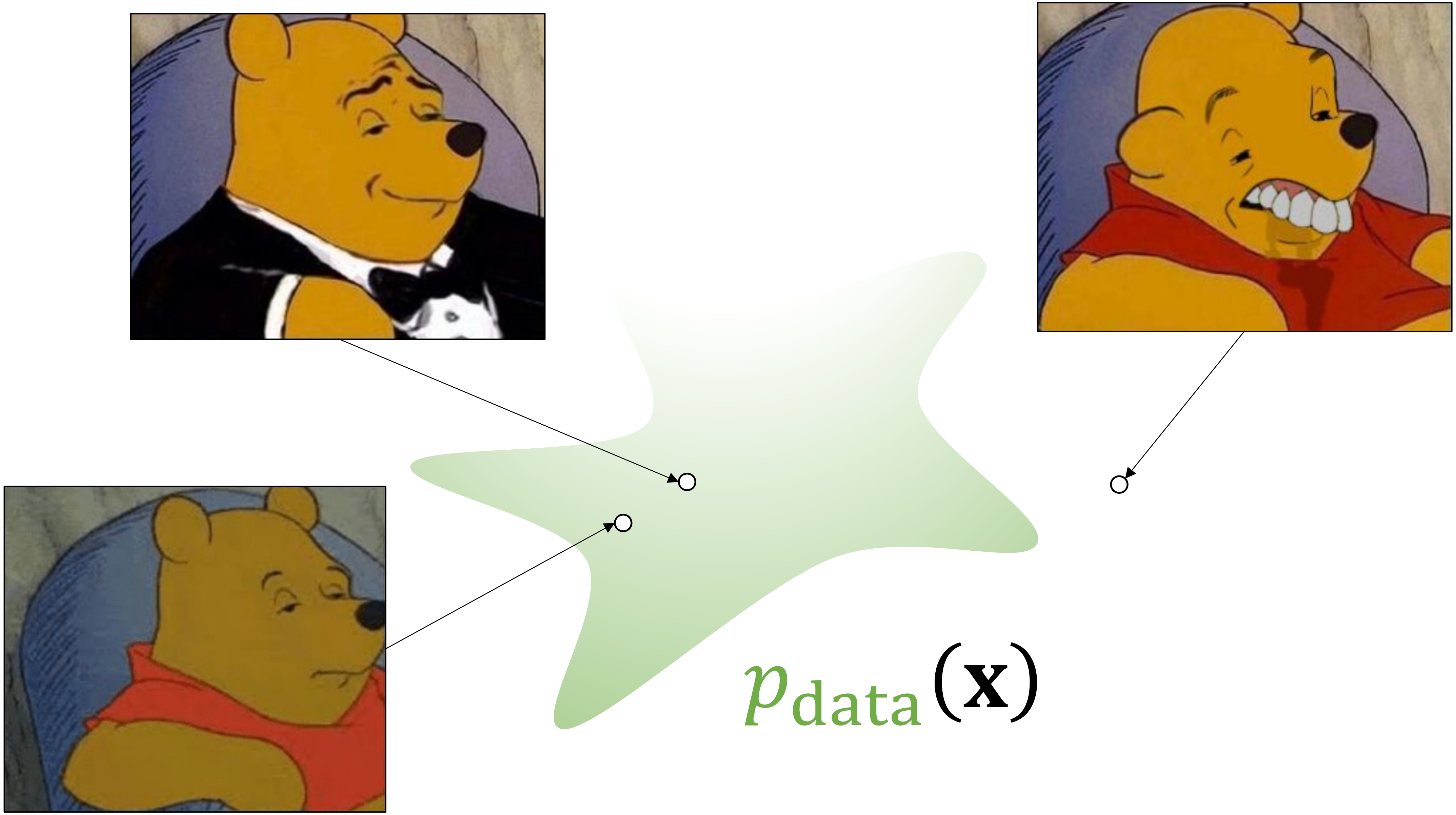
Data Distribution

Slide from Jia-Bin Huang



Data Distribution

Slide from Jia-Bin Huang

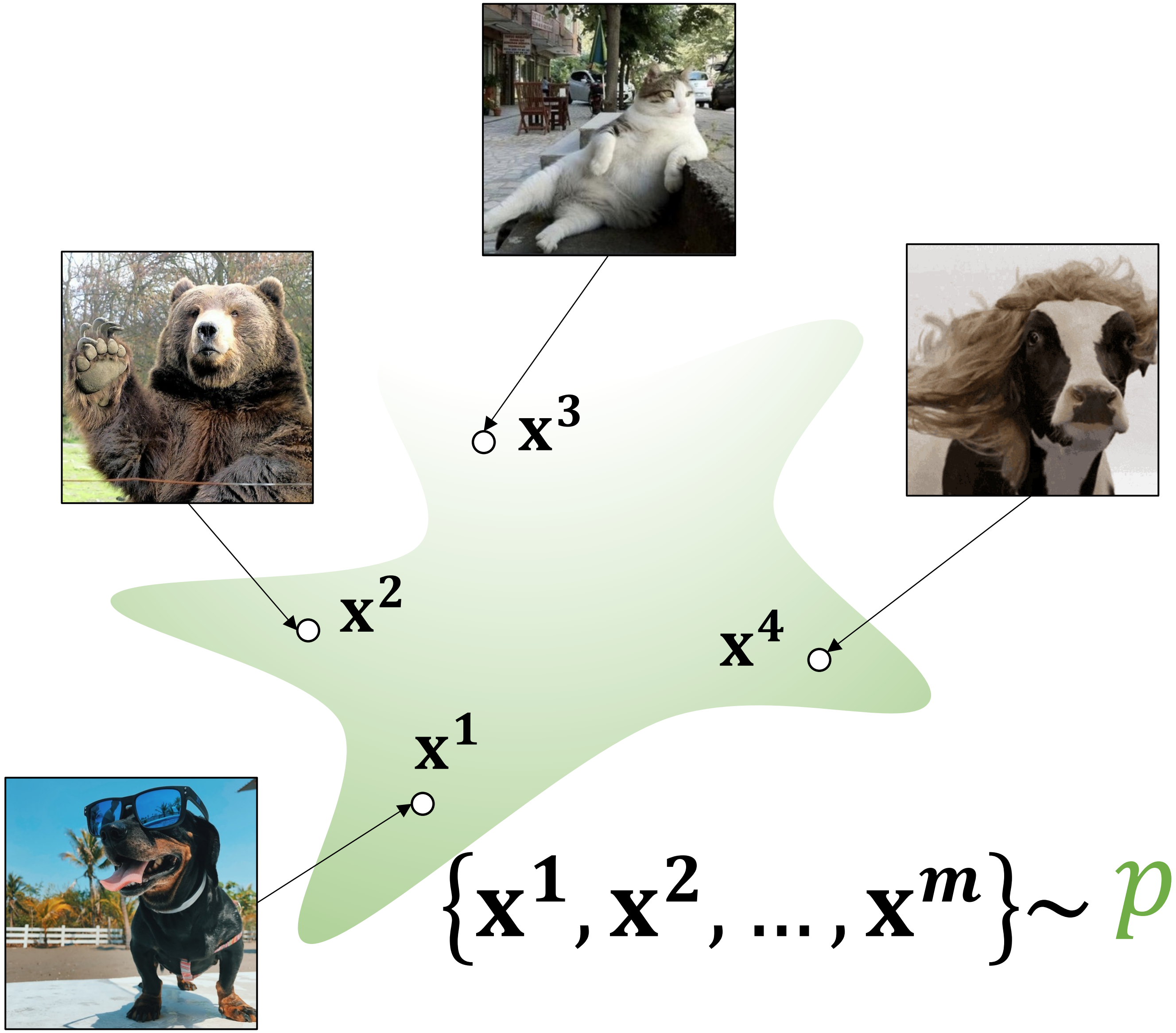


Data Distribution

Slide from Jia-Bin Huang

Data Distribution

Slide from Jia-Bin Huang



$$\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^m\} \sim p_{\text{data}}(\mathbf{x})$$

Data Distribution

Slide from Jia-Bin Huang

$\mathbf{x}^3$

$\mathbf{x}^2$

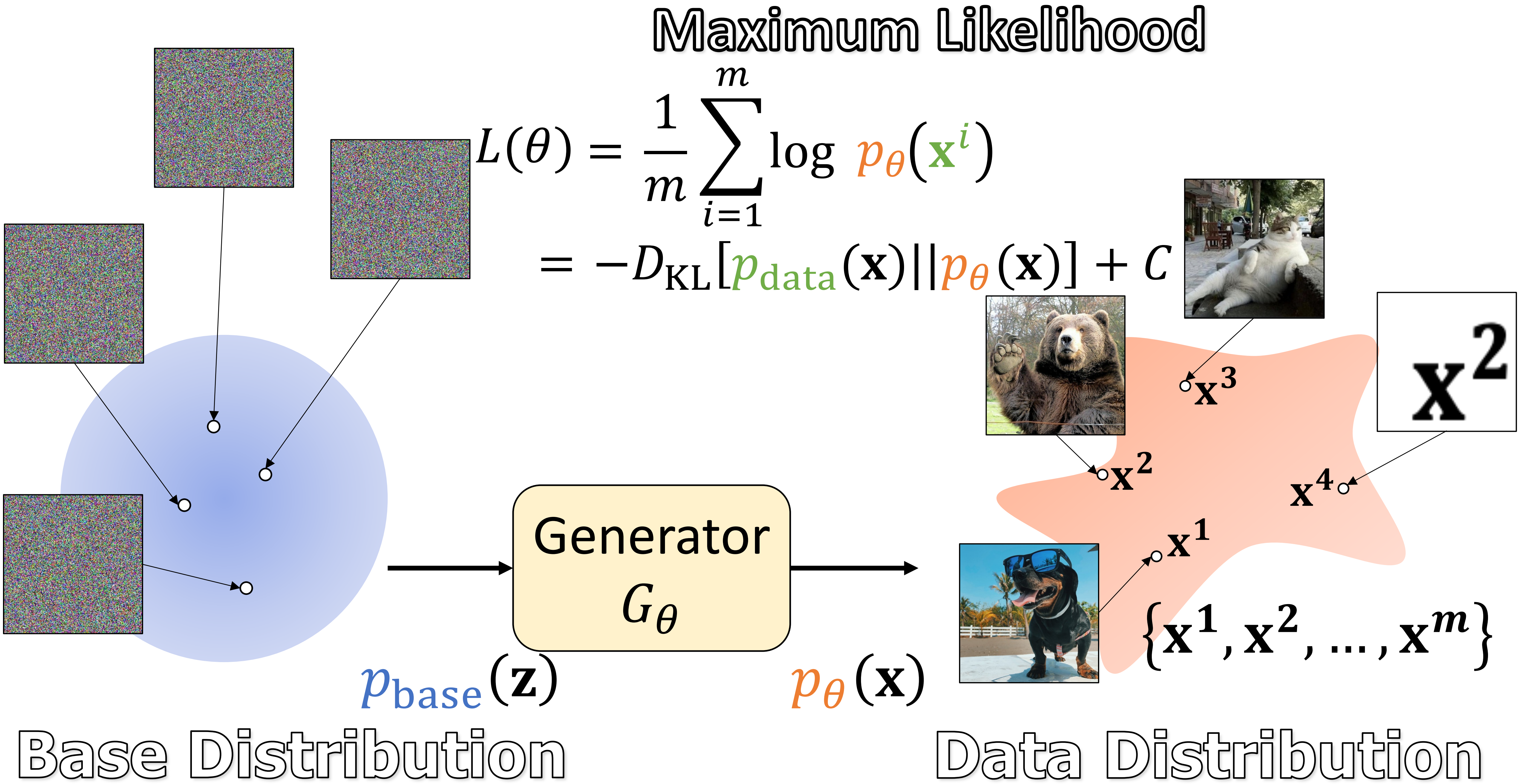
$\mathbf{x}^4$

$\mathbf{x}^1$

$$\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^m\}$$

Data Distribution

Slide from Jia-Bin Huang

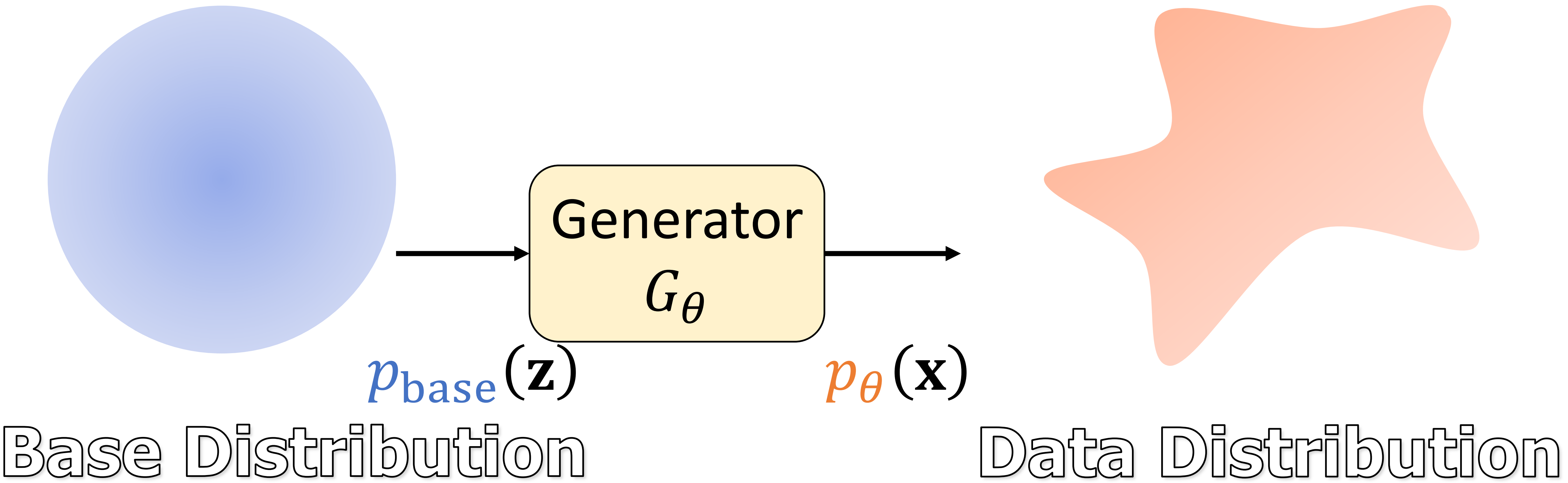


Slide from Jia-Bin Huang

Maximum Likelihood

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\, p_\theta(\mathbf{x}^i)$$

$$= -D_{\mathrm{KL}}[p_{\mathrm{data}}(\mathbf{x})||p_\theta(\mathbf{x})] + C$$

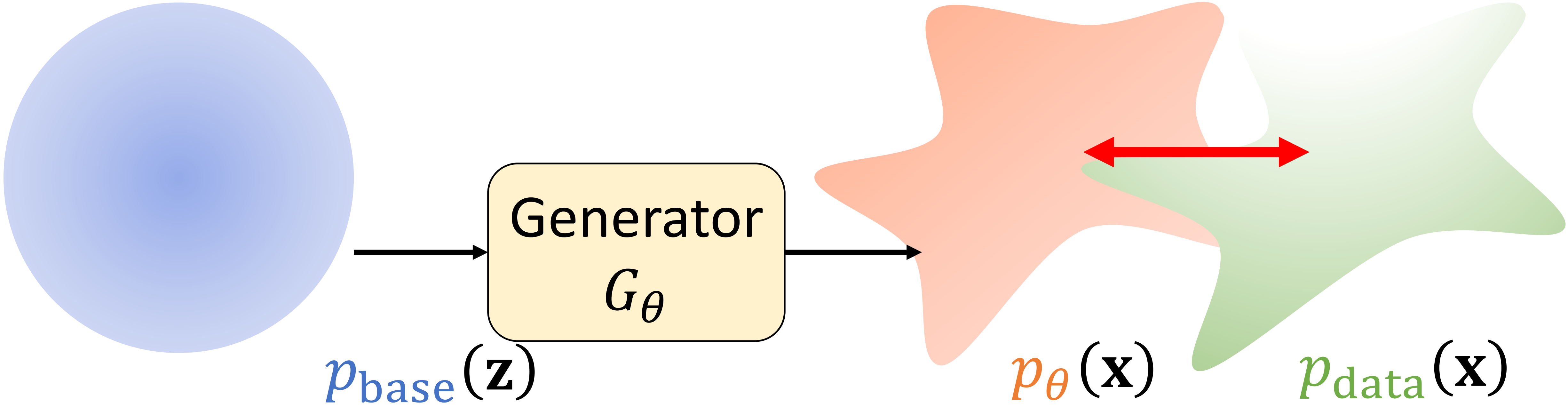


Slide from Jia-Bin Huang

Maximum Likelihood

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\ p_\theta(\mathbf{x}^i)$$

$$= -D_{\mathrm{KL}}[p_{\mathrm{data}}(\mathbf{x})||p_\theta(\mathbf{x})] + C$$



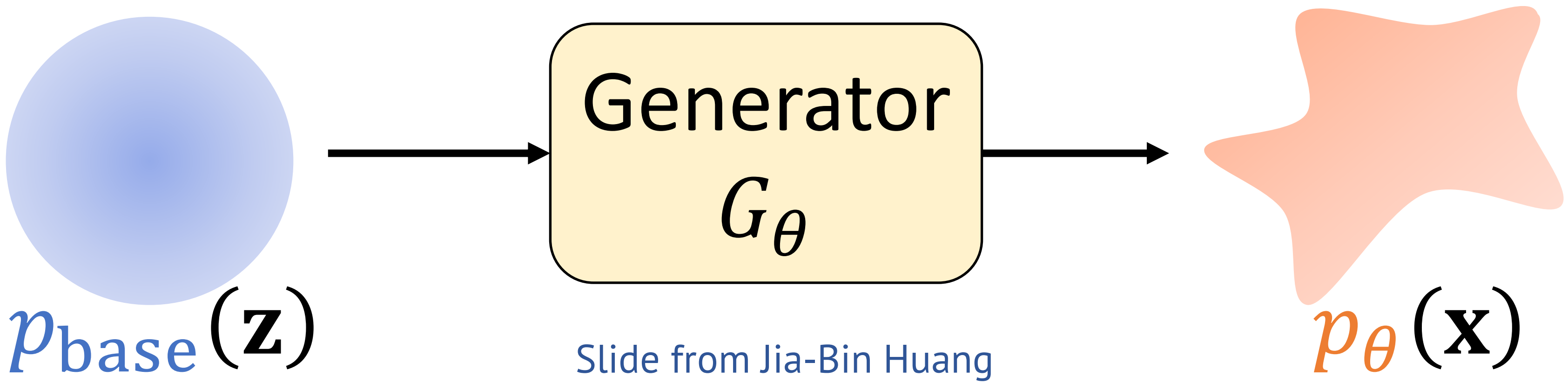
Base Distribution

Data Distribution

Slide from Jia-Bin Huang

Maximum Likelihood

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\ p_\theta(\mathbf{x}^i)$$

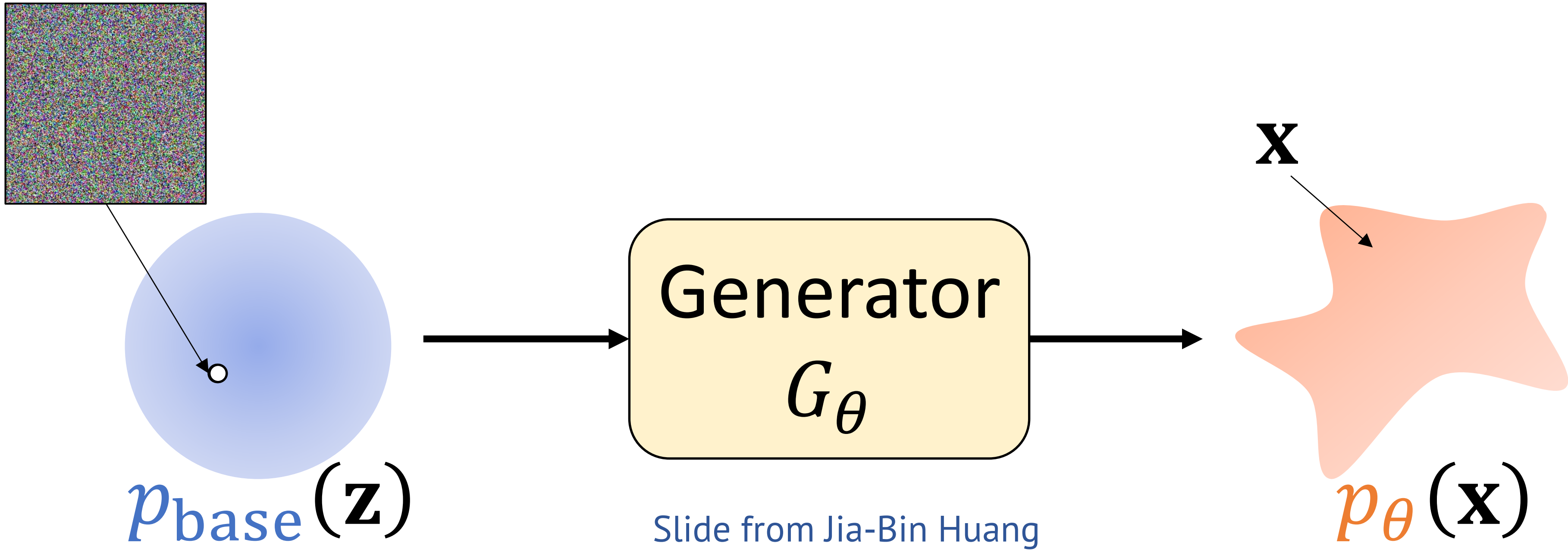


Slide from Jia-Bin Huang

Maximum Likelihood

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\ p_\theta(\mathbf{x}^i)$$

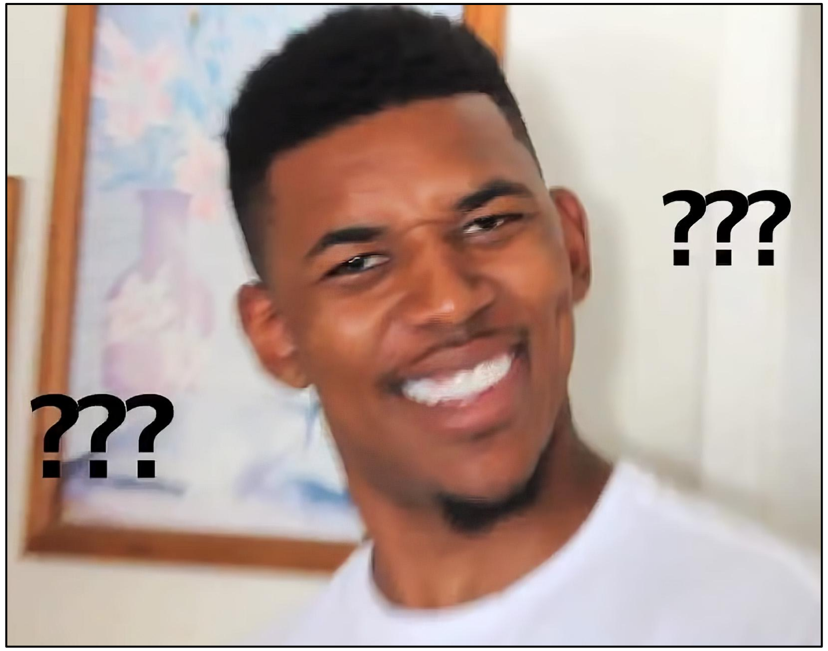
$$\mathbf{z} \sim p_{\text{base}}(\mathbf{z}) \qquad \mathbf{x} = G_\theta(\mathbf{z})$$



Slide from Jia-Bin Huang

Maximum Likelihood

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\ p_\theta(\mathbf{x}^i)$$

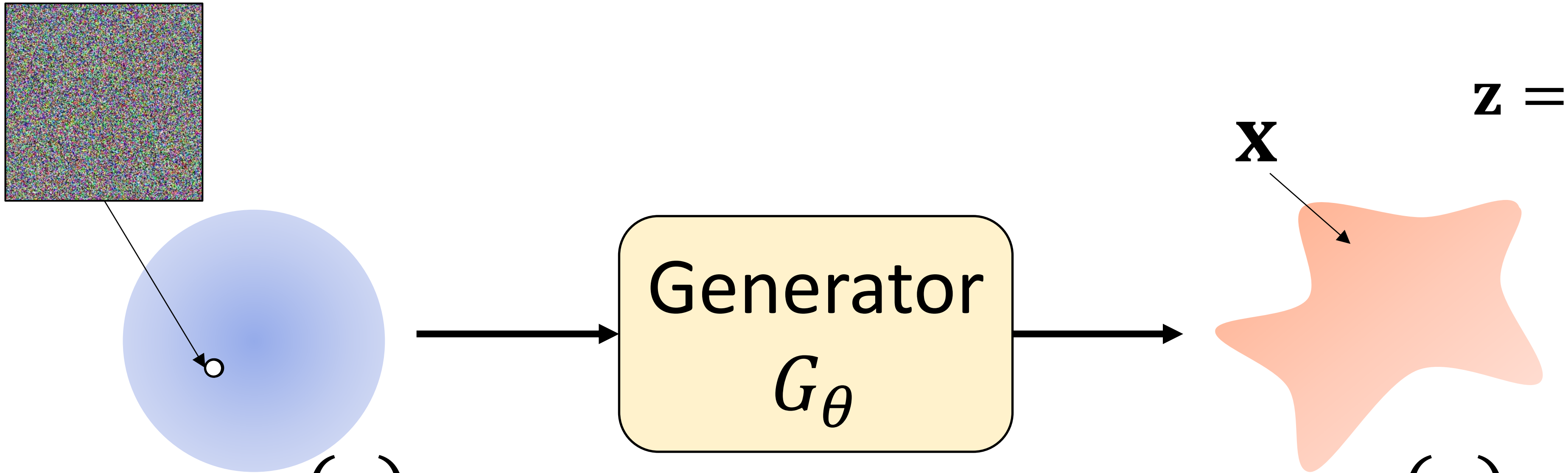


$\mathbf{z} \sim p_{\text{base}}(\mathbf{z})$

$\mathbf{x} = G_\theta(\mathbf{z})$

$p_\theta(\mathbf{x}) = p_{\text{base}}(\mathbf{z})$

$\mathbf{z} = G_\theta^{-1}(\mathbf{x})$

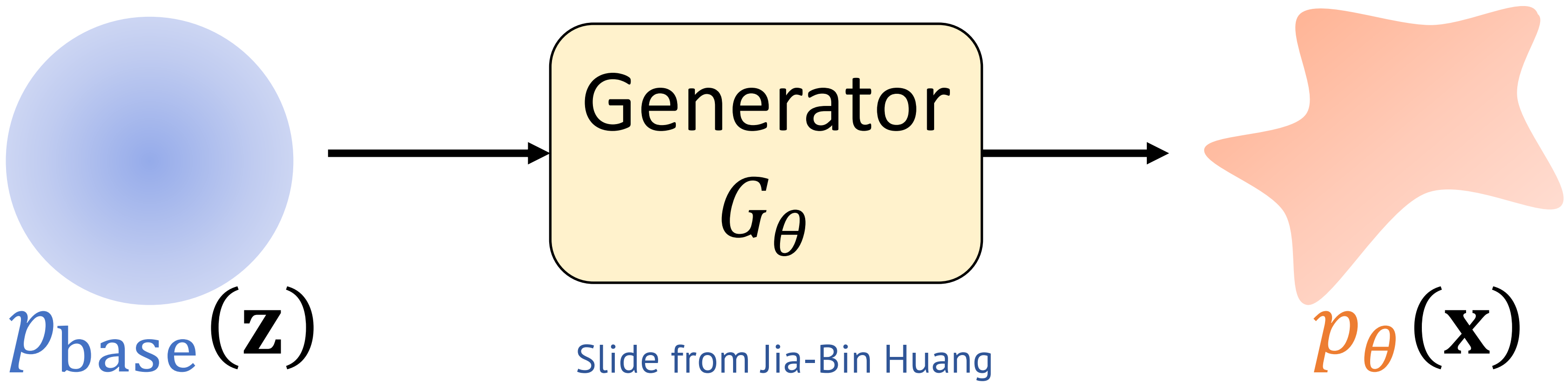


$p_{\text{base}}(\mathbf{z})$

$p_\theta(\mathbf{x})$

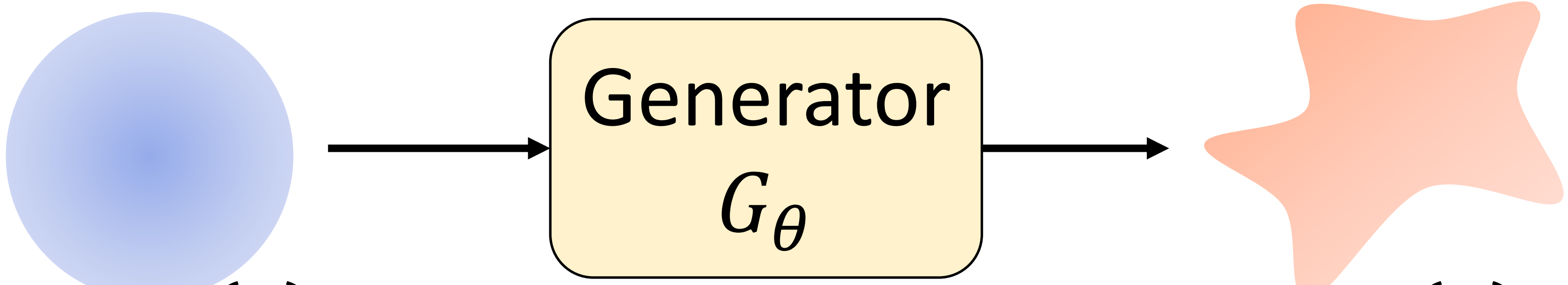
Slide from Jia-Bin Huang

$\mathbf{z} \sim p_{\text{base}}(\mathbf{z})$ $\qquad$ $\mathbf{x} = G_\theta(\mathbf{z})$ $\qquad$ $p_\theta(\mathbf{x}) =$?



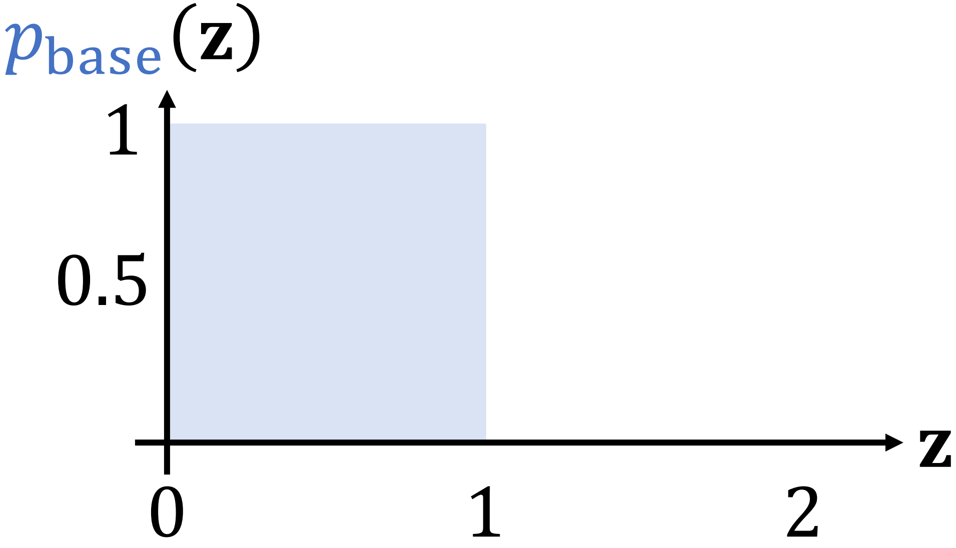
Slide from Jia-Bin Huang

$\mathbf{z} \sim p_{\text{base}}(\mathbf{z})$ $\quad\quad$ $\mathbf{x} = G_\theta(\mathbf{z})$ $\quad\quad$ $p_\theta(\mathbf{x}) = \,?$

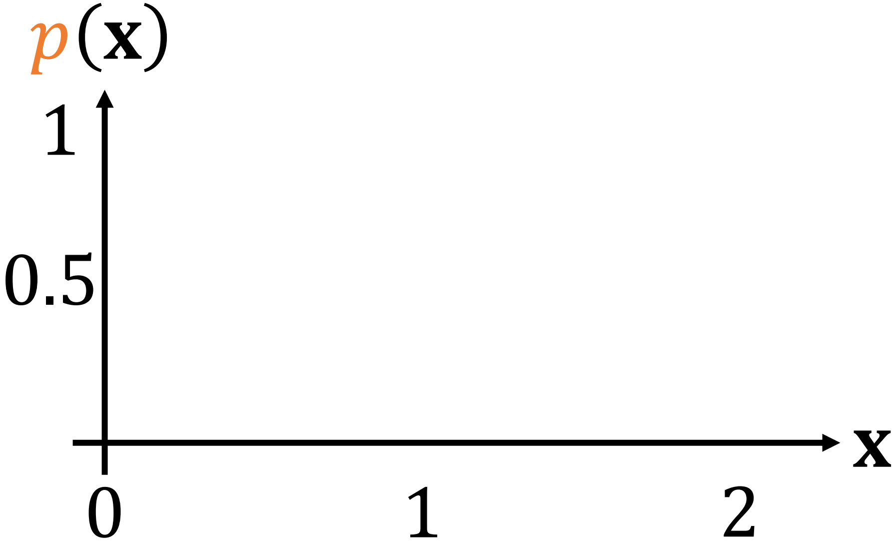


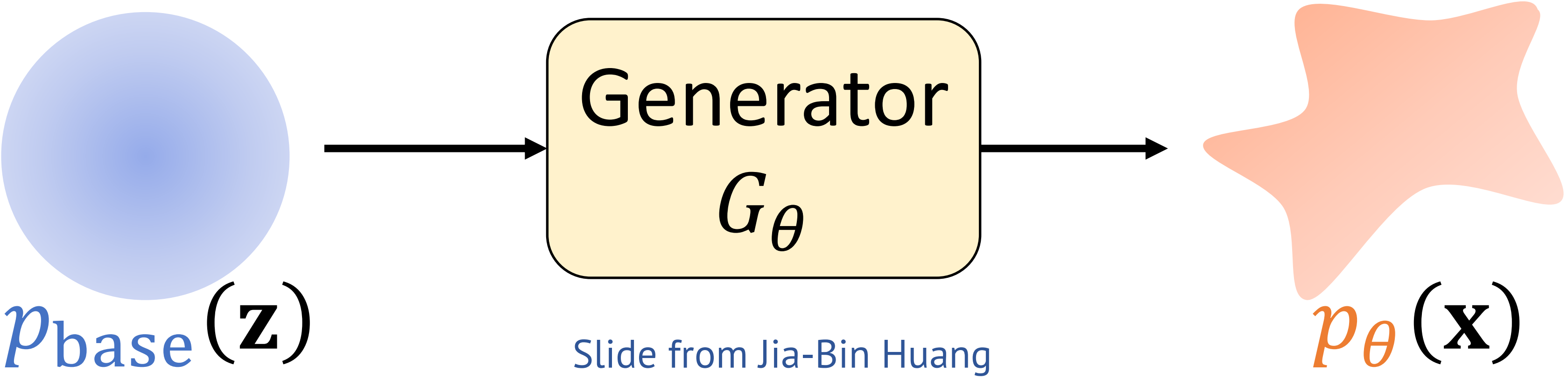
Slide from Jia-Bin Huang

$\mathbf{z} \sim p_{\text{base}}(\mathbf{z})$ $\qquad$ $\mathbf{x} = G(\mathbf{z})$ $\qquad$ $p(\mathbf{x}) = ?$



$\mathbf{x} = 2\mathbf{z}$

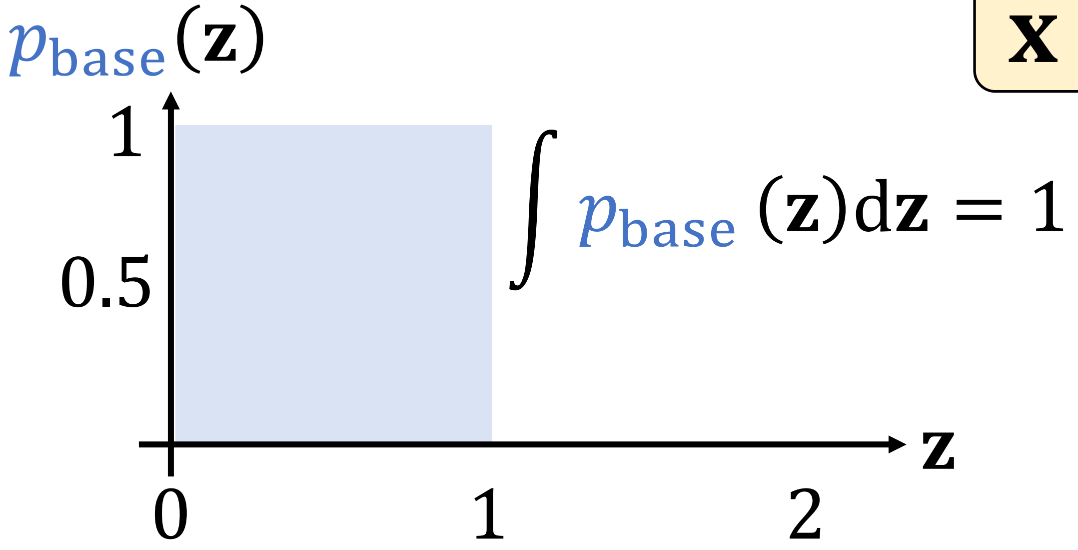


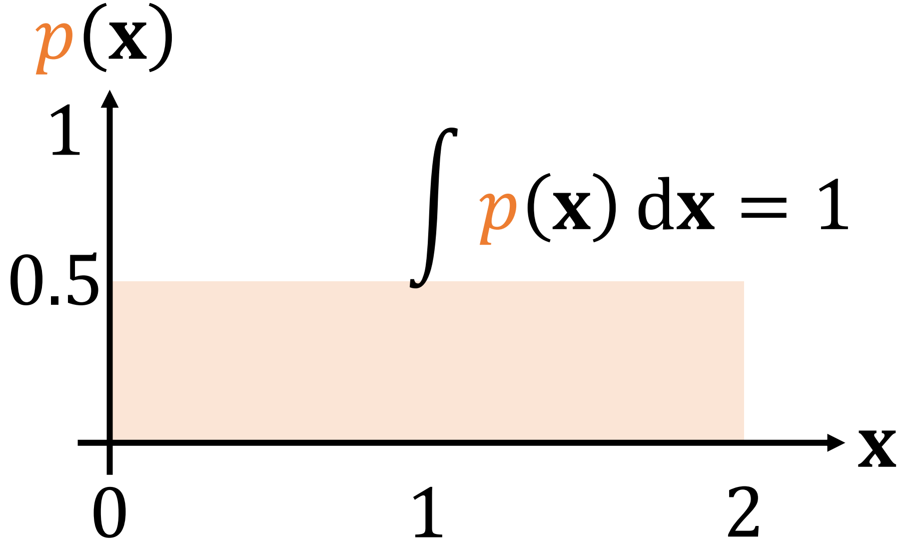


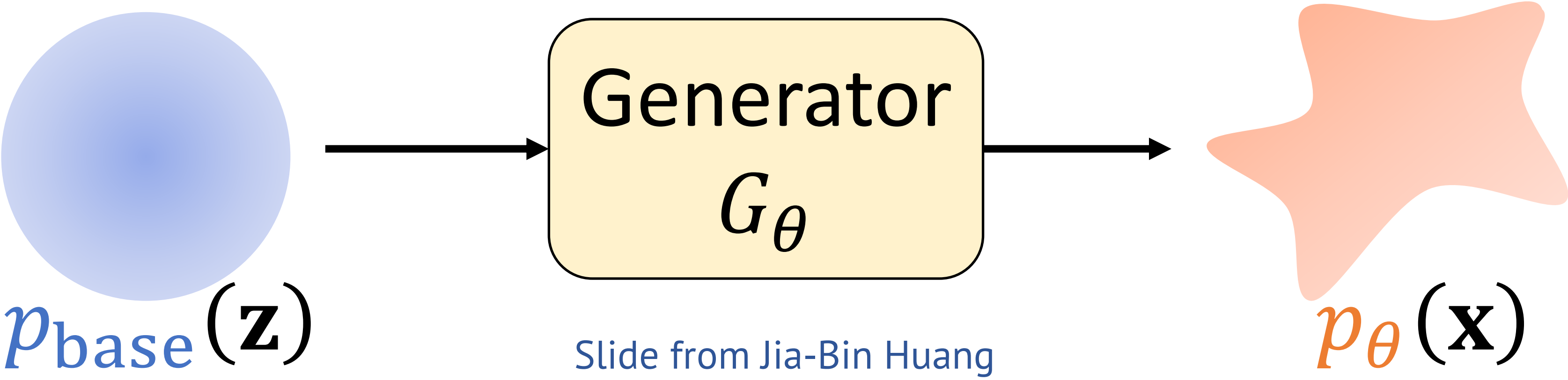
Slide from Jia-Bin Huang

$\mathbf{z} \sim p_{\text{base}}(\mathbf{z})$ $\quad \mathbf{x} = G(\mathbf{z})$ $\quad p(\mathbf{x}) = ?$

$\boxed{\mathbf{x} = 2\mathbf{z}}$



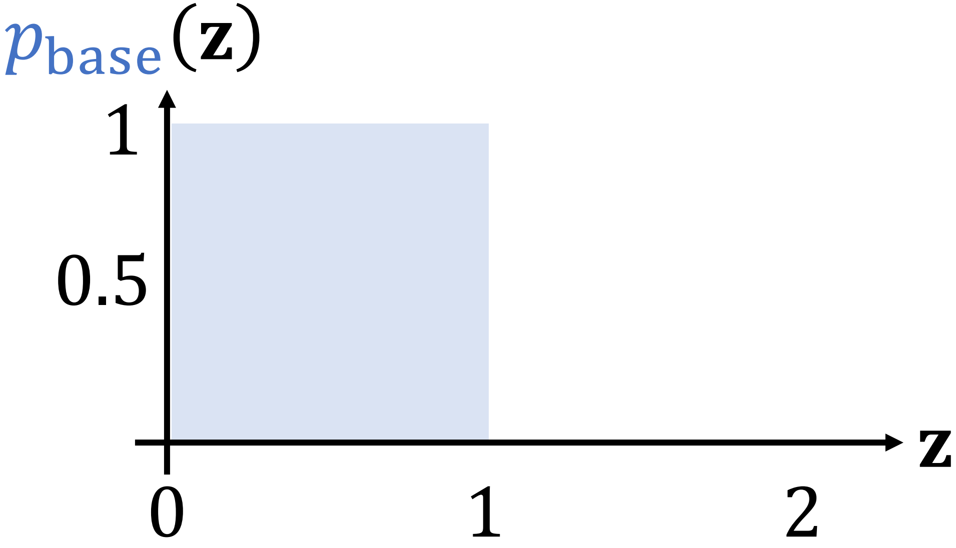




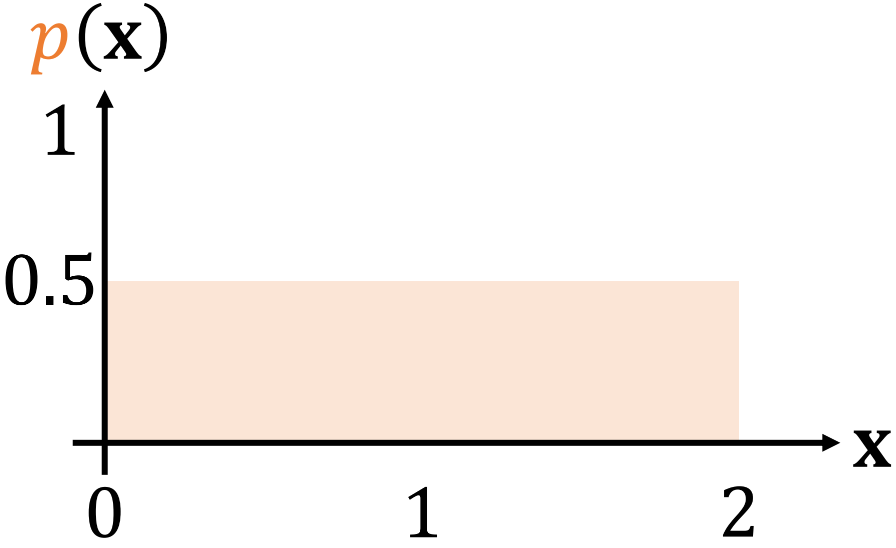
Slide from Jia-Bin Huang

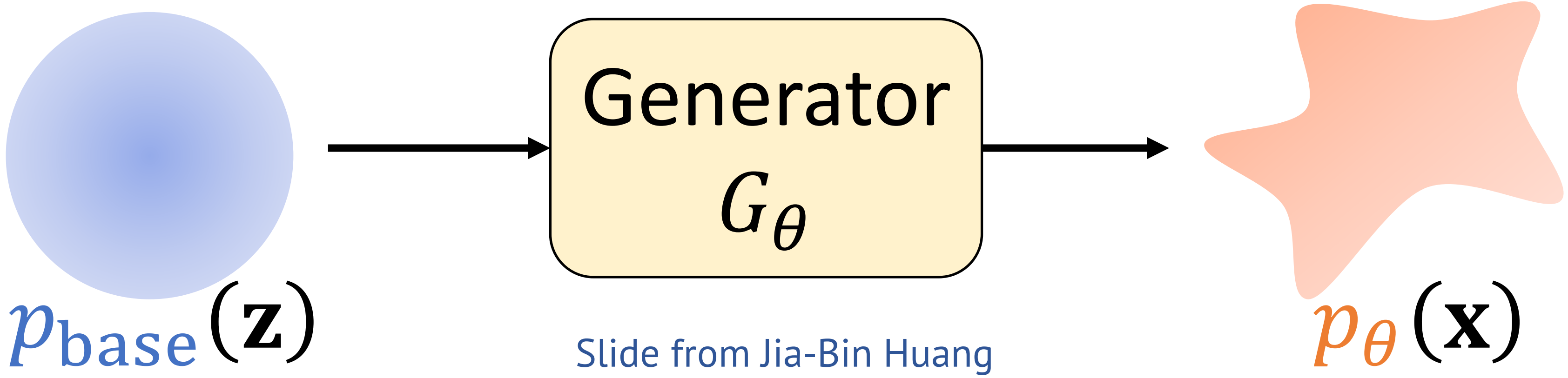
$\mathbf{z} \sim p_{\text{base}}(\mathbf{z})$ $\qquad$ $\mathbf{x} = G(\mathbf{z})$ $\qquad$ $p(\mathbf{x}) = \frac{1}{2} p_{\text{base}}(\mathbf{z})$

$\mathbf{z} = G^{-1}(\mathbf{x})$



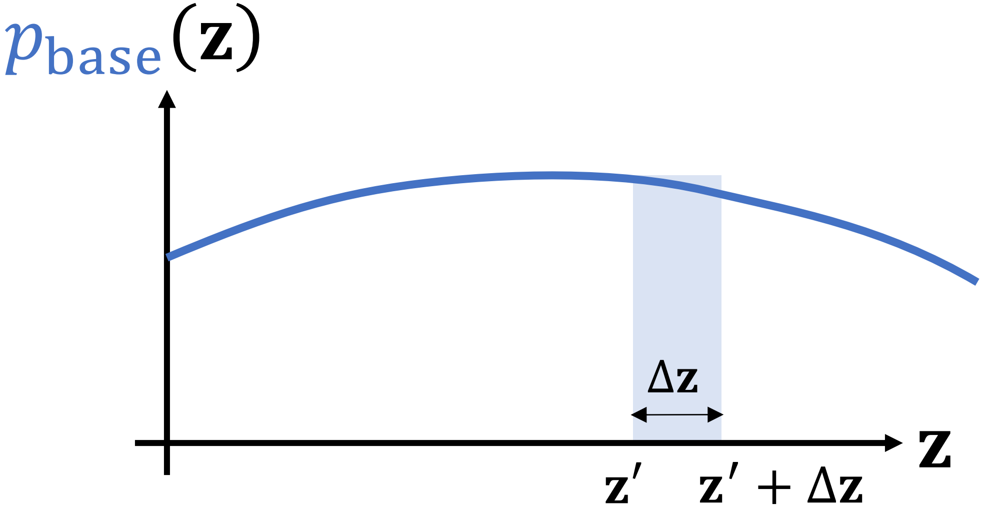
$\boxed{\mathbf{x} = 2\mathbf{z}}$

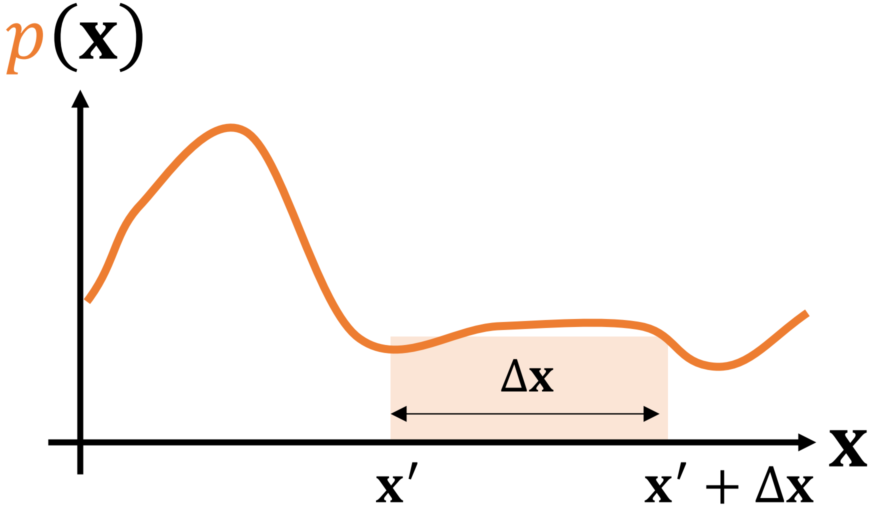


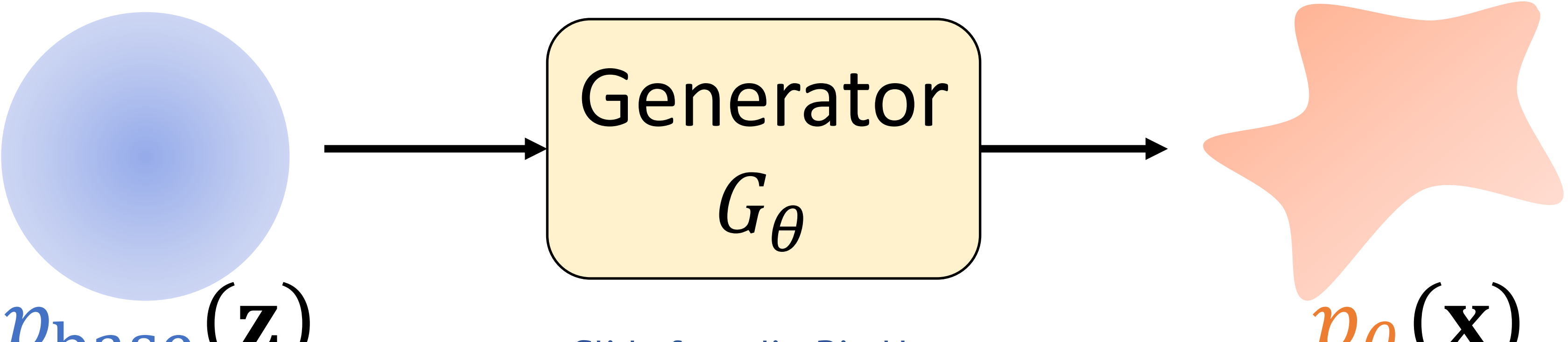


Slide from Jia-Bin Huang

$\mathbf{z} \sim p_{\text{base}}(\mathbf{z}) \qquad \mathbf{x} = G(\mathbf{z}) \qquad p(\mathbf{x}')\Delta\mathbf{x} = p_{\text{base}}(\mathbf{z}')\Delta\mathbf{z}$

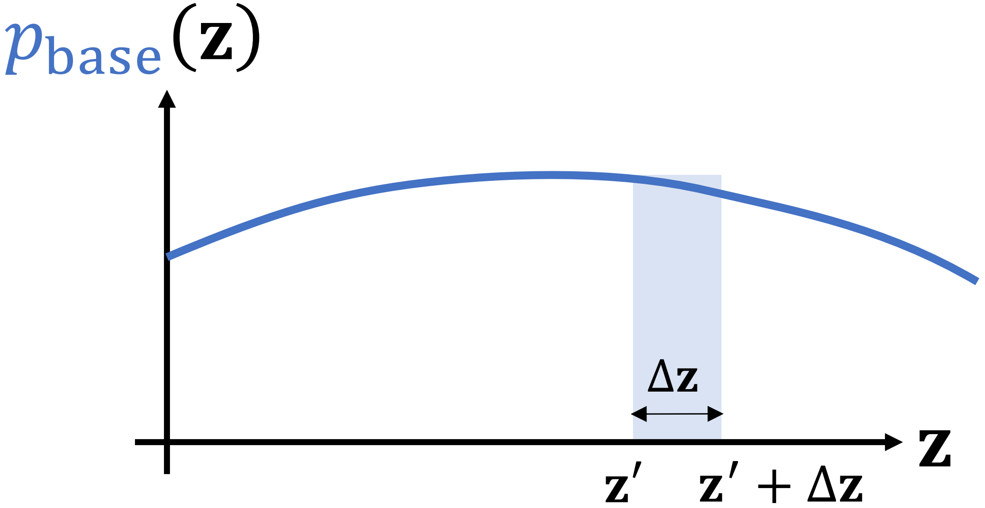


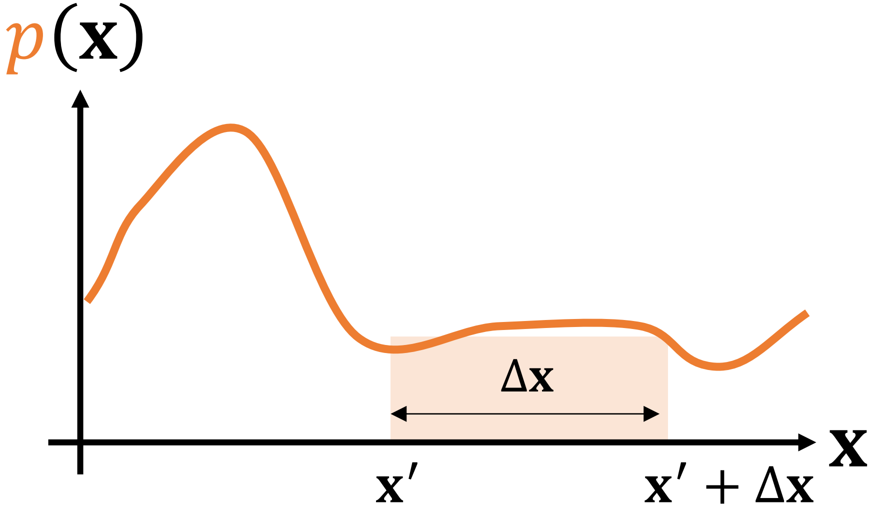


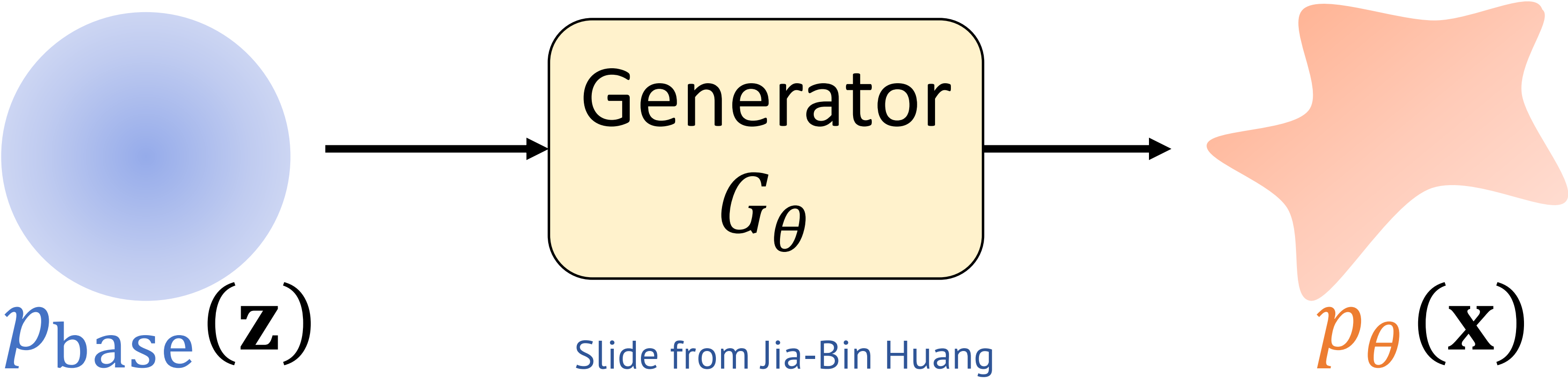


Slide from Jia-Bin Huang

$$\mathbf{z} \sim p_{\text{base}}(\mathbf{z}) \qquad \mathbf{x} = G(\mathbf{z}) \qquad p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}')\,\frac{\Delta \mathbf{z}}{\Delta \mathbf{x}}$$

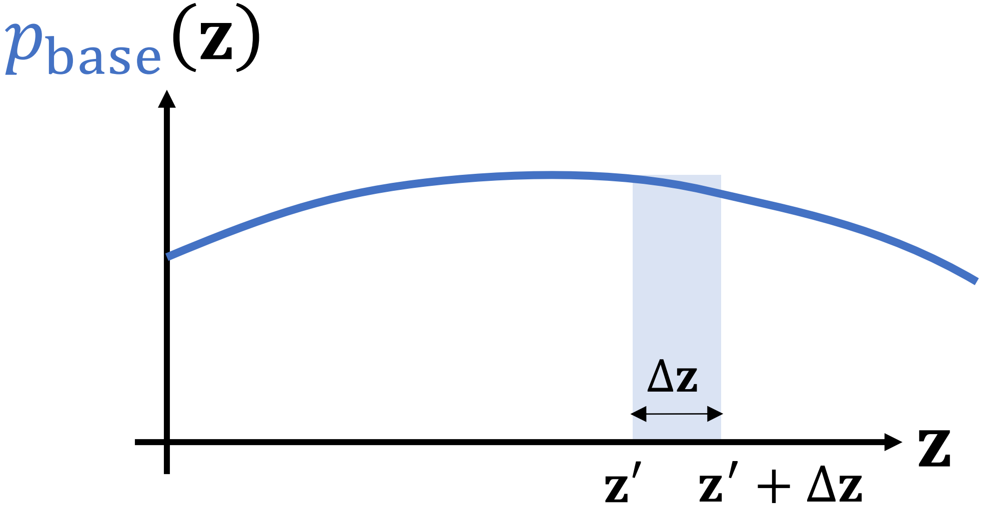


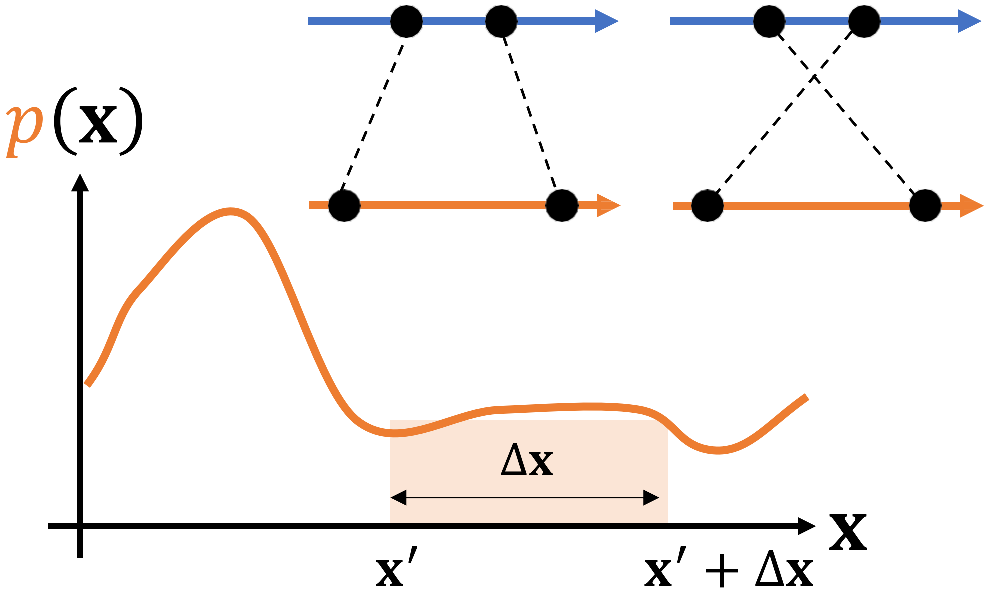


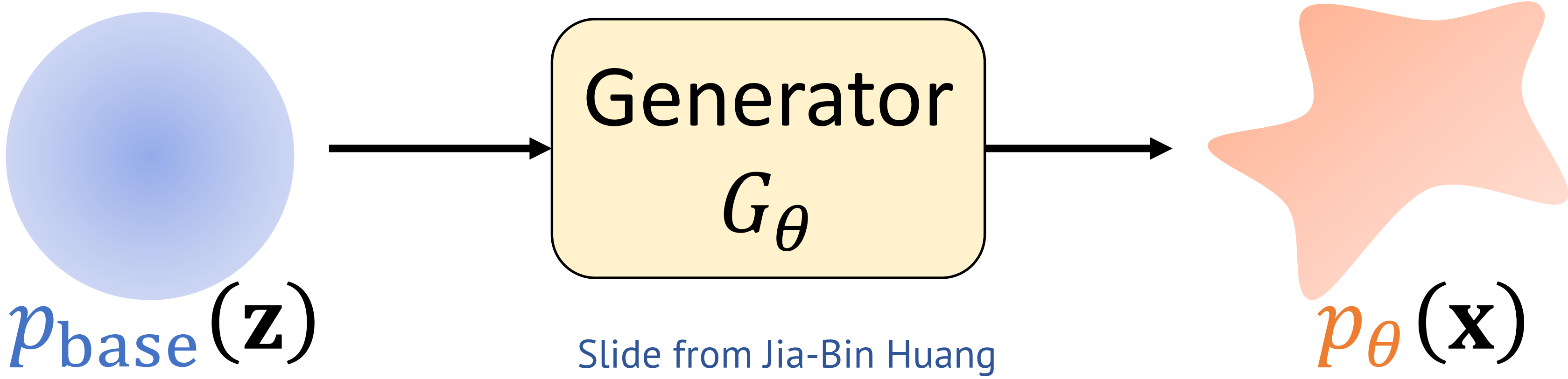


Slide from Jia-Bin Huang

$\mathbf{z} \sim p_{\text{base}}(\mathbf{z})$ $\qquad$ $\mathbf{x} = G(\mathbf{z})$ $\qquad$ $p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}')\left|\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right|$





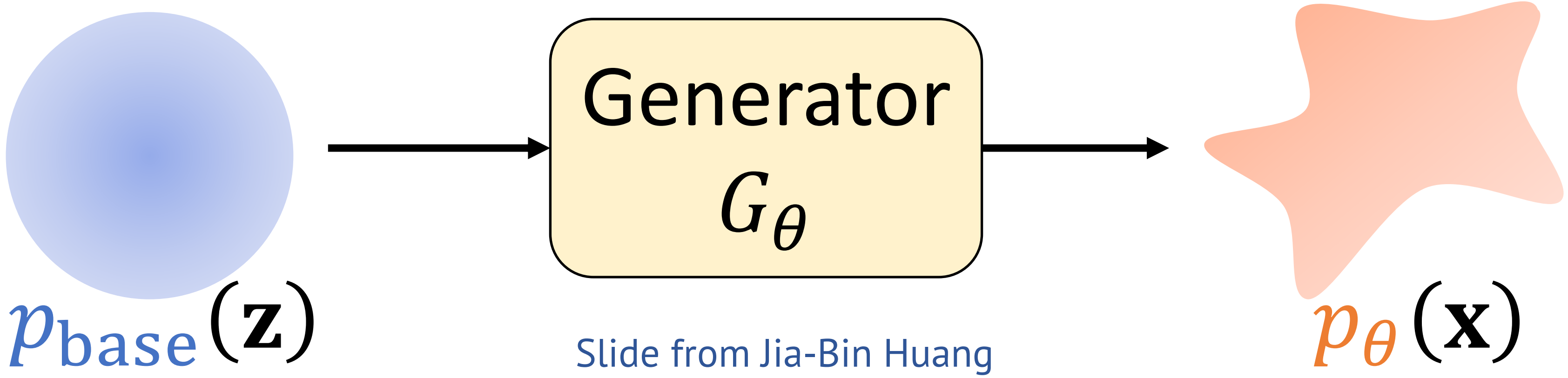


Slide from Jia-Bin Huang

$$\mathbf{z} \sim p_{\text{base}}(\mathbf{z}) \qquad \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \end{bmatrix} = G\left(\begin{bmatrix} \mathrm{z}_1 \\ \mathrm{z}_2 \end{bmatrix}\right)$$

Change-of-variable Formula

$$p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}')\,|\det[\,\boldsymbol{J}_{G^{-1}}]|$$



Slide from Jia-Bin Huang

Maximum Likelihood

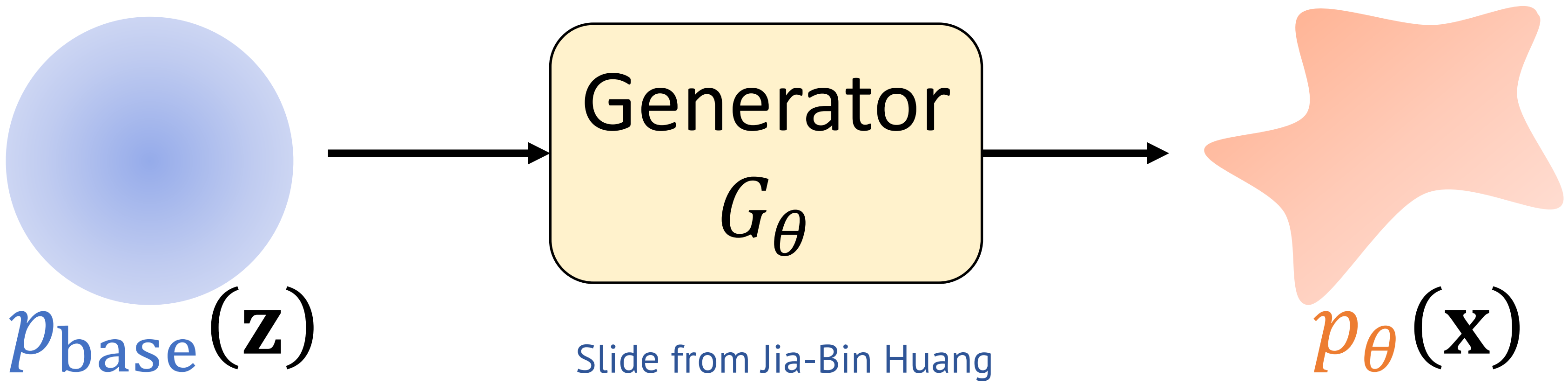
$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\ p_\theta(\mathbf{x}^i)$$

$$p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}')\,|\det[\boldsymbol{J}_{G^{-1}}]|$$

$$\log p_\theta(\mathbf{x}^i) = \log p_{\text{base}}(\mathbf{z}^i) + \log|\det[\boldsymbol{J}_{G^{-1}}]|$$

1 $\boldsymbol{G}$ is invertible $\quad \mathbf{z}^i = \boldsymbol{G}^{-1}(\mathbf{x}^i)$

2 Can compute $|\det[\boldsymbol{J}_{G^{-1}}]|$

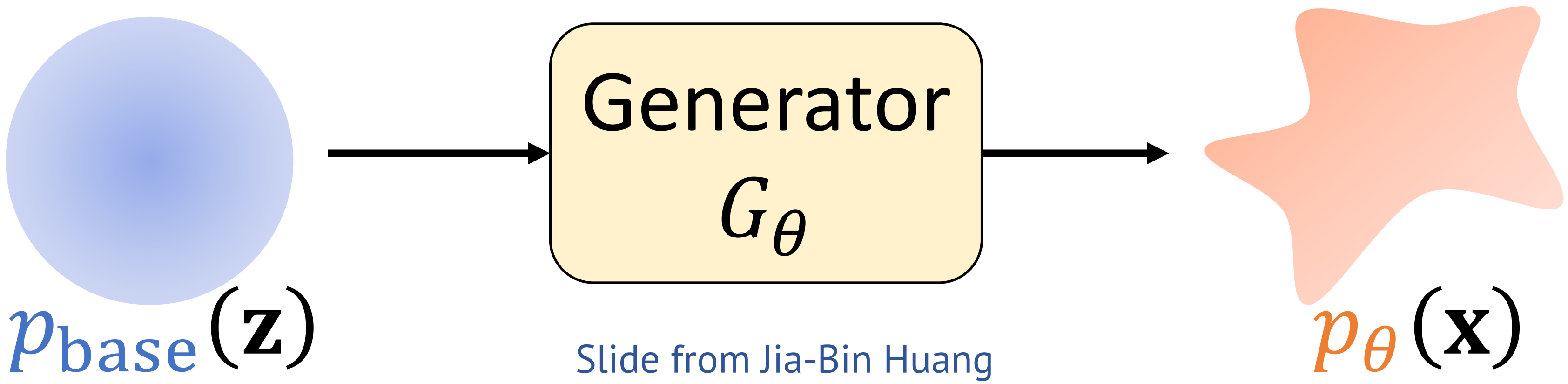


Slide from Jia-Bin Huang

Maximum Likelihood

$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log\, p_\theta(\mathbf{x}^i)$$

$$p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}')\, |\det[\boldsymbol{J}_{\boldsymbol{G}^{-1}}]|$$

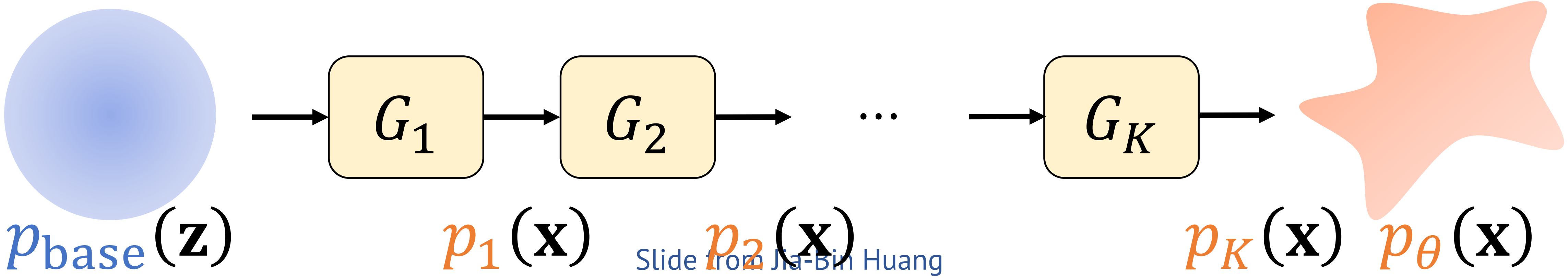


Slide from Jia-Bin Huang

Maximum Likelihood

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\ p_\theta(\mathbf{x}^i)$$

$$p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}')\,|\det[\boldsymbol{J}_{\boldsymbol{G}^{-1}}]|$$



Slide from Jia-Bin Huang

Maximum Likelihood

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log \ p_\theta(\mathbf{x}^i)$$

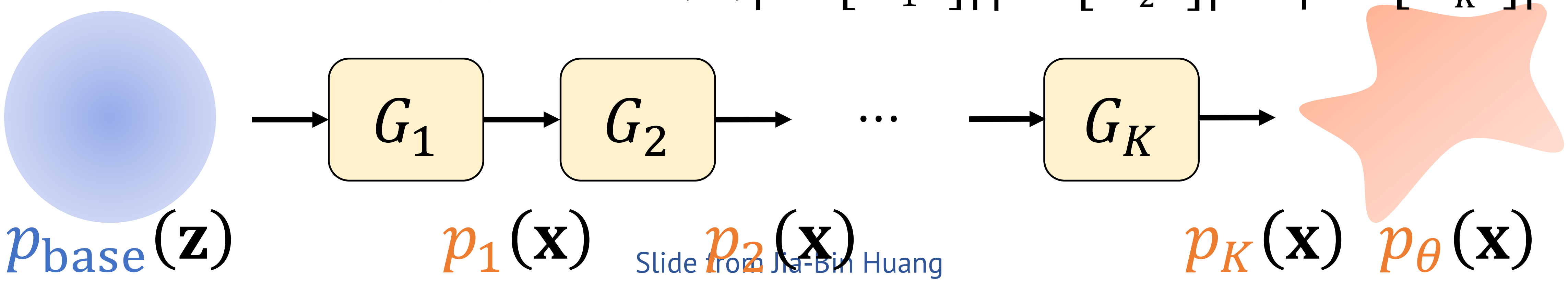
$$p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}')\,\left|\det[\,J_{G^{-1}}]\right|$$

$$p_1(\mathbf{x}^i) = p_{\text{base}}(\mathbf{z}^i)\left|\det\left[J_{G_1^{-1}}\right]\right|$$

$$p_2(\mathbf{x}^i) = p_{\text{base}}(\mathbf{z}^i)\left|\det\left[J_{G_1^{-1}}\right]\right|\left|\det\left[J_{G_2^{-1}}\right]\right|$$

$$\vdots$$

$$p_K(\mathbf{x}^i) = p_{\text{base}}(\mathbf{z}^i)\left|\det\left[J_{G_1^{-1}}\right]\right|\left|\det\left[J_{G_2^{-1}}\right]\right| \cdots \left|\det\left[J_{G_K^{-1}}\right]\right|$$



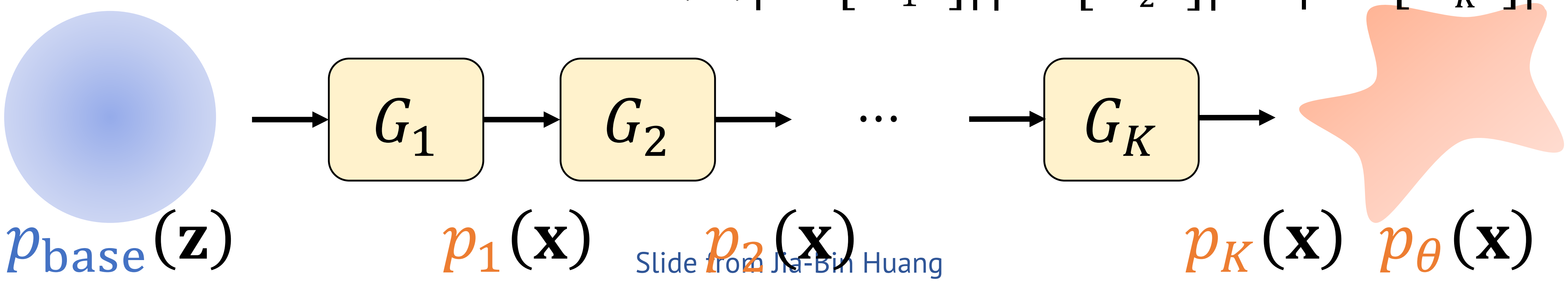
Slide from Jia-Bin Huang

Maximum Likelihood

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\, p_{\theta}(\mathbf{x}^{i})$$

$$p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}')\,\left|\det[\boldsymbol{J}_{G^{-1}}]\right|$$

$$p_K(\mathbf{x}^{i}) = p_{\text{base}}(\mathbf{z}^{i})\left|\det\left[J_{G_1^{-1}}\right]\right|\left|\det\left[J_{G_2^{-1}}\right]\right| \cdots \left|\det\left[J_{G_K^{-1}}\right]\right|$$



Slide from Jia-Bin Huang

Maximum Likelihood

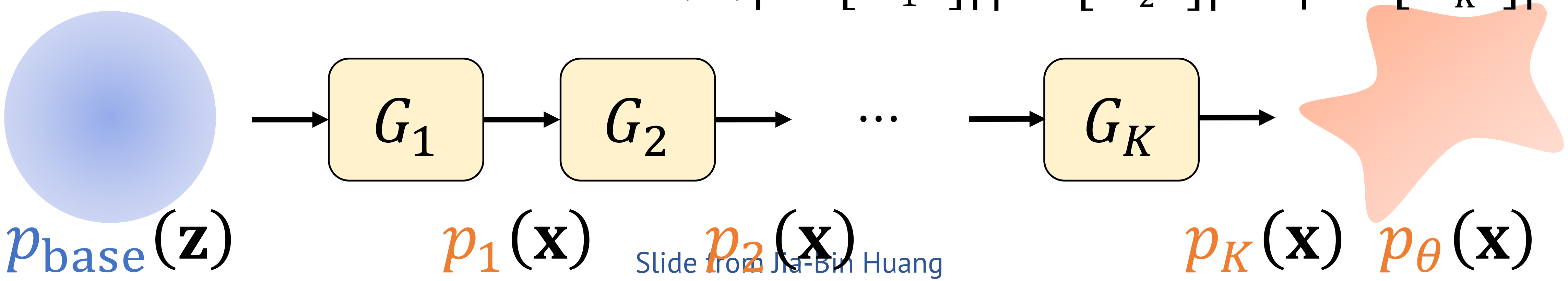
$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} \log \, p_\theta(\mathbf{x}^i)$$

$$p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}') \left|\det[\boldsymbol{J}_{G^{-1}}]\right|$$

$$\log \, p_\theta(\mathbf{x}^i) = \log \, p_{\text{base}}(\mathbf{z}^i) + \sum_{k=1}^{K} \log \left|\det\left[\boldsymbol{J}_{G_k^{-1}}\right]\right|$$

$$\mathbf{z}^i = G_1^{-1}\left(\cdots\left(G_K^{-1}(\mathbf{x}^i)\right)\right)$$

$$p_K(\mathbf{x}^i) = p_{\text{base}}(\mathbf{z}^i)\left|\det\left[\boldsymbol{J}_{G_1^{-1}}\right]\right|\left|\det\left[\boldsymbol{J}_{G_2^{-1}}\right]\right| \cdots \left|\det\left[\boldsymbol{J}_{G_K^{-1}}\right]\right|$$



Slide from Jia-Bin Huang

Maximum Likelihood

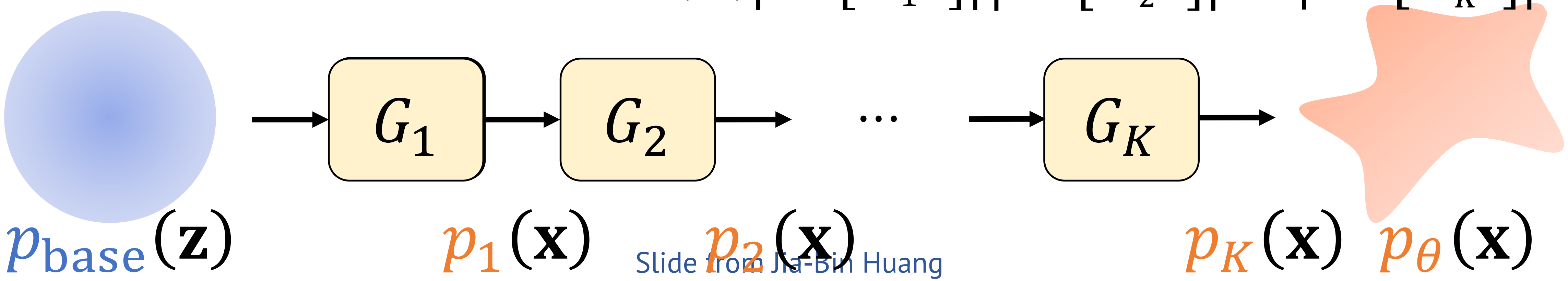
$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} \log\ p_\theta(\mathbf{x}^i)$$

$$p(\mathbf{x}') = p_{\text{base}}(\mathbf{z}')\,\left|\det[\,\boldsymbol{J}_{\boldsymbol{G}^{-1}}]\right|$$

$$\log\ p_\theta(\mathbf{x}^i) = \log\ p_{\text{base}}(\mathbf{z}^i) + \sum_{k=1}^{K} \log \left|\det\left[\boldsymbol{J}_{G_k^{-1}}\right]\right|$$

$$\mathbf{z}^i = G_1^{-1}\left(\cdots\left(G_K^{-1}(\mathbf{x}^i)\right)\right)$$

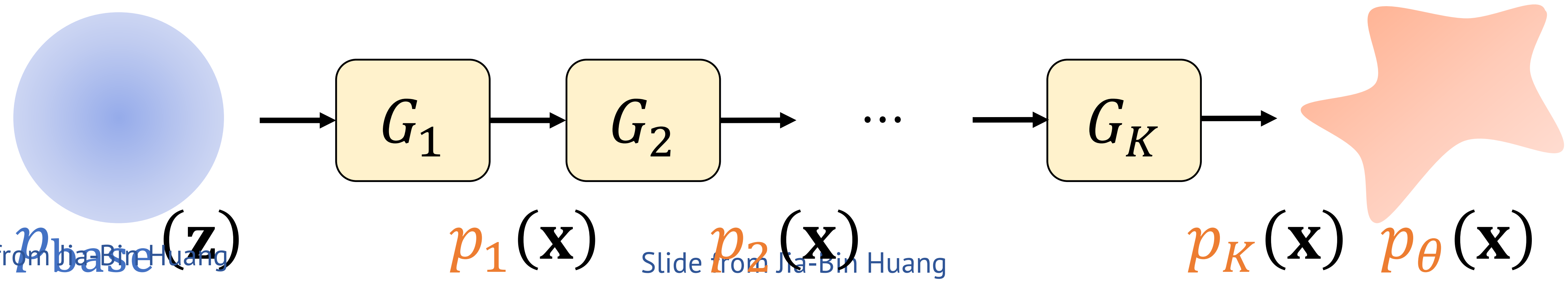
$$p_K(\mathbf{x}^i) = p_{\text{base}}(\mathbf{z}^i)\left|\det\left[\boldsymbol{J}_{G_1^{-1}}\right]\right|\left|\det\left[\boldsymbol{J}_{G_2^{-1}}\right]\right| \cdots \left|\det\left[\boldsymbol{J}_{G_K^{-1}}\right]\right|$$



Slide from Jia-Bin Huang

Residual flows

$$\mathbf{x}_{\mathrm{k+1}} = \mathbf{x}_{\mathrm{k}} + \delta\, u(\mathbf{x}_{\mathrm{k}})$$

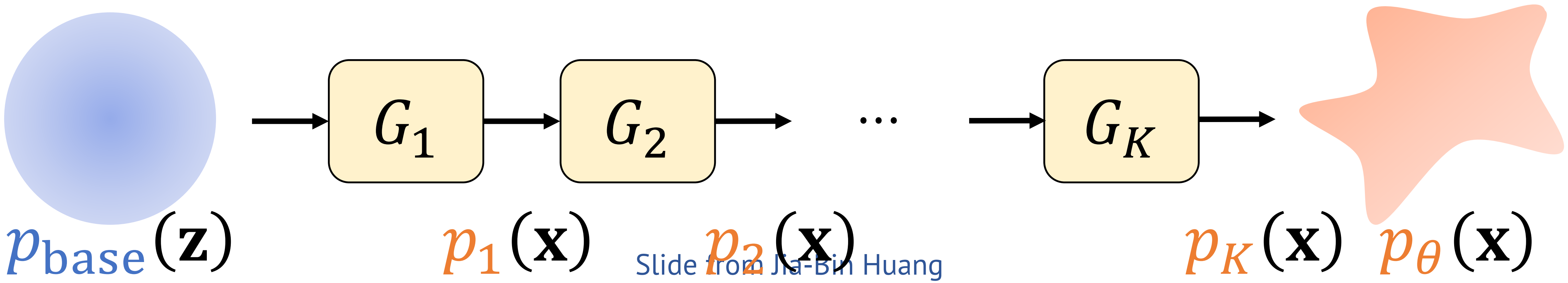


Slide from Jia-Bin Huang

Slide from Jia-Bin Huang

Residual flows

$$\mathbf{x}_{\mathrm{k}+1} - \mathbf{x}_{\mathrm{k}} = \delta\, u(\mathbf{x}_{\mathrm{k}})$$



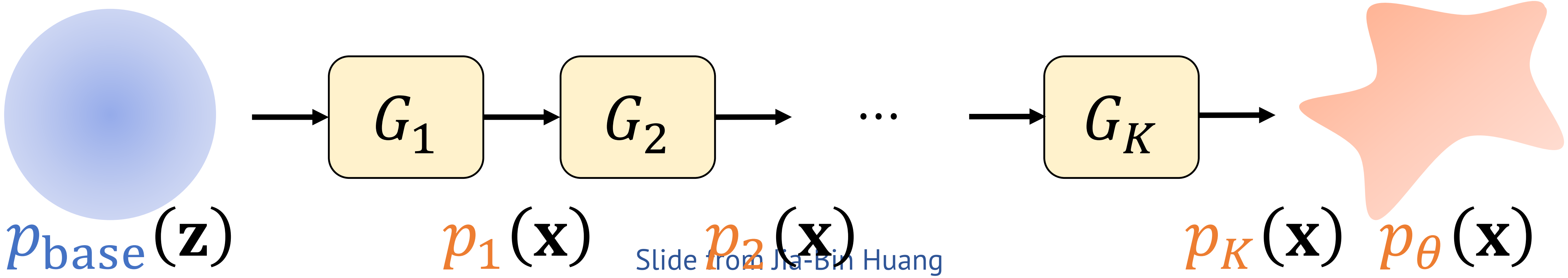
Slide from Jia-Bin Huang

Residual flows

$$\frac{\mathbf{x}_{\mathrm{k+1}} - \mathbf{x}_{\mathrm{k}}}{\delta} = u(\mathbf{x}_{\mathrm{k}}) \qquad \delta = 1/K \qquad K \to \infty$$

$$\underbrace{\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t}}_{\text{Changes in position}} = \underbrace{u_t(\ \mathbf{x}_t\ )}_{\text{Vector field}}$$

Ordinary Differential Equation



Slide from Jia-Bin Huang

Residual flows

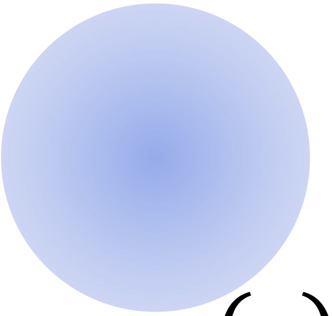
$$\frac{\mathbf{x}_{k+1} - \mathbf{x}_{k}}{\delta} = u(\mathbf{x}_{k}) \qquad \delta = 1/K \qquad K \to \infty$$

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = u_t(\mathbf{x}_t, \theta)$$

Changes in position   Vector field

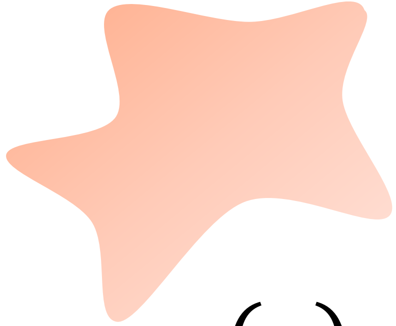
Neural Ordinary Differential Equation

$t = 0$



$p_{\text{base}}(\mathbf{z})$

$t = 1$



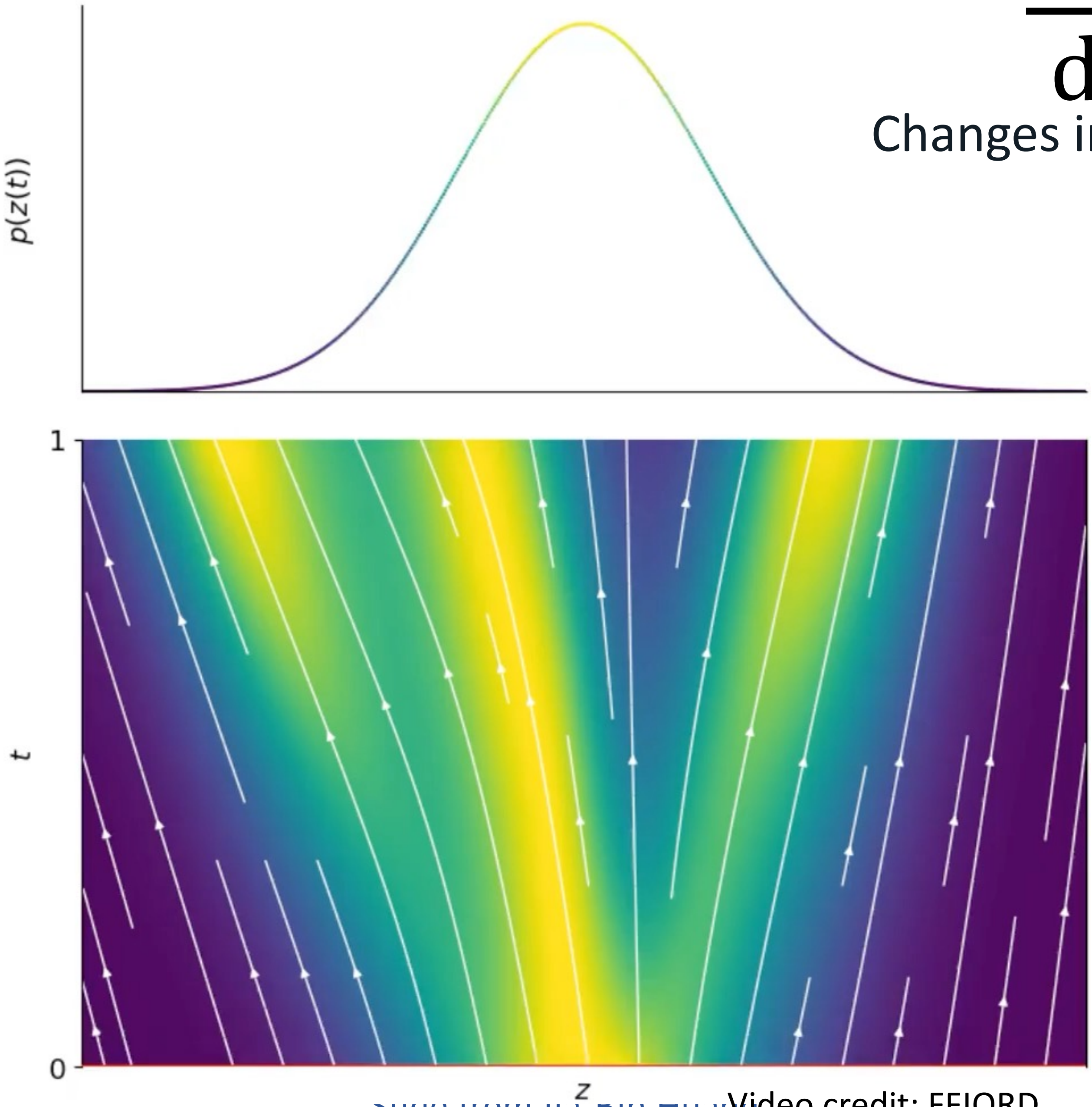
$p_\theta(\mathbf{x})$

Slide from Jia-Bin Huang

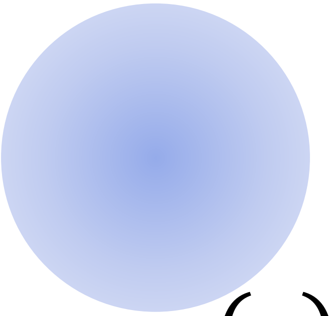
Neural Ordinary Differential Equation

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = u_t(\mathbf{x}_t, \theta)$$

Changes in position   Vector field

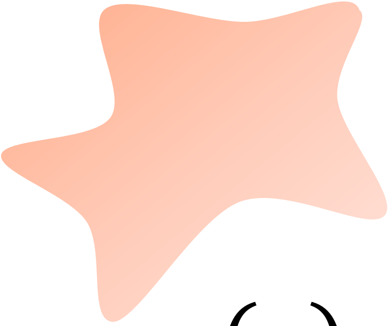


$t = 0$



$p_{\text{base}}(\mathbf{z})$

$t = 1$



$p_\theta(\mathbf{x})$

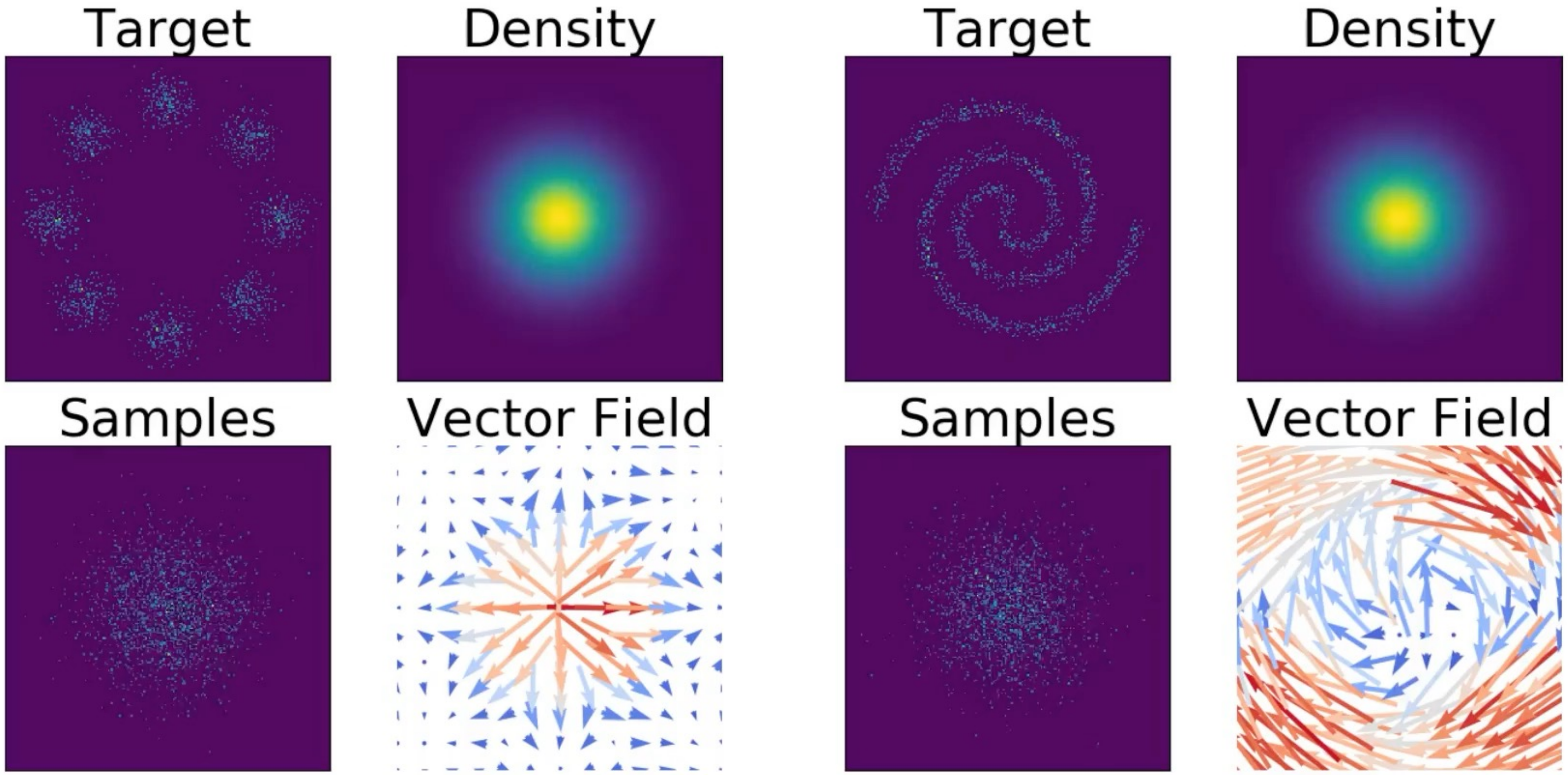
Slide from Jia-Bin Huang

Video credit: FFJORD

Neural Ordinary Differential Equation

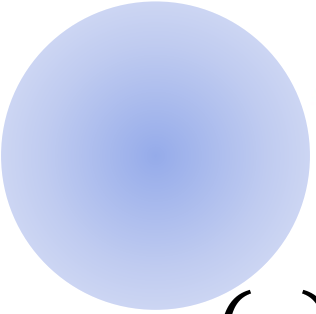
$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = u_t(\mathbf{x}_t, \theta)$$

Changes in position   Vector field



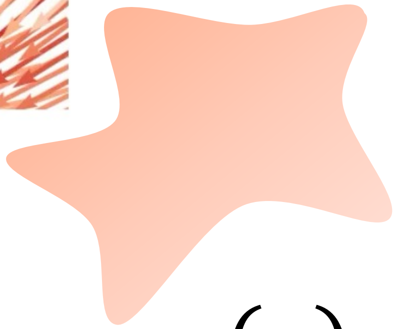
Video credit: FFJORD

$t = 0$



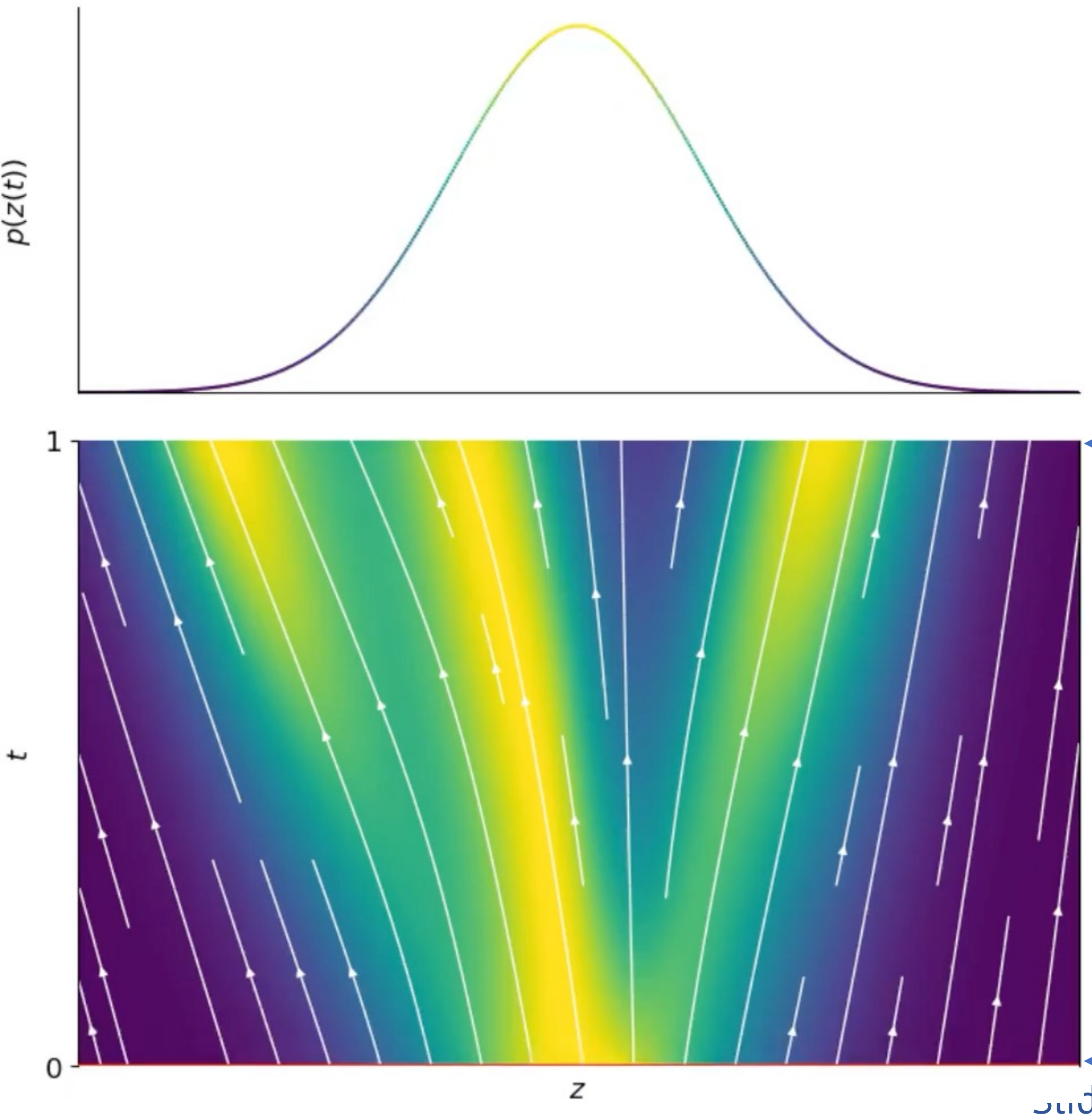
$p_{\text{base}}(\mathbf{z})$

$t = 1$



$p_\theta(\mathbf{x})$

Continuity equation / Transport equation



p_θ

Target distribution

p_{base}

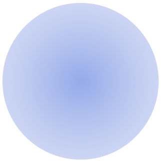
Match the “flow” $u_t(\mathbf{x}_t)$!

Slide from Jia-Bin Huang

Match the “flow” $u_t(\mathbf{x}_\mathrm{t})$!

$$\mathrm{L}_{\mathrm{FM}} = \mathrm{E}_{t,\, p_t(\mathbf{x}_t)}\left[|| v_t(\mathbf{x}_\mathrm{t}, \theta) - u_t(\mathbf{x}_\mathrm{t}) ||_2^2 \right]$$

Probability path Neural network Vector field

$p_0 = p_{\text{base}}(\mathbf{z})$ 

$p_1 = p_\theta(\mathbf{x})$

We don’t know

Probability path $p_t(\mathbf{x}_t)$

Vector field $u_t(\mathbf{x}_\mathrm{t})$

Slide from Jia-Bin Huang

Flow matching (FM) objective

$$\mathrm{L}_{\mathrm{FM}} = \mathrm{E}_{t,\, p_t(\mathbf{x}_t)}\left[|| v_t(\mathbf{x}_\mathrm{t}, \theta) - u_t(\mathbf{x}_\mathrm{t}) ||_2^2 \right]$$

Probability path Neural network Vector field

Probability path $p_t(\mathbf{x}_t) = \int p_t(\mathbf{x}_t | \mathbf{z})\; q(\mathbf{z})\, d\mathbf{z}$

Conditional probability path

Slide from Jia-Bin Huang

[Lipman et al. 2023]

Flow matching (FM) objective

$$L_{FM} = E_{t,\, p_t(\mathbf{x}_t)}\left[\| v_t(\mathbf{x}_t, \theta) - u_t(\mathbf{x}_t) \|_2^2 \right]$$

Probability path   Neural network   Vector field

Probability path $p_t(\mathbf{x}_t) = \int p_t(\mathbf{x}_t|\mathbf{z})\; q(\mathbf{z})\, d\mathbf{z}$

Conditional probability path

Example: Let $\mathbf{z} = \mathbf{x}_1$

$$p_t(\mathbf{x}|\mathbf{z}) = N(x|\, t\mathbf{x}_1, (t\sigma - t + 1)^2)$$

[Lipman et al. 2023]

Flow matching (FM) objective

$$L_{FM} = E_{t,\, p_t(\mathbf{x}_t)}\left[\, ||v_t(\mathbf{x}_t, \theta) - u_t(\mathbf{x}_t)||_2^2 \,\right]$$

Probability path   Neural network   Vector field

Probability path $p_t(\mathbf{x}_t) = \int p_t(\mathbf{x}_t|\mathbf{z})\; q(\mathbf{z})\, dz$

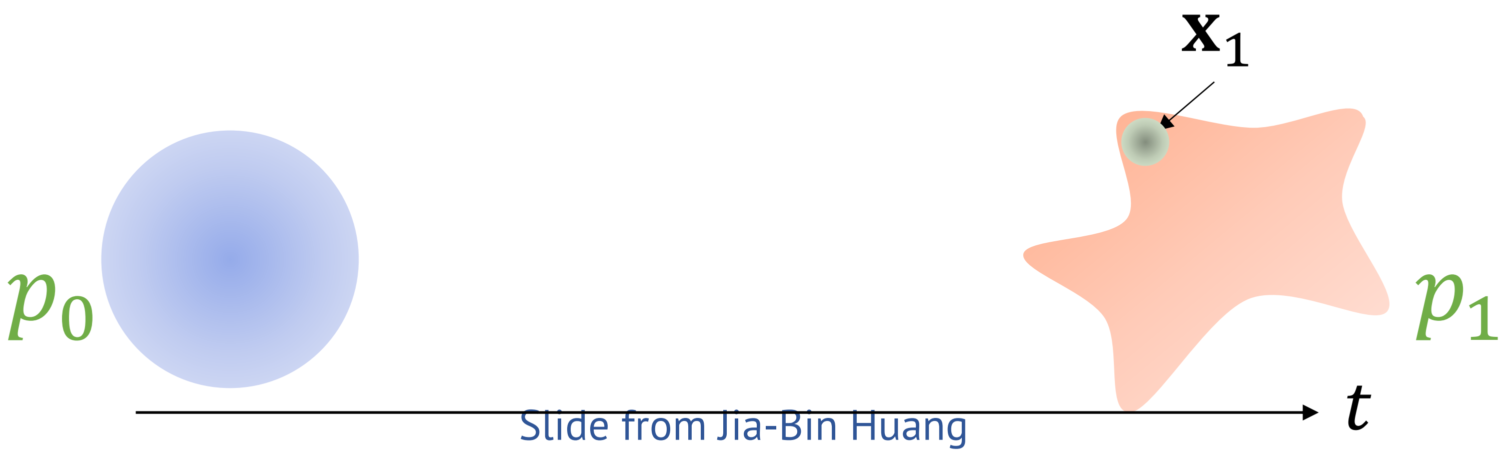
Conditional probability path

Example 1: Let $\mathbf{z} = \mathbf{x}_1$

$$p_t(\mathbf{x}|\mathbf{z}) = N(x|\, t\mathbf{x}_1, (t\sigma - t + 1)^2)$$

"just move toward $\mathbf{x}_1$"

$$u_t(\mathbf{x}|\mathbf{z}) = \frac{\mathbf{x}_1 - (1-\sigma)\mathbf{x}}{1 - (1-\sigma)t}$$



Slide from Jia-Bin Huang

[Lipman et al. 2023]

Flow matching (FM) objective

$$L_{FM} = E_{t,\, p_t(\mathbf{x}_t)}[||v_t(\mathbf{x}_\mathrm{t}, \theta) - u_t(\mathbf{x}_\mathrm{t})||_2^2\,]$$

Probability path Neural network Vector field

Conditional Flow matching (CFM) objective

$$L_{CFM} = E_{t,\, \boxed{q(\mathbf{z})},\, p_t(\mathbf{x}_t|\mathbf{z})}[||v_t(\mathbf{x}_\mathrm{t}, \theta) - u_t(\mathbf{x}_\mathrm{t}|\mathbf{z})||_2^2\,]$$

Slide from Jia-Bin Huang

[Lipman et al. 2023]

Flow matching (FM) objective

$$L_{FM} = E_{t,\, p_t(\mathbf{x}_t)}[||v_t(\mathbf{x}_t, \theta) - u_t(\mathbf{x}_t)||_2^2\,]$$

Probability path Neural network Vector field

Conditional Flow matching (CFM) objective

$$L_{CFM} = E_{t,\, q(\mathbf{z}),\, \boxed{p_t(\mathbf{x}_t|\mathbf{z})}}[||v_t(\mathbf{x}_t, \theta) - u_t(\mathbf{x}_t|\mathbf{z})||_2^2\,]$$

Slide from Jia-Bin Huang

[Lipman et al. 2023]

Flow matching (FM) objective

$$\mathrm{L}_{\mathrm{FM}} = \mathrm{E}_{t,\, p_t(\mathbf{x}_t)}\left[\| v_t(\mathbf{x}_\mathrm{t}, \theta) - u_t(\mathbf{x}_\mathrm{t}) \|_2^2 \right]$$

Probability path Neural network Vector field

Conditional Flow matching (CFM) objective

$$\mathrm{L}_{\mathrm{CFM}} = \mathrm{E}_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t | \mathbf{z})}\left[\| v_t(\mathbf{x}_\mathrm{t}, \theta) - \boxed{u_t(\mathbf{x}_\mathrm{t} | \mathbf{z})} \|_2^2 \right]$$

Slide from Jia-Bin Huang

Flow matching (FM) objective

$$\mathrm{L_{FM}} = \mathrm{E}_{t,\, p_t(\mathbf{x}_t)}\left[\| v_t(\mathbf{x}_\mathrm{t}, \theta) - u_t(\mathbf{x}_\mathrm{t}) \|_2^2 \right]$$

Probability path Neural network Vector field

Conditional Flow matching (CFM) objective

$$\mathrm{L_{CFM}} = \mathrm{E}_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t|\mathbf{z})}\left[\| v_t(\mathbf{x}_\mathrm{t}, \theta) - u_t(\mathbf{x}_\mathrm{t}|\mathbf{z}) \|_2^2 \right]$$

$$\nabla_\theta \mathrm{L_{FM}}(\theta) = \nabla_\theta \mathrm{L_{CFM}}(\theta)$$

Slide from Jia-Bin Huang

[Lipman et al. 2023]

Conditional Flow matching (CFM) objective

$$L_{\mathrm{CFM}} = \mathrm{E}_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t|\mathbf{z})}\left[|| v_t(\mathbf{x}_t, \theta) - u_t(\mathbf{x}_t|\mathbf{z}) ||_2^2 \right]$$

Example 2: Independent coupling

Conditional distribution $q(\mathbf{z}) = q(\mathbf{x_0})q(\mathbf{x_1})$

Conditional probability path $p_t(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|t\mathbf{x}_1 + (1-t)\mathbf{x}_0, \sigma^2)$

Rectified flow

Stochastic interpolant

Indep. CFM [Tong et al. 2023]

Slide from Jia-Bin Huang

Conditional Flow matching (CFM) objective

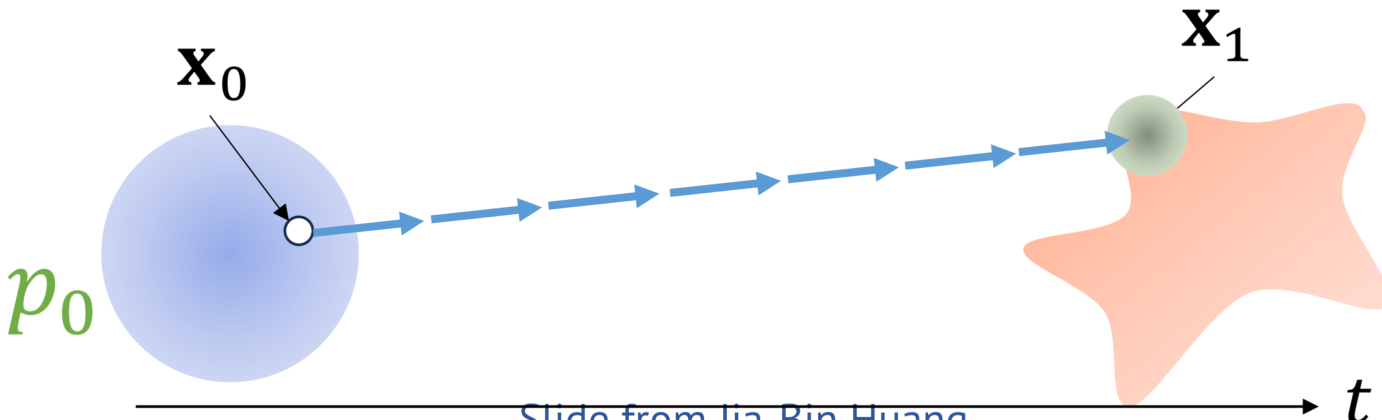
$$\mathrm{L_{CFM}} = \mathrm{E}_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t|\mathbf{z})}\left[||v_t(\mathbf{x}_t, \theta) - u_t(\mathbf{x}_t|\mathbf{z})||_2^2 \right]$$

Example 2: Independent coupling

Conditional distribution $q(\mathbf{z}) = q(\mathbf{x_0})q(\mathbf{x_1})$

Conditional probability path $p_t(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|t\mathbf{x}_1 + (1-t)\mathbf{x}_0, \sigma^2)$

Conditional vector field $u_t(\mathbf{x}|\mathbf{z}) = \mathbf{x_1} - \mathbf{x_0}$



Rectified flow

Stochastic interpolant

Indep. CFM [Tong et al. 2023]

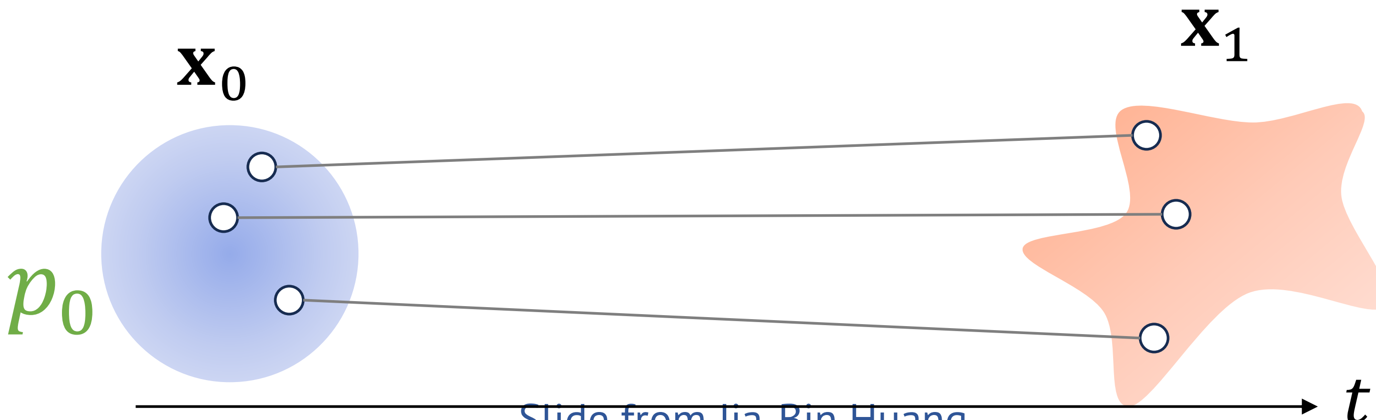
Slide from Jia-Bin Huang

Conditional Flow matching (CFM) objective

$$\mathrm{L}_{\mathrm{CFM}} = \mathrm{E}_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t|\mathbf{z})}\left[||v_t(\mathbf{x}_{\mathrm{t}}, \theta) - u_t(\mathbf{x}_{\mathrm{t}}|\mathbf{z})||_2^2 \right]$$

Example 3: Optimal transport CFM

Conditional distribution $\quad q(\mathbf{z}) = \pi(\mathbf{x}_0, \mathbf{x}_1)$



Multisample FM [Pooladian et al. 2023]
OT-CFM [Tong et al. 2023]

Slide from Jia-Bin Huang

Conditional Flow matching (CFM) objective

$$\mathrm{L}_{\mathrm{CFM}} = \mathrm{E}_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t|\mathbf{z})}\left[||v_t(\mathbf{x}_\mathrm{t}, \theta) - u_t(\mathbf{x}_\mathrm{t}|\mathbf{z})||_2^2\right]$$

Example 3: Optimal transport CFM

Conditional distribution $q(\mathbf{z}) = \pi(\mathbf{x}_0, \mathbf{x}_1)$

Conditional probability path $p_t(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|t\mathbf{x}_1 + (1-t)\mathbf{x}_0, \sigma^2)$

Conditional vector field $u_t(\mathbf{x}|\mathbf{z}) = \mathbf{x}_1 - \mathbf{x}_0$

Slide from Jia-Bin Huang

Multisample FM [Pooladian et al. 2023]
OT-CFM [Tong et al. 2023]

Conditional Flow matching (CFM) objective

$$L_{CFM} = E_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t|\mathbf{z})}\left[\| v_t(\mathbf{x}_t, \theta) - u_t(\mathbf{x}_t|\mathbf{z}) \|_2^2 \right]$$

Example 3: Optimal transport CFM

Conditional distribution $q(\mathbf{z}) = \pi(\mathbf{x}_0, \mathbf{x}_1)$

Conditional probability path $p_t(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|t\mathbf{x}_1 + (1-t)\mathbf{x}_0, \sigma^2)$

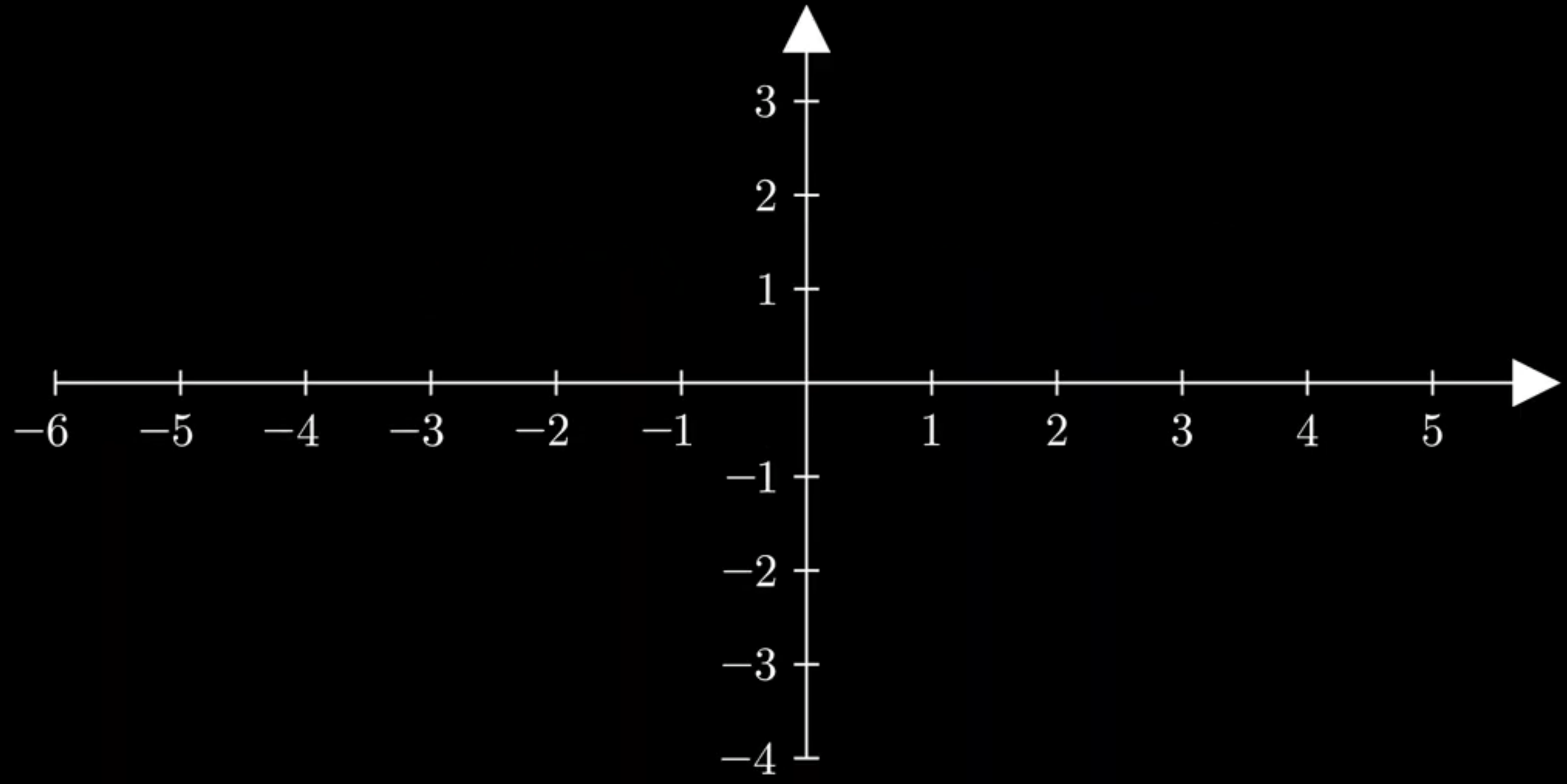
Conditional vector field $u_t(\mathbf{x}|\mathbf{z}) = \mathbf{x}_1 - \mathbf{x}_0$

Multisample FM [Pooladian et al. 2023]
OT-CFM [Tong et al. 2023]

Slide from Jia-Bin Huang

Flow Matching in Action

credits: Jia-Bin Huang; Link: https://www.youtube.com/watch?v=swKdn-qT47Q


3
2
1
−6
−5
−4
−3
−2
−1
1
2
3
4
5
−1
−2
−3
−4

Flow Matching Dynamics

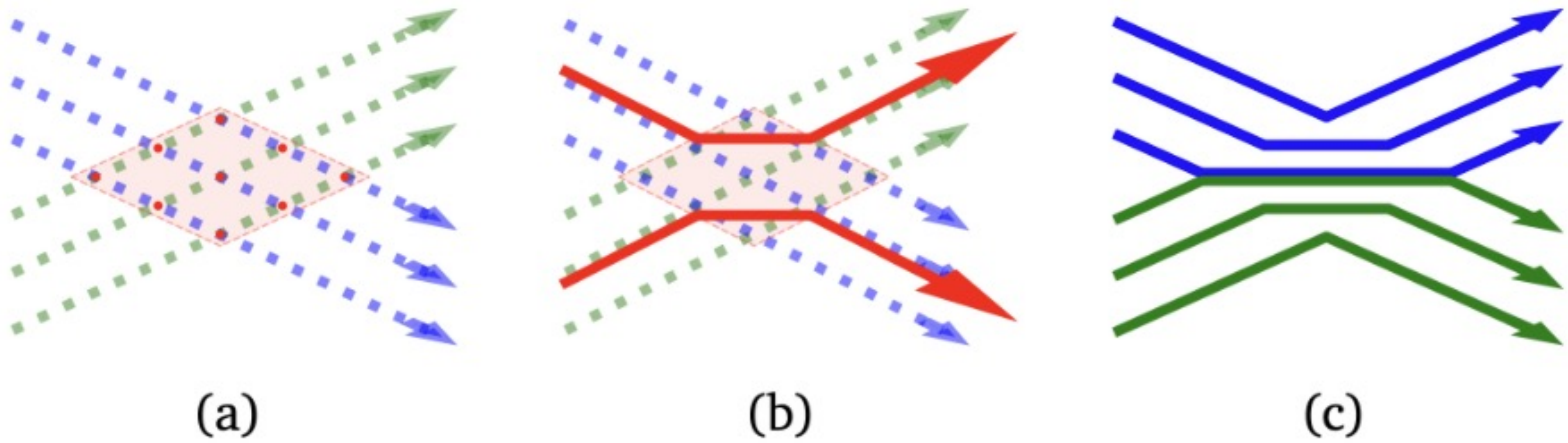
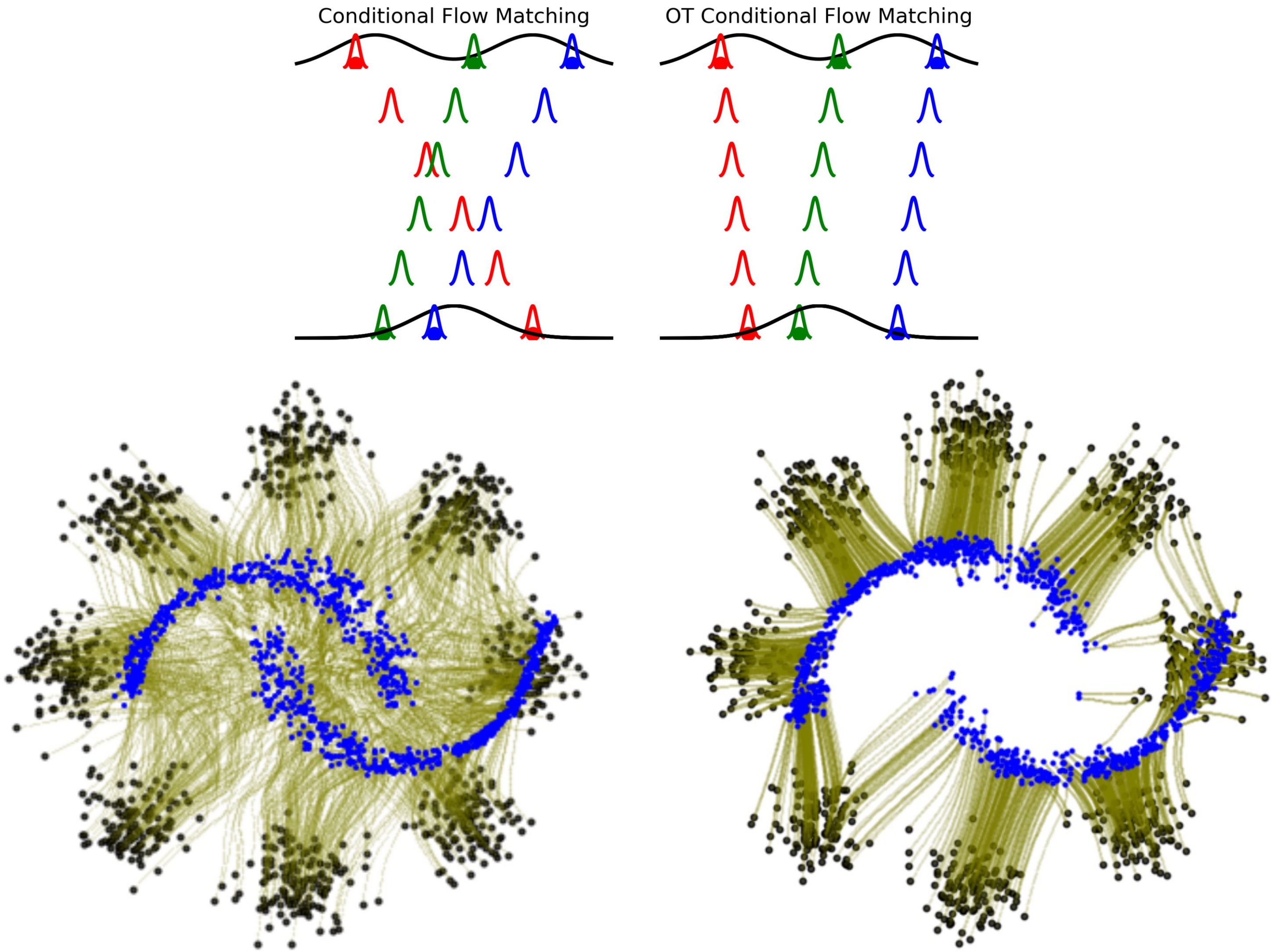


Figure 3. A close-up view of how rectification "rewires" interpolation trajectories. (a) Interpolation trajectories with intersections. (b) Averaged velocity directions at intersection points (red arrows). (c) Trajectories of the resulting rectified flow.

Source: https://rectifiedflow.github.io/blog/2024/intro/

Conditional Flow matching (CFM) objective

$$\mathrm{L}_{\mathrm{CFM}} = \mathrm{E}_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t|\mathbf{z})}\left[|| v_t(\mathbf{x}_\mathrm{t}, \theta) - u_t(\mathbf{x}_\mathrm{t}|\mathbf{z}) ||_2^2 \right]$$



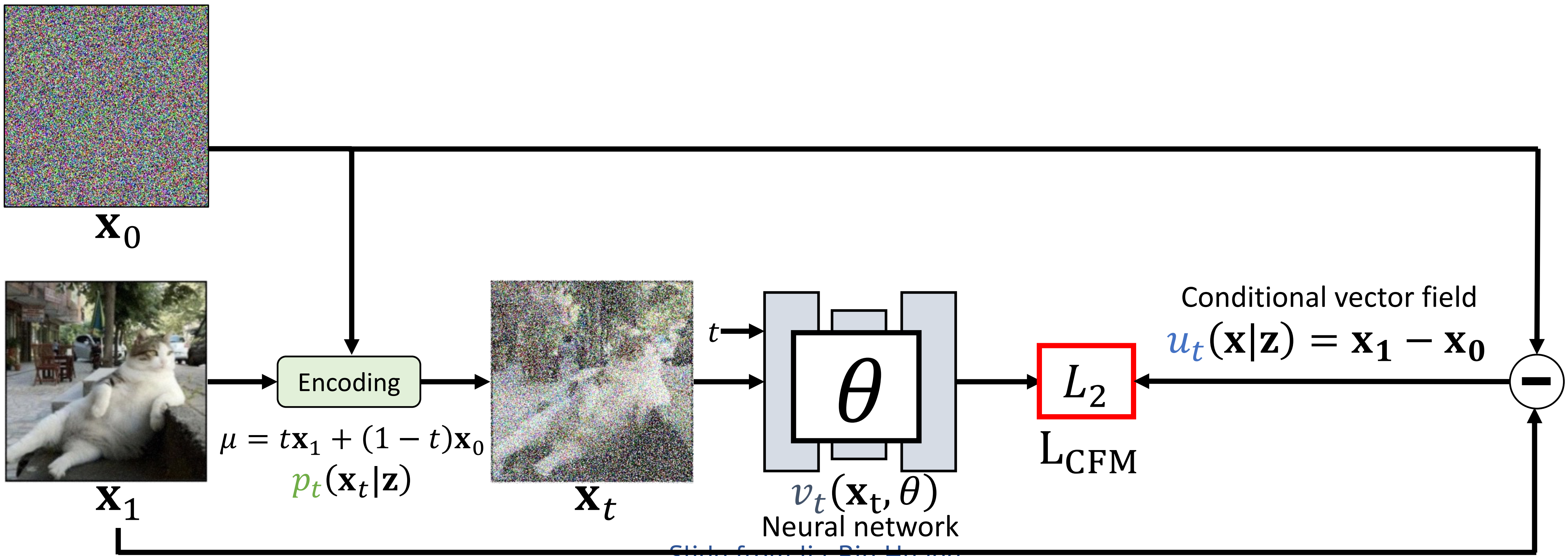
OT-CFM [Tong et al. 2023]

Conditional Flow matching (CFM) objective

$$L_{CFM} = E_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t|\mathbf{z})}\left[\| v_t(\mathbf{x}_t, \theta) - u_t(\mathbf{x}_t|\mathbf{z}) \|_2^2 \right]$$

Conditional distribution

$q(\mathbf{z}) = q(\mathbf{x}_0)q(\mathbf{x}_1)$



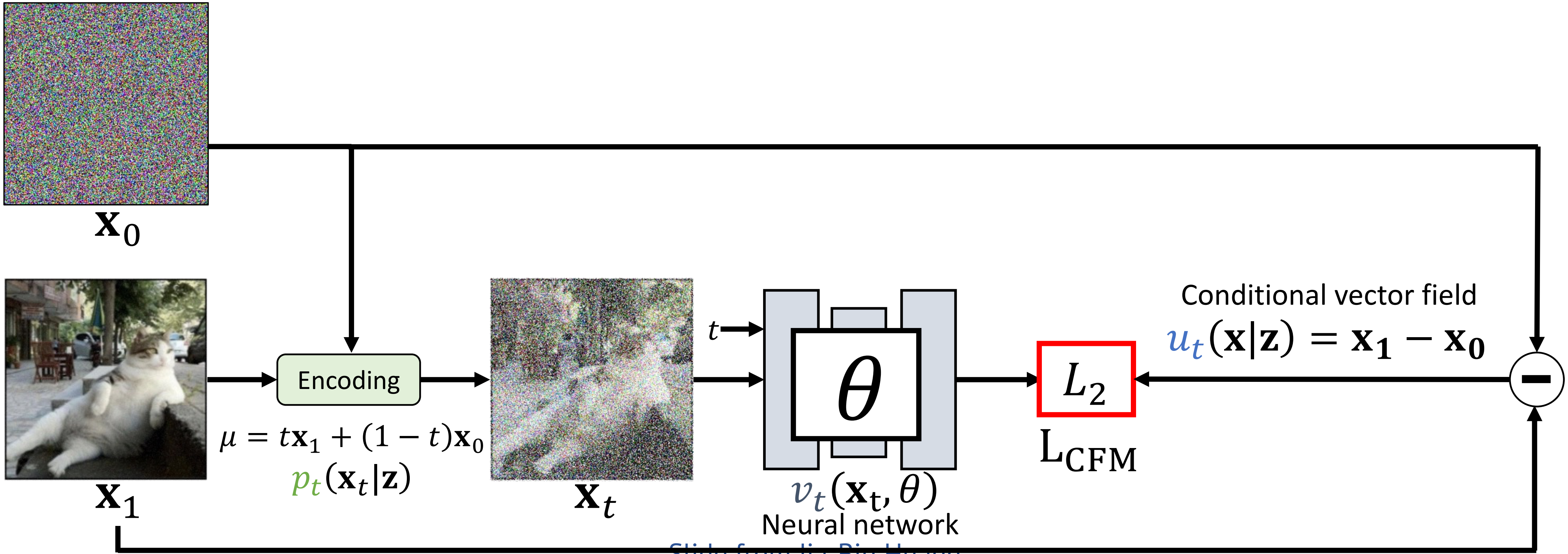
Slide from Jia-Bin Huang

Conditional Flow matching (CFM) objective

$$L_{CFM} = E_{t,\, q(\mathbf{z}),\, p_t(\mathbf{x}_t|\mathbf{z})}\left[||v_t(\mathbf{x}_t, \theta) - u_t(\mathbf{x}_t|\mathbf{z})||_2^2 \right]$$

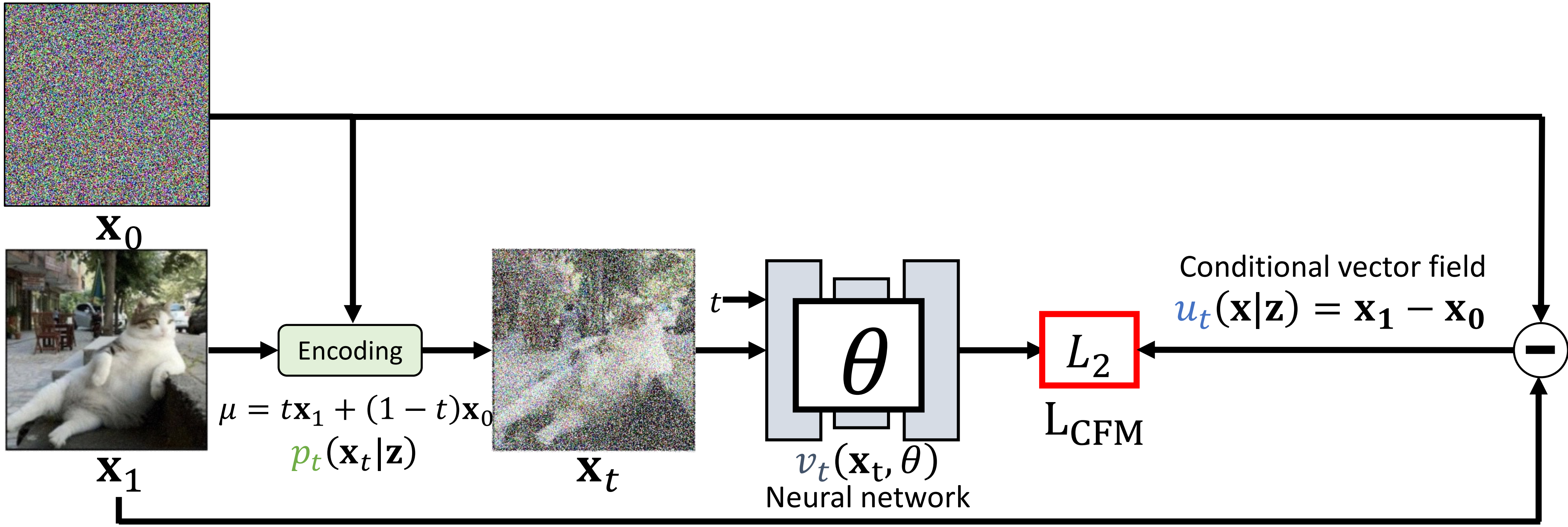
Conditional distribution

$q(\mathbf{z}) = q(\mathbf{x_0})q(\mathbf{x_1})$

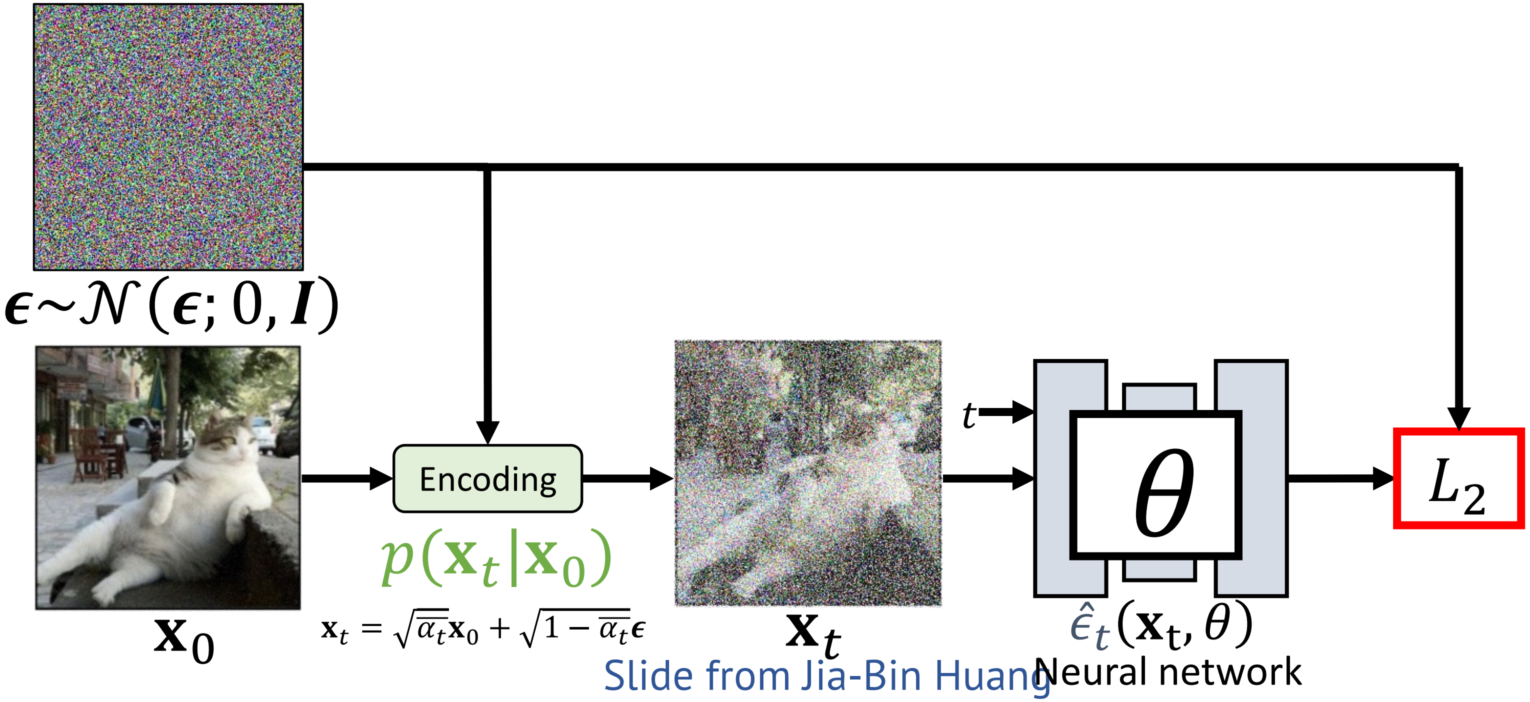


Slide from Jia-Bin Huang

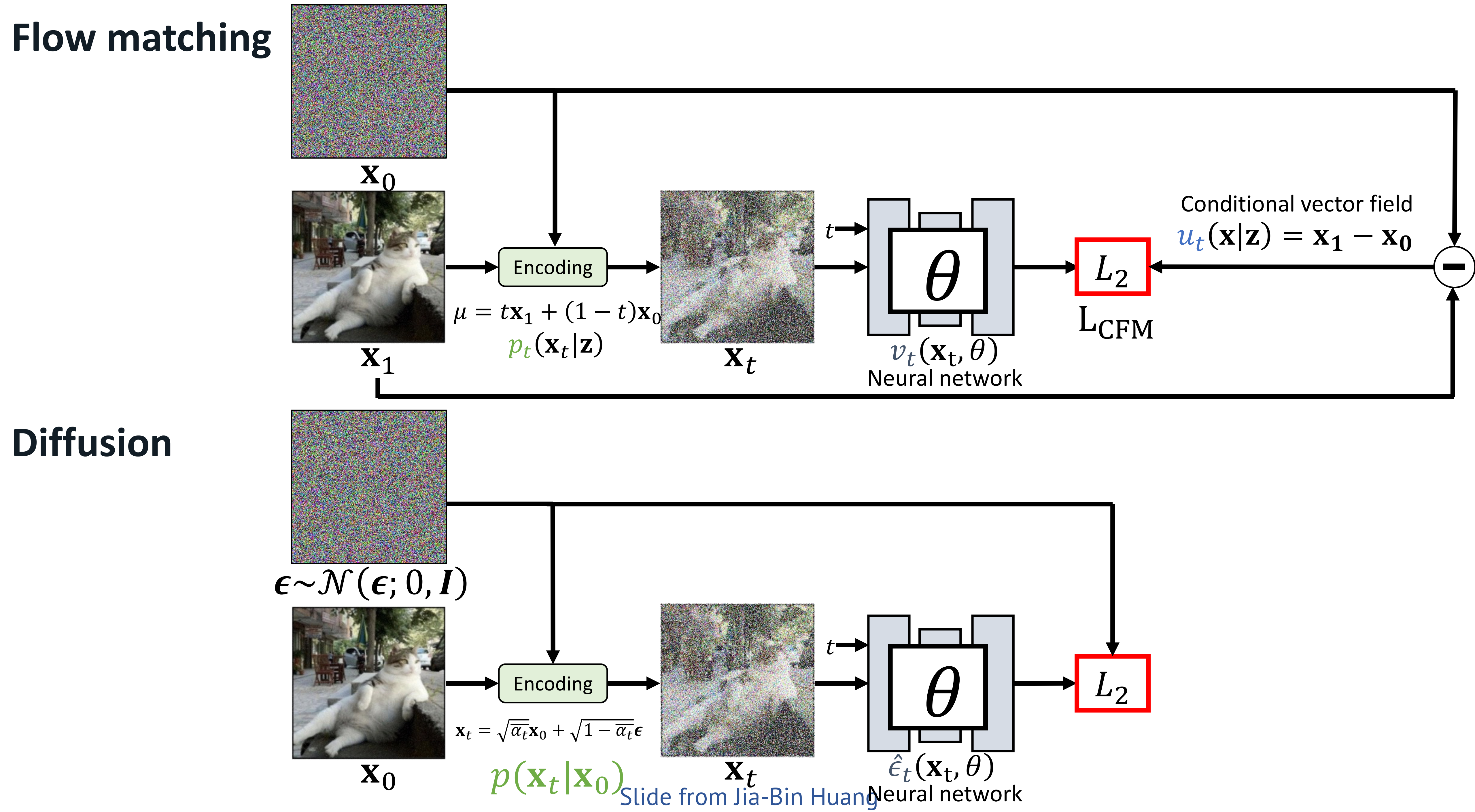
Flow matching



Diffusion



Slide from Jia-Bin Huang



Slide from Jia-Bin Huang

Connection between Diffusion and Flow matching

- Similar training objectives derived using different outputs from the network

Network Output	Formulation	MSE on Network Output
$\hat{\boldsymbol{\epsilon}}$-prediction	$\hat{\boldsymbol{\epsilon}}$	$\lVert\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\rVert_2^2$
$\hat{\mathbf{x}}$-prediction	$\hat{\mathbf{x}} = (\mathbf{z}_t - \sigma_t \hat{\boldsymbol{\epsilon}})/\alpha_t$	$\lVert\hat{\mathbf{x}} - \mathbf{x}\rVert_2^2 = e^{-\lambda}\lVert\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\rVert_2^2$
$\hat{\mathbf{v}}$-prediction	$\hat{\mathbf{v}} = \alpha_t \hat{\boldsymbol{\epsilon}} - \sigma_t \hat{\mathbf{x}}$	$\lVert\hat{\mathbf{v}} - \mathbf{v}\rVert_2^2 = \alpha_t^2 (e^{-\lambda} + 1)^2 \lVert\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\rVert_2^2$
$\hat{\mathbf{u}}$-flow matching vector field	$\hat{\mathbf{u}} = \hat{\boldsymbol{\epsilon}} - \hat{\mathbf{x}}$	$\lVert\hat{\mathbf{u}} - \mathbf{u}\rVert_2^2 = (e^{-\lambda/2} + 1)^2 \lVert\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}\rVert_2^2$

Source: https://diffusionflow.github.io

Imagine you want to use your trained denoiser model to transform random noise into a datapoint. Recall that the DDIM update is given by $\mathbf{z}_s = \alpha_s \hat{\mathbf{x}} + \sigma_s \hat{\boldsymbol{\epsilon}}$. Interestingly, by rearranging terms it can be expressed in the following formulation, with respect to several sets of network outputs and reparametrizations:

$$\tilde{\mathbf{z}}_s = \tilde{\mathbf{z}}_t + \textbf{Network output} \cdot (\eta_s - \eta_t) \tag{5}$$

Network Output	Reparametrization
$\hat{\mathbf{x}}$-prediction	$\tilde{\mathbf{z}}_t = \mathbf{z}_t / \sigma_t$ and $\eta_t = \alpha_t / \sigma_t$
$\hat{\boldsymbol{\epsilon}}$-prediction	$\tilde{\mathbf{z}}_t = \mathbf{z}_t / \alpha_t$ and $\eta_t = \sigma_t / \alpha_t$
$\hat{\mathbf{u}}$-flow matching vector field	$\tilde{\mathbf{z}}_t = \mathbf{z}_t / (\alpha_t + \sigma_t)$ and $\eta_t = \sigma_t / (\alpha_t + \sigma_t)$

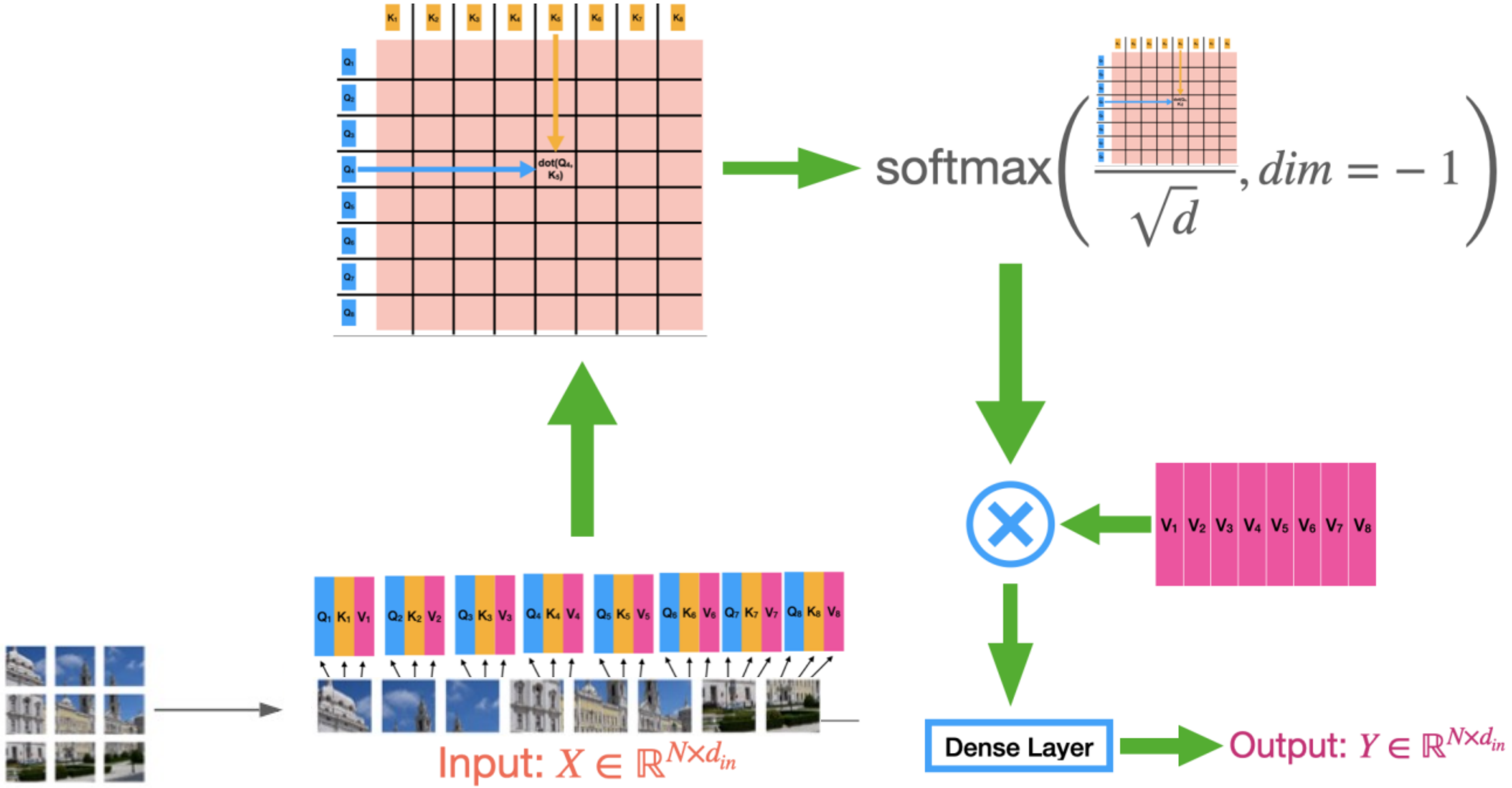
Remember the flow matching update in Equation (4)? This should look similar. If we set the network output as $\hat{\mathbf{u}}$ in the last line and let $\alpha_t = 1 - t$, $\sigma_t = t$, we have $\tilde{\mathbf{z}}_t = \mathbf{z}_t$ and $\eta_t = t$, which is the flow matching update! More formally, the flow matching update is a Euler sampler of the sampling ODE (i.e., $\mathrm{d}\mathbf{z}_t = \hat{\mathbf{u}} \mathrm{d}t$), and with the flow matching noise schedule,

> *Diffusion with DDIM sampler == Flow matching sampler (Euler).*

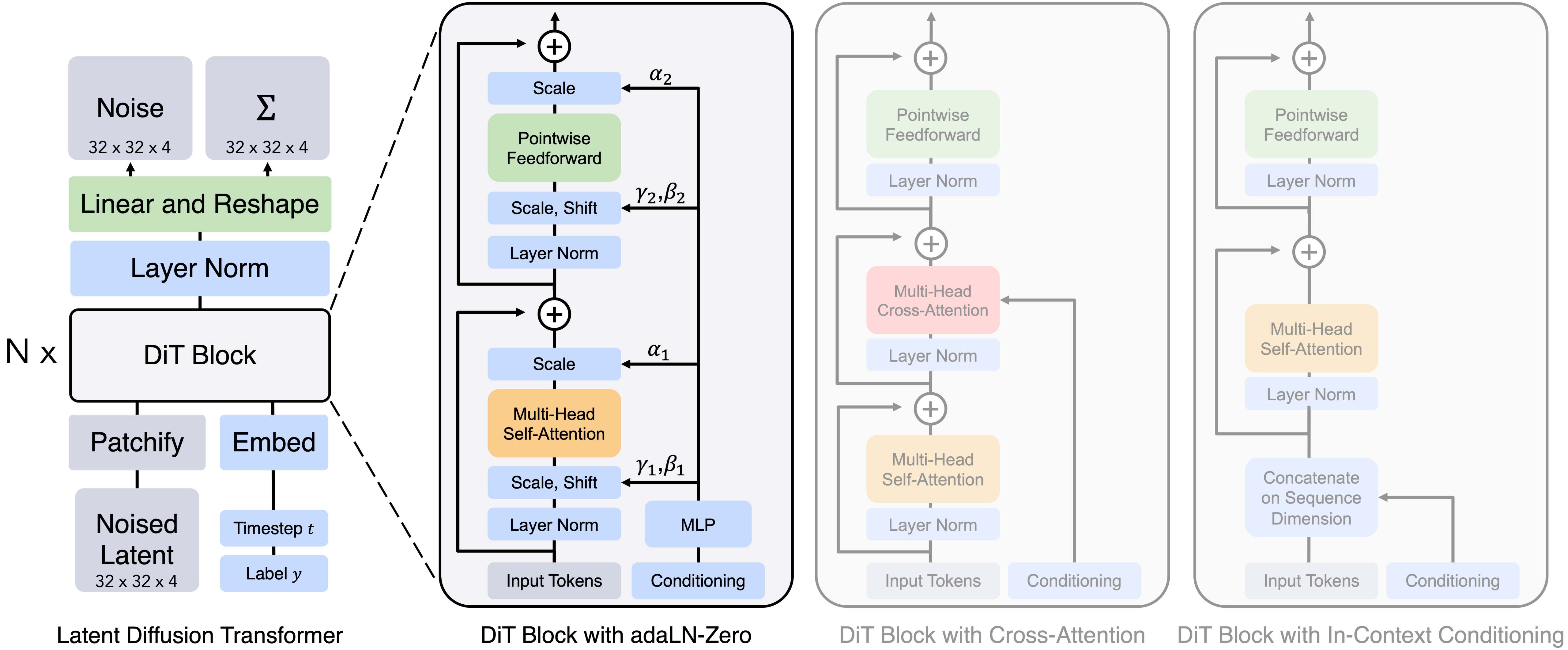
Source: https://diffusionflow.github.io

DiT Architecture

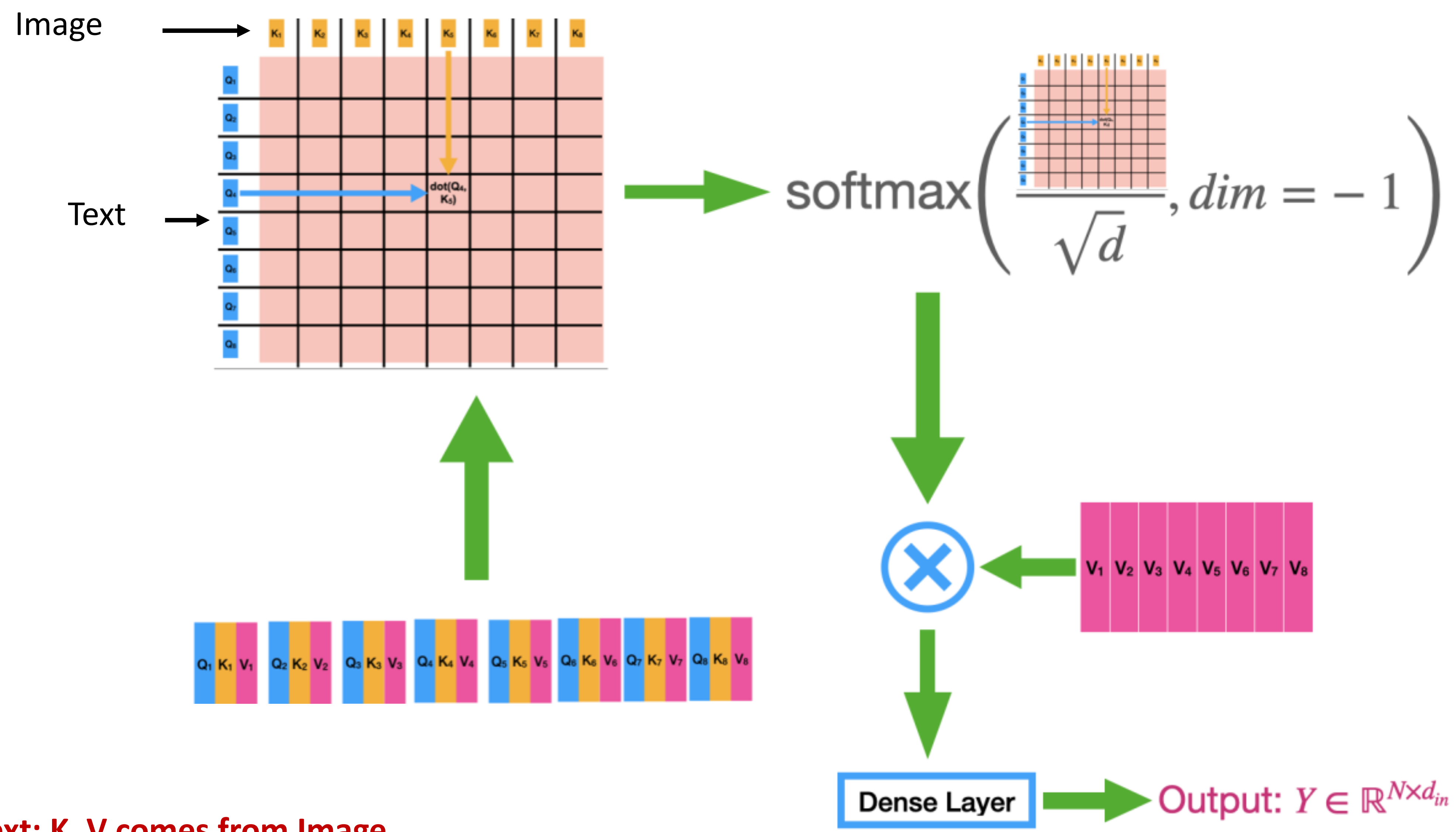
Attention Module



Diffusion Transformers Architecture (DiT)



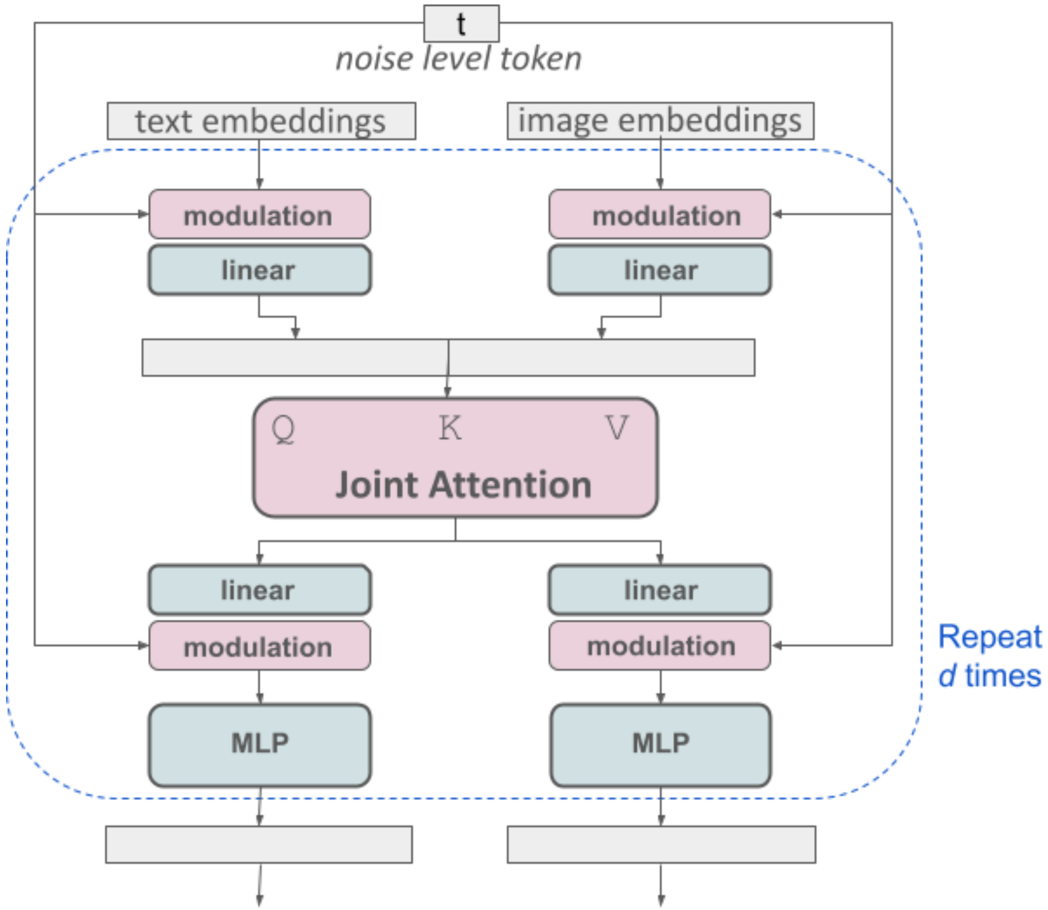
Cross Attention Module



Q comes from Text; K, V comes from Image

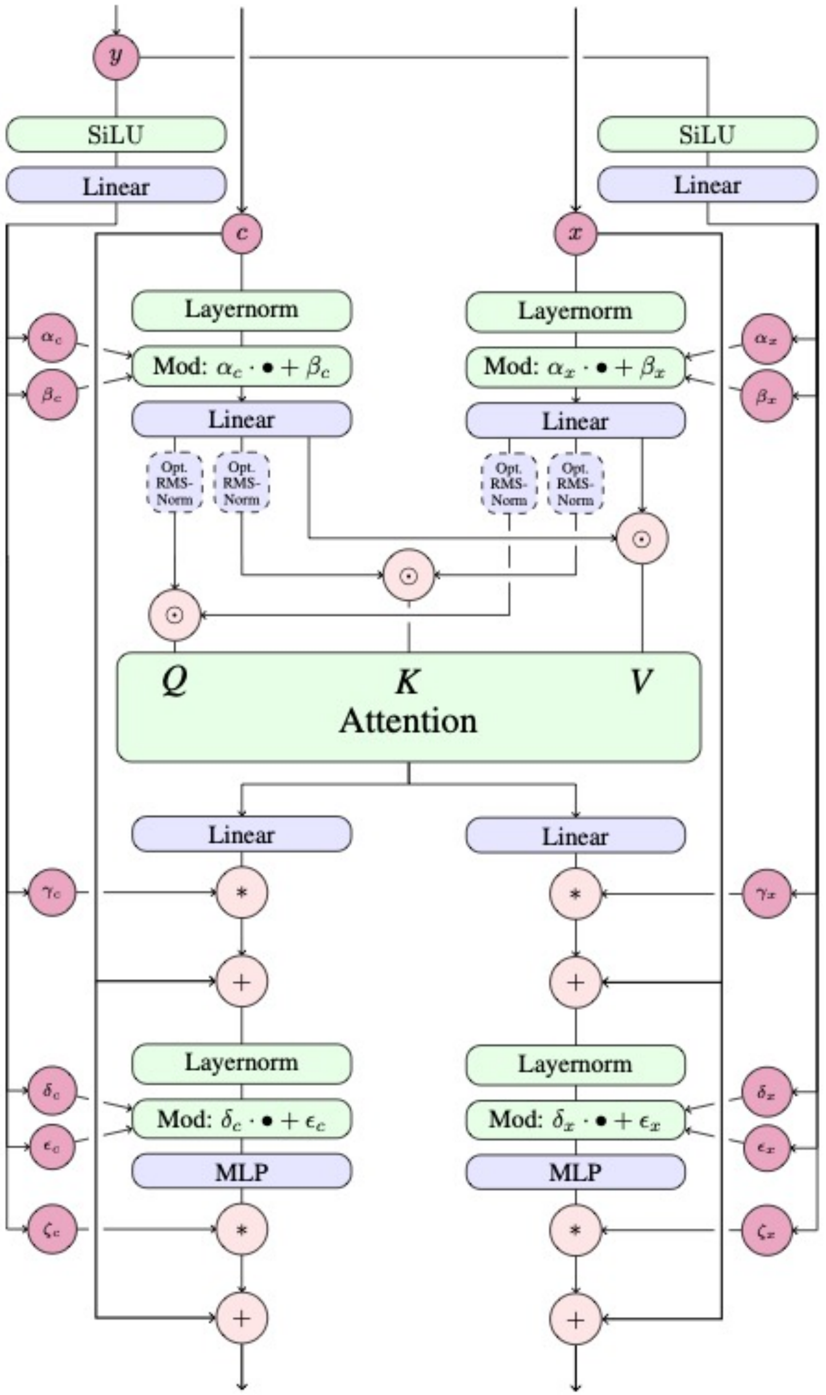
MMDiT architecture

- Separate paths for text and image tokens
- Both representations vary across the network

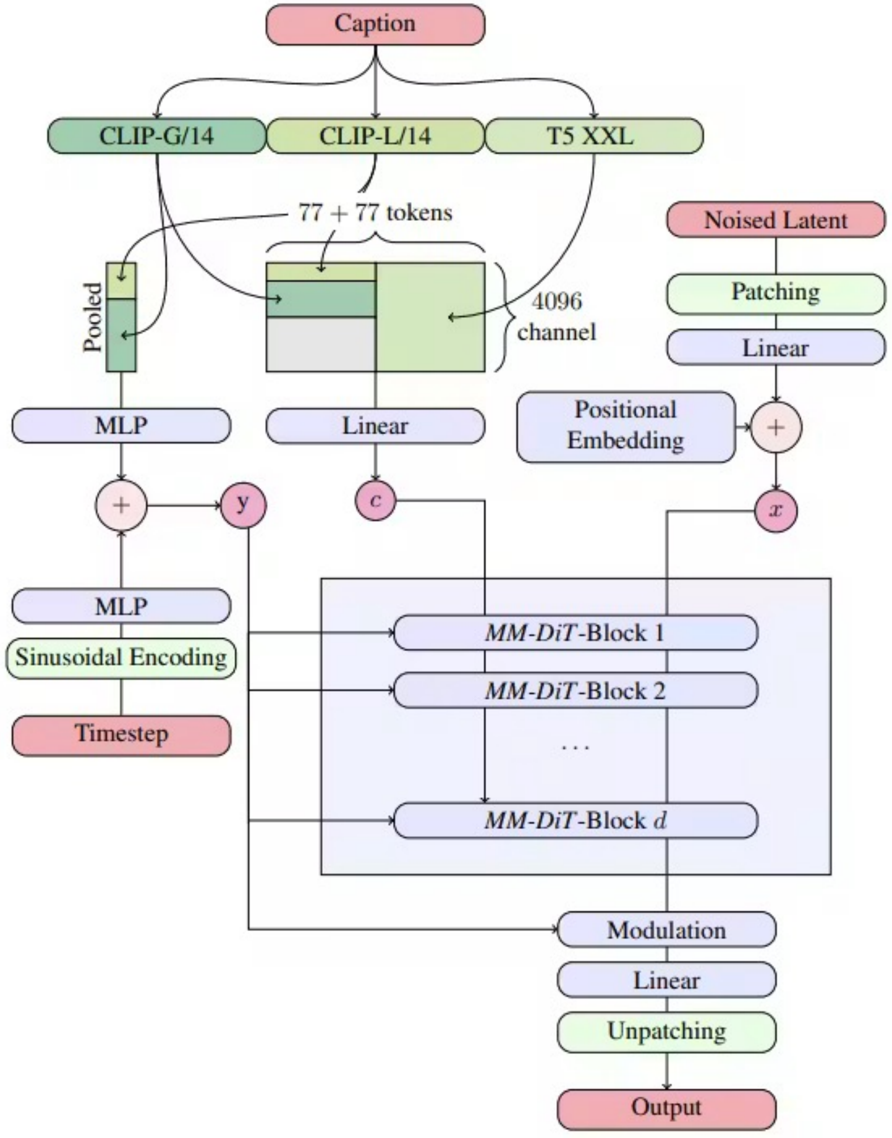


MMDiT architecture in SD 3.5

- Separate paths for text and image tokens
- Both representations vary across the network



MMDiT architecture in SD 3.5



Please fill in the Feedback Form at:

https://virajshah.com/sc395-feedback5

Thank You!