Diffusion Model

A brief introduction & Applications

Huawei Fan Cheng Wan Anqi Zeng Jiasheng Xu

Inspiration

Structure

Introduction

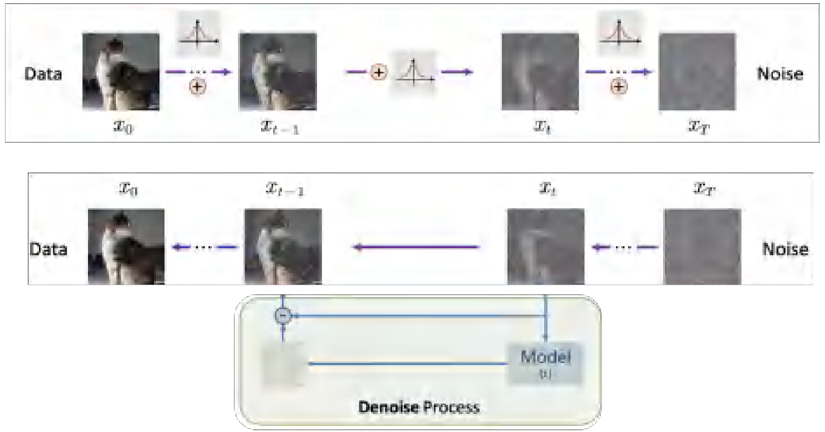
Application

What is Diffusion Model?

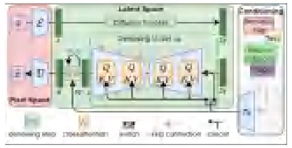
前向加噪

反向去噪

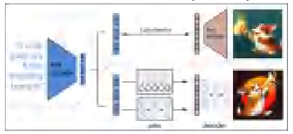
模型训练



How Diffusion Model can be applied in different fields?

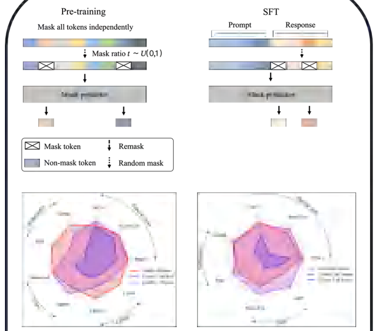


Stable Diffusio(LDM)

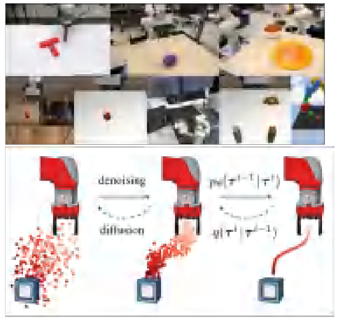


DALLE 2 By OpenAI

TXT2IMG



LLM



Robotics

Brief Introduction

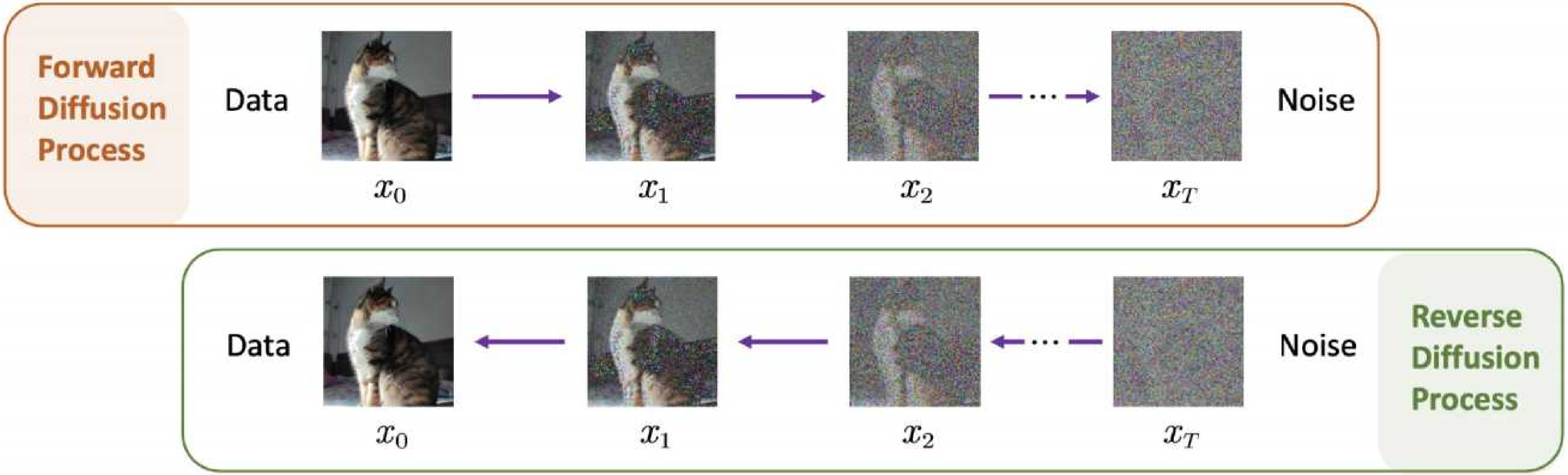
前向加噪

反向去噪

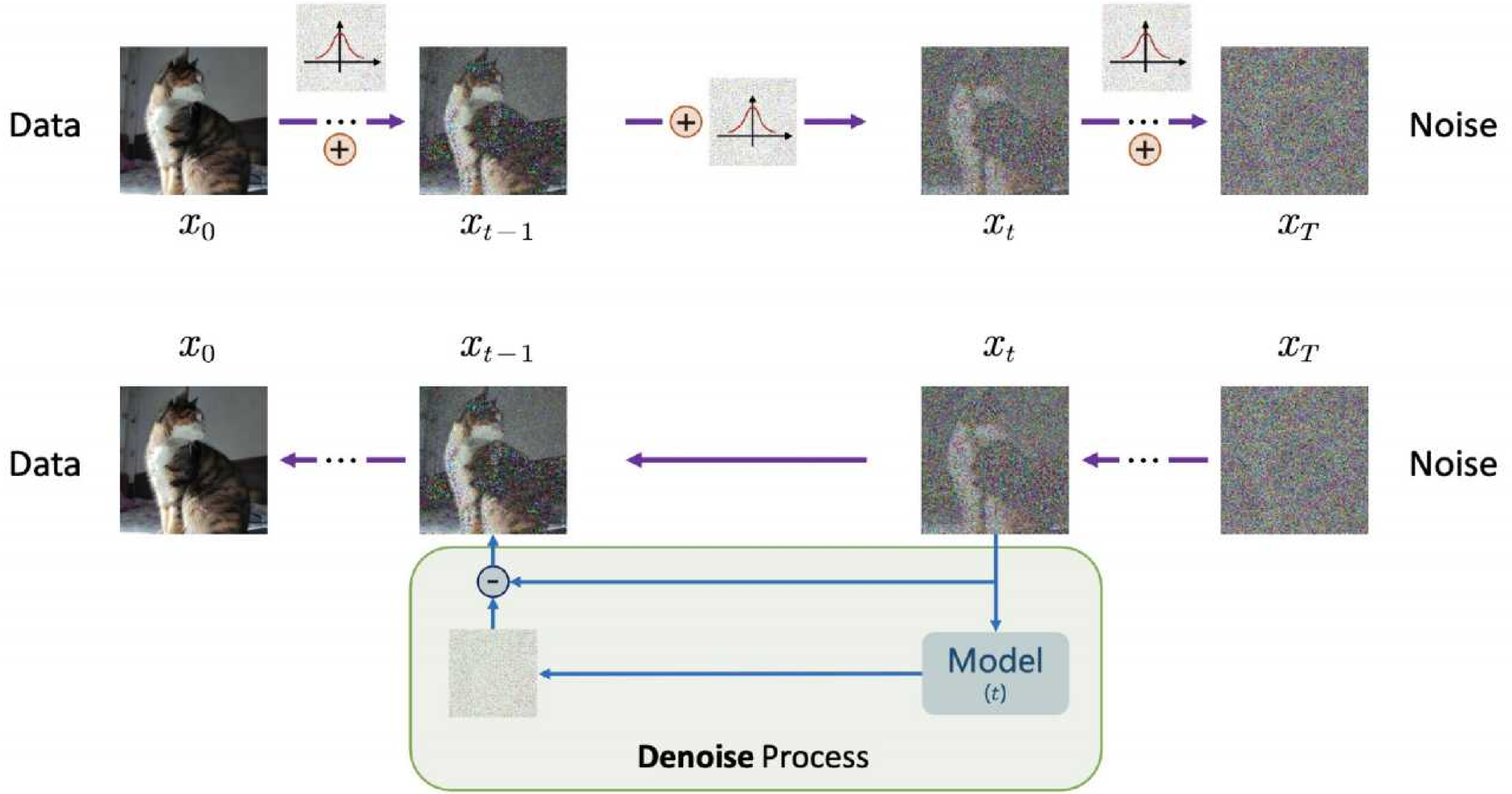
模型训练

What is Diffusion Model?

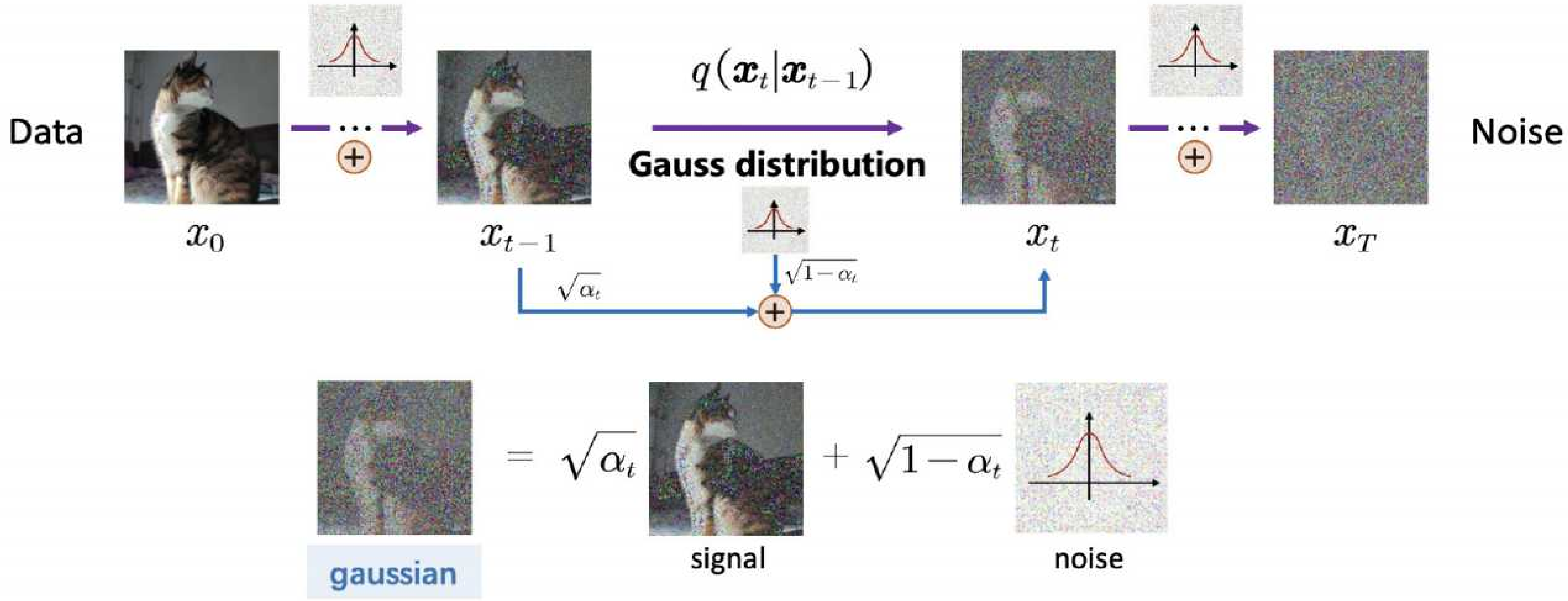
Denoising Diffusion Probabilistic Models



What is Diffusion Model?



Forward Diffusion Process



$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_{t-1}, \quad \varepsilon \sim \mathcal{N}(0, I)$$

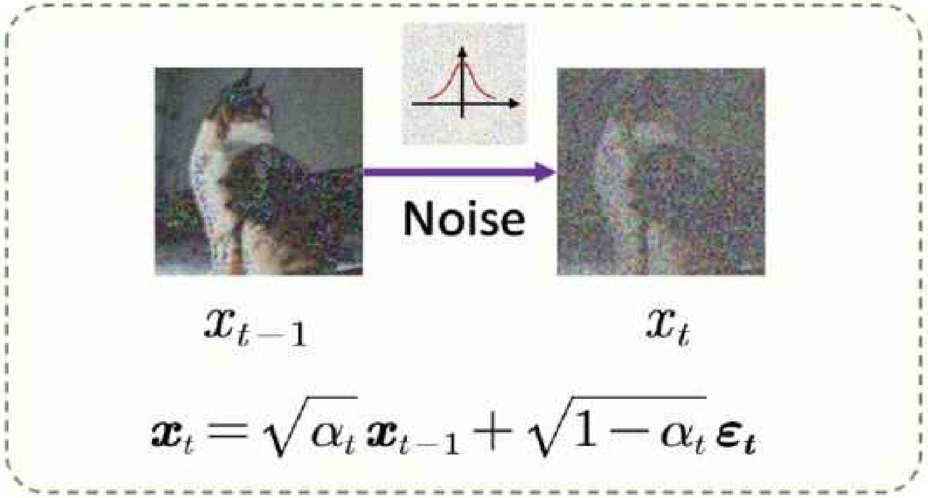
$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t) I)$$

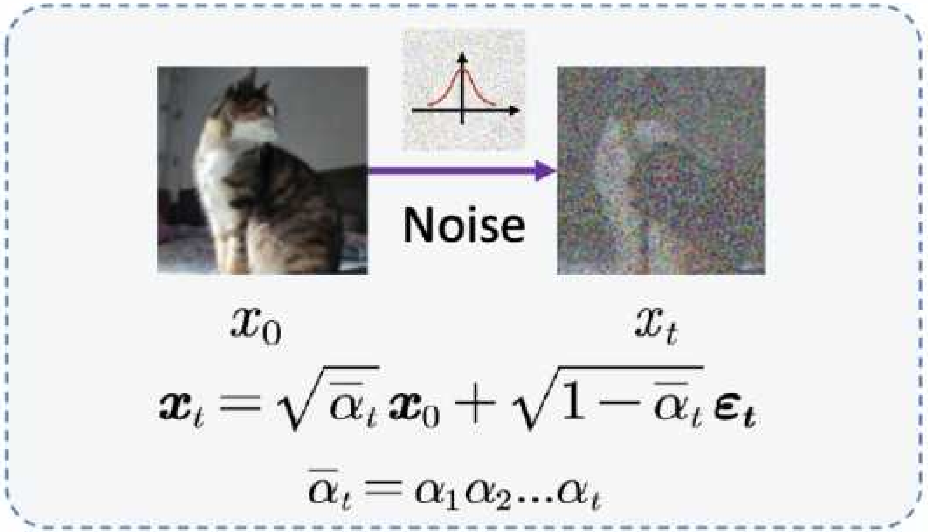
Forward Diffusion Process

①Forward (closed-form)

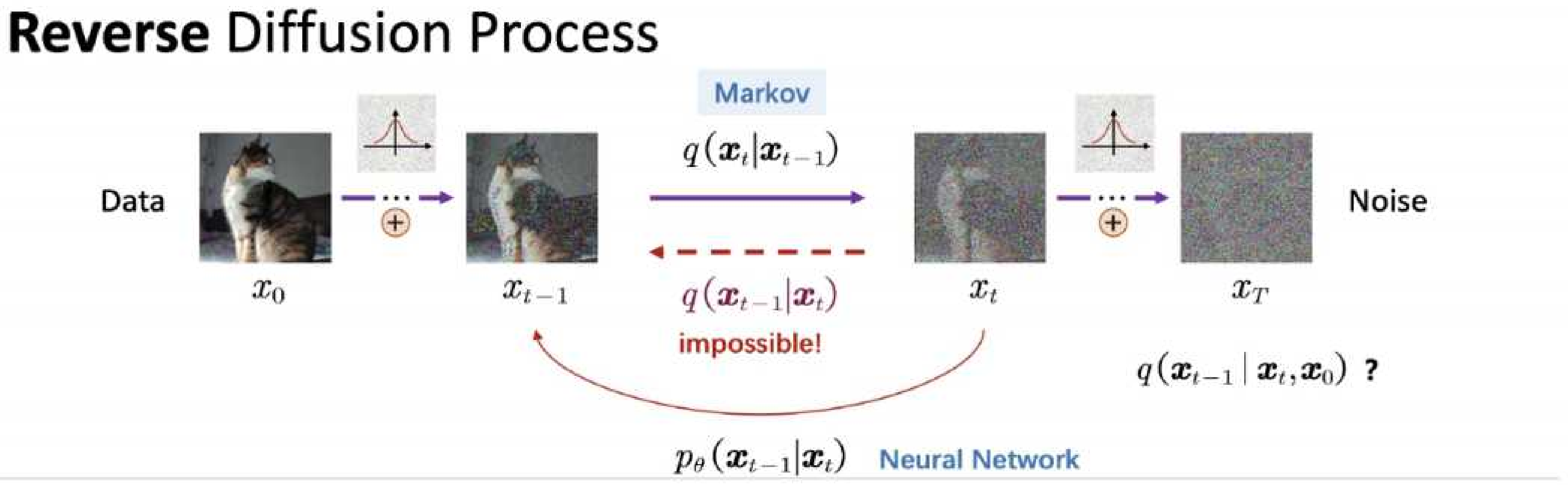
$$
\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\varepsilon_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}\varepsilon_{t-2}\right) + \sqrt{1-\alpha_t}\varepsilon_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \left(\sqrt{\alpha_t(1-\alpha_{t-1})}\varepsilon_{t-2} + \sqrt{1-\alpha_t}\varepsilon_{t-1}\right) \\
&\;\;\vdots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I)
\end{aligned}
$$

Where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$





Reverse Diffusion Process

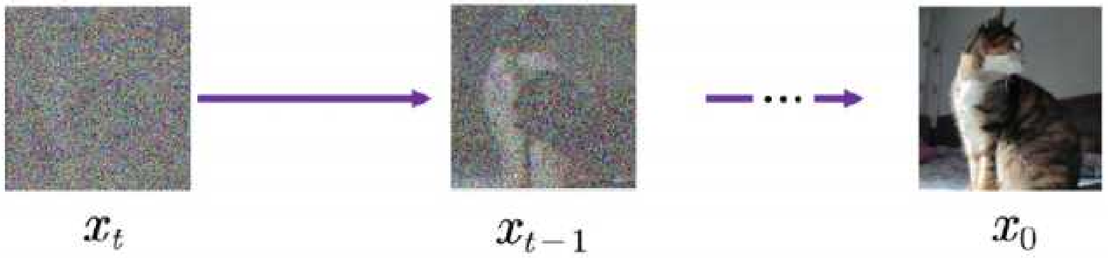


Assume: the output is gaussian

Target Distribution: $q(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_t(x_t), \Sigma_t(x_t))$

Approximated Distribution: $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$

What is $q(x_{t-1} \mid x_t, x_0)$



If we know x_0 and x_t

$q(x_{t-1}|x_t, x_0)$ is deterministic

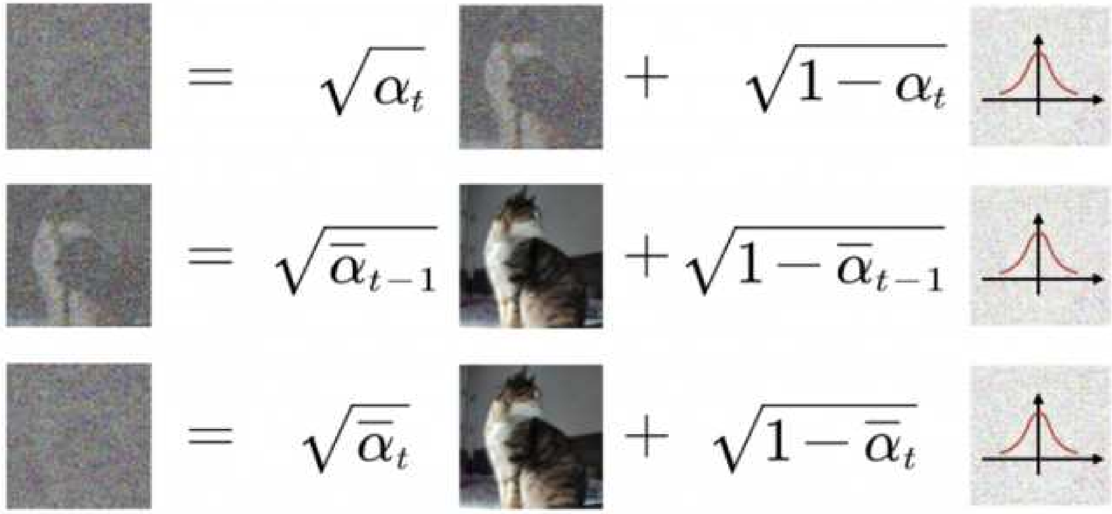
Assume: Markov

$$q(x_{t-1} \mid x_t, x_0) = \frac{q(x_{t-1}, x_t, x_0)}{q(x_t, x_0)} = \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)q(x_0)}{q(x_t|x_0)q(x_0)} = \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$$q(x_t \mid x_{t-1}) \sim \mathcal{N}\left(x_t; \sqrt{\alpha_t}x_{t-1}, 1-\alpha_t\right)$$

$$q(x_{t-1} \mid x_0) \sim \mathcal{N}\left(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, 1-\bar{\alpha}_{t-1}\right)$$

$$q(x_t \mid x_0) \sim \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, 1-\bar{\alpha}_t\right)$$



Remove x_0

③**Reverse** If we know x_0

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}}_{\mu_q(\mathbf{x}_t,t)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}_{\sum_q(t)} I\right)$$

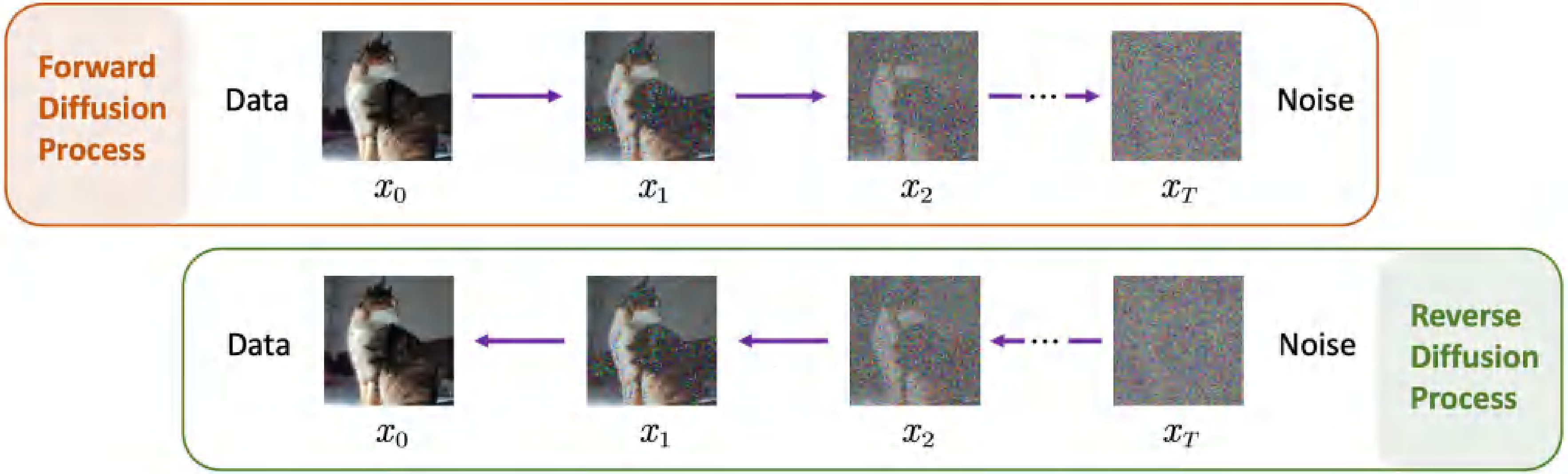
① **Forward (close-form)**

$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon_t \Rightarrow x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_t}{\sqrt{\bar{\alpha}_t}}$

$\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t} \Rightarrow \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_t\right)$ noise predictor $\varepsilon_t(x_0 \to x_t)$

Why not predict x_0 directly?

Summary

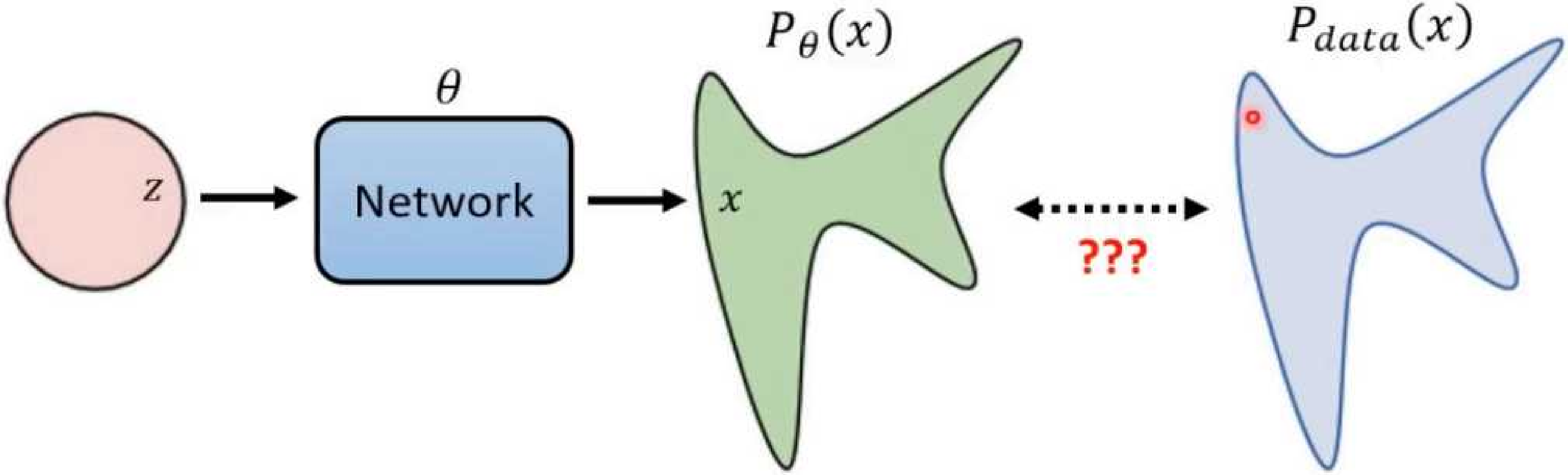


$$Forward\ Process:\ \ q(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$$

$$Reverse\ Process:\ \ p(x_{t-1}|x_t) = N(x_{t-1}; \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_t), \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t I)$$

Reverse Diffusion Process

Maximum Likelihood Estimation



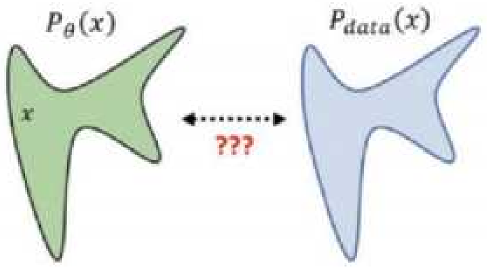
Sample $\{x^1, x^2, \ldots, x^m\}$ from $P_{data}(x)$

We can compute $P_\theta(x^i)$

???

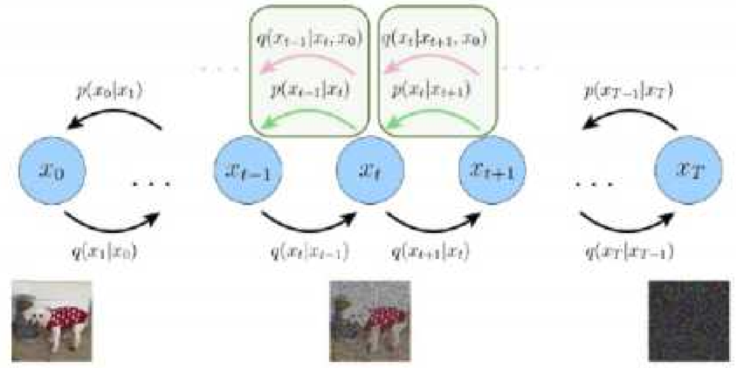
$$\theta^* = arg \max_\theta \prod_{i=1}^{m} P_\theta(x^i)$$

Maximum Likelihood Estimation



$$\arg\max_{\theta} \prod_{i=1}^{t} p_\theta(\boldsymbol{x}_i)$$

$$\arg\max_{\theta} \sum_{i=1}^{t} \log p_\theta(\boldsymbol{x}_i)$$



②Optimization (view 1)

$$\min \; -\log p_\theta(x_0) \leq -\log p_\theta(x_0) + D_{KL}\left(q(x_{1:T} \mid x_0) \,\|\, p_\theta(x_{1:T} \mid x_0)\right)$$

$$\min \; -\log p_\theta(x_0) \leq \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{q(x_{1:T} \mid x_0)}{p_\theta(x_{0:T})}\right] \quad \textbf{(ELBO)}$$

$$\min \; -\log p_\theta(x_0) \leq \mathbb{E}_{q(x_{1:T}|x_0)}\left[\underbrace{D_{KL}\left(q(x_T \mid x_0) \,\|\, p_\theta(x_T)\right)}_{\text{prior term}}\right] + \sum_{t=2}^{T} \underbrace{D_{KL}\left(q(x_{t-1} \mid x_t, x_0) \,\|\, p_\theta(x_{t-1} \mid x_t)\right)}_{\text{reconstruction term}} - \log p_\theta(x_0 \mid x_1)$$

Derivation Process

$$
\begin{aligned}
L_{\mathrm{VLB}} &= \mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big] \\
&= \mathbb{E}_q\Big[\log \frac{\prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_T)\prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=1}^T \log \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \Big(\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \cdot \frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}\Big) + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big] \\
&= \mathbb{E}_q\Big[-\log p_\theta(\mathbf{x}_T) + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{q(\mathbf{x}_1|\mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\Big] \\
&= \mathbb{E}_q\Big[\log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)\Big] \\
&= \mathbb{E}_q[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_T))}_{L_T} + \sum_{t=2}^T \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}]
\end{aligned}
$$

Derivation Process

$$L_{t-1} = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2\right] + C$$

$$L_{t-1} - C = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}), \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}) - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon})\right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}), t)\right\|^2\right]$$

$$= \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0,\boldsymbol{\epsilon}), t)\right\|^2\right]$$

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t))\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)$$

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

$$L_{\mathrm{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

Diffusion Model Test by Pytorch

Applications

TXT2IMG

LDM&DALLE

LLM

LLaDA

Robotics

Diffusion Policy

Application 01 —— 图像生成 txt2img

什么是txt2img?

通过给定文本提示词（text prompt）输出一张匹配提示词的图片。

Stable Diffusio(LDM)

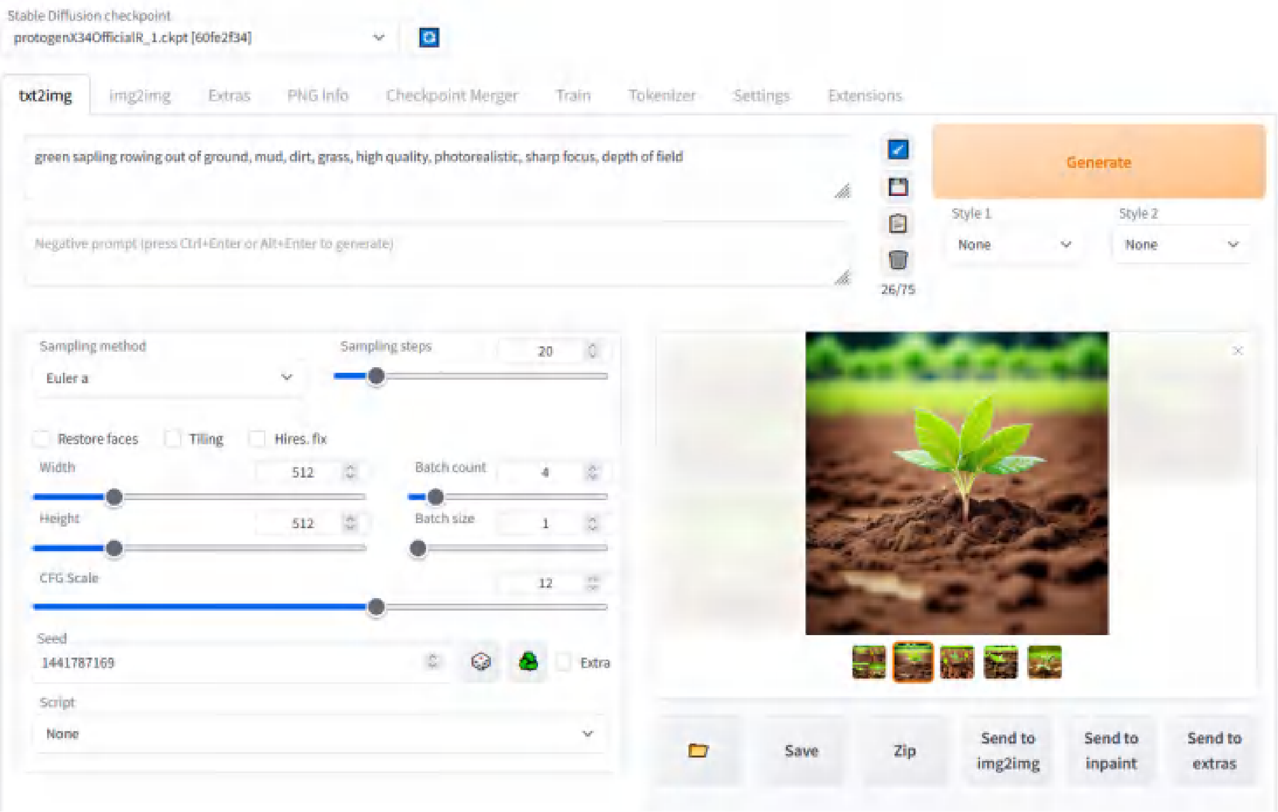
High-Resolution Image Synthesis with Latent Diffusion Models

DALLE 2 By OpenAI

Hierarchical Text-Conditional Image Generation with CLIP Latents

Application 01 —— 图像生成 LDM

Have a try! Open Source

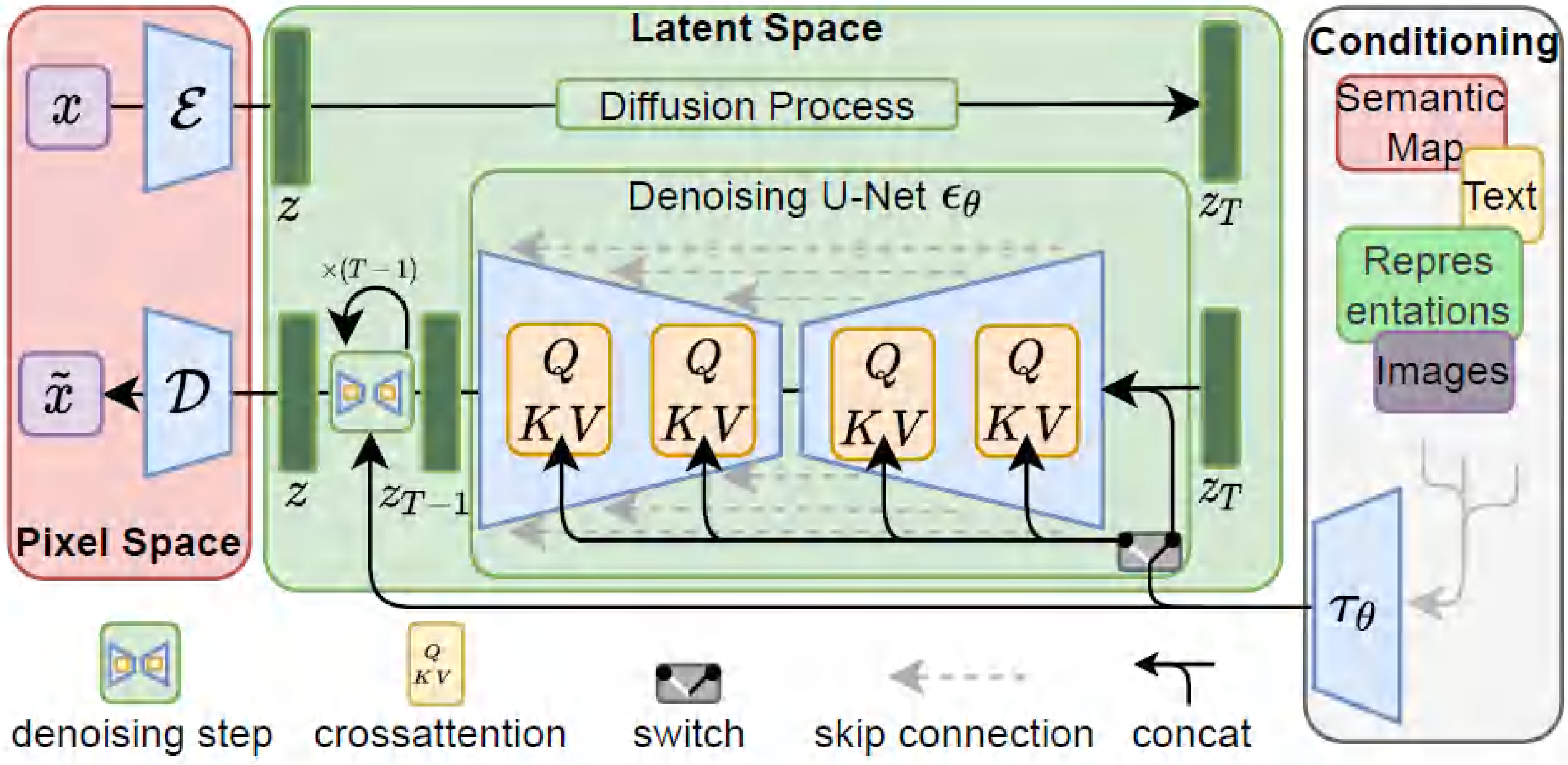


本地部署

https://github.com/AUTOMATIC1111/stable-diffusion-webui

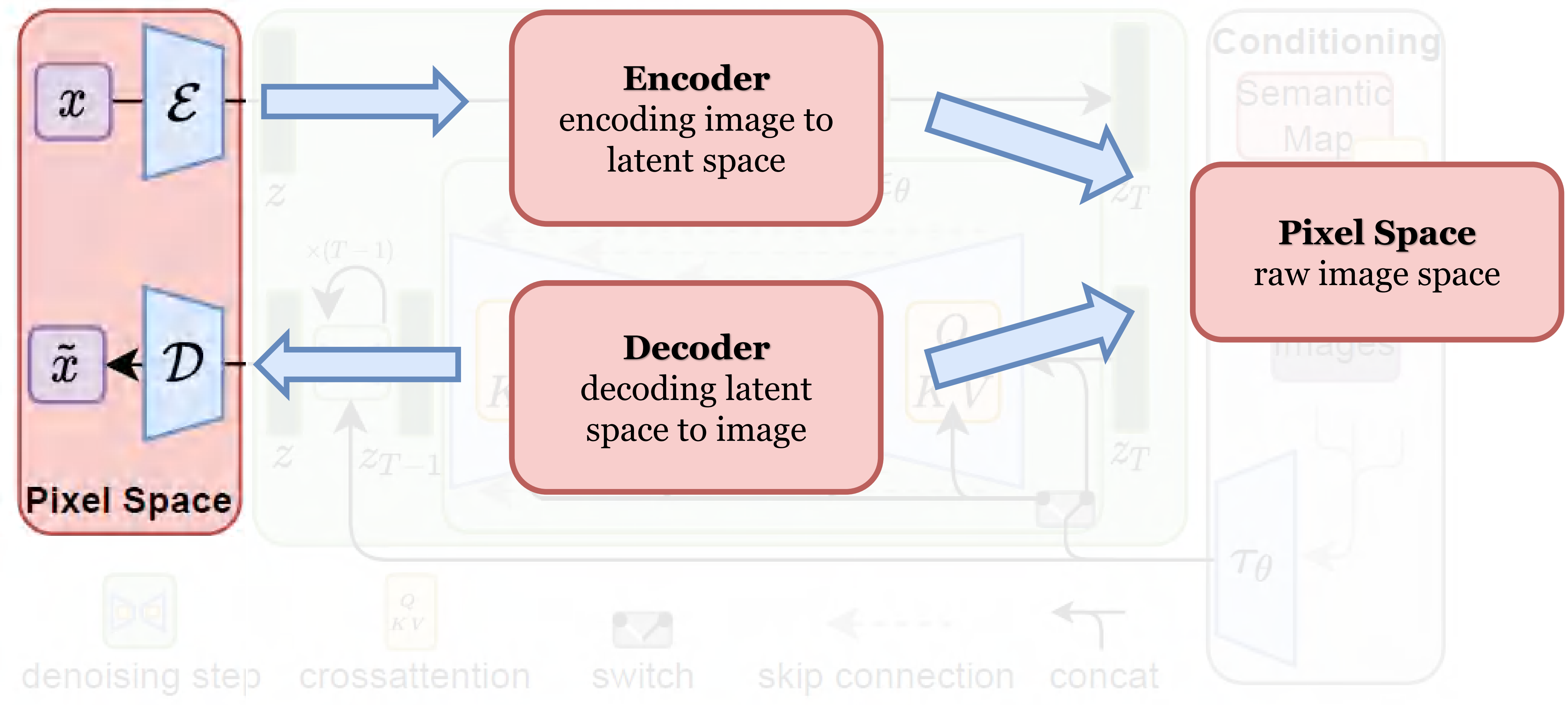
网站访问

Application 01 —— 图像生成 LDM



High-Resolution Image Synthesis with Latent Diffusion Models

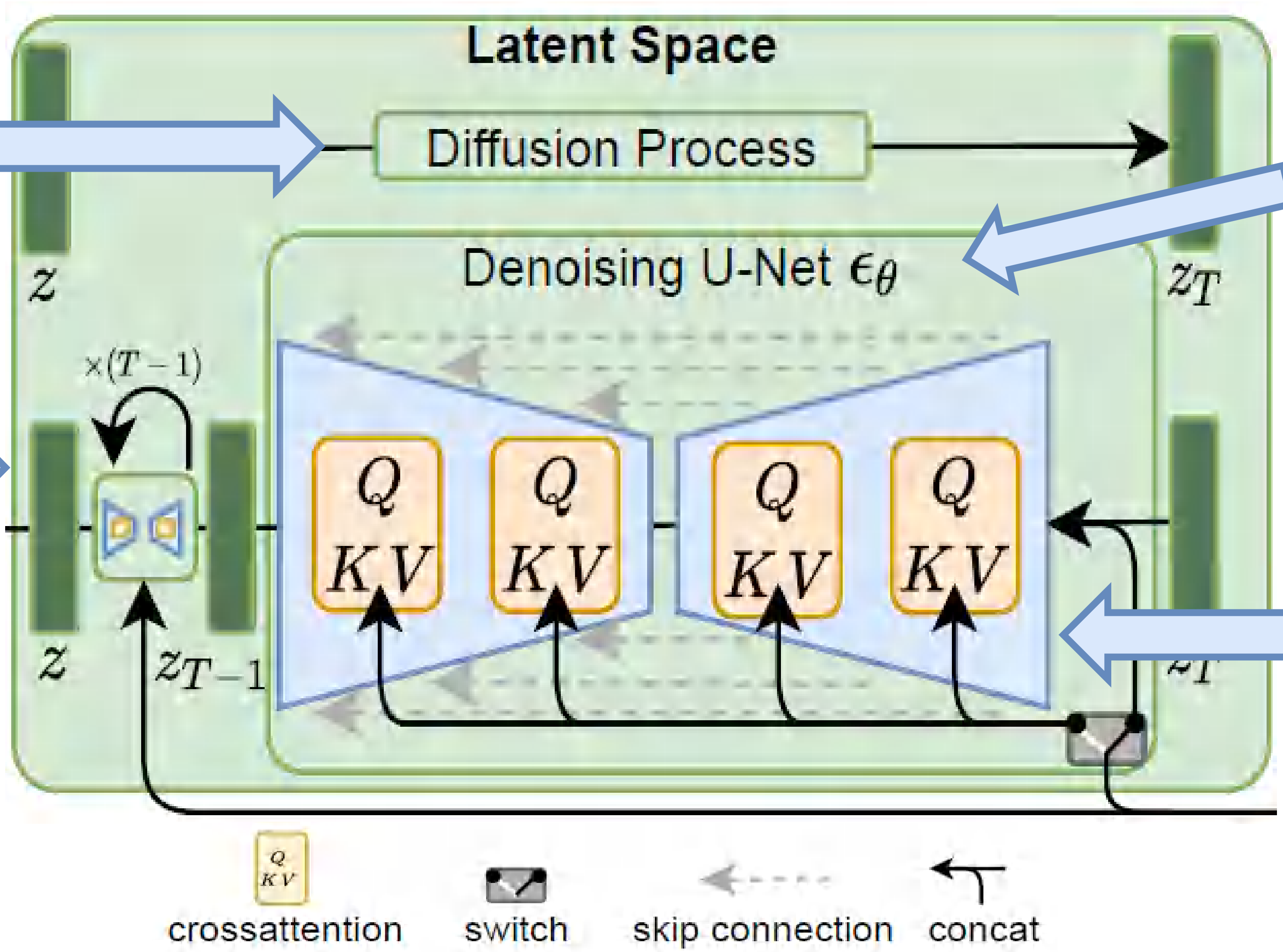
Application 01 —— 图像生成 LDM



Application 01 —— 图像生成 LDM

Diffusion Process
gradual denoising of data

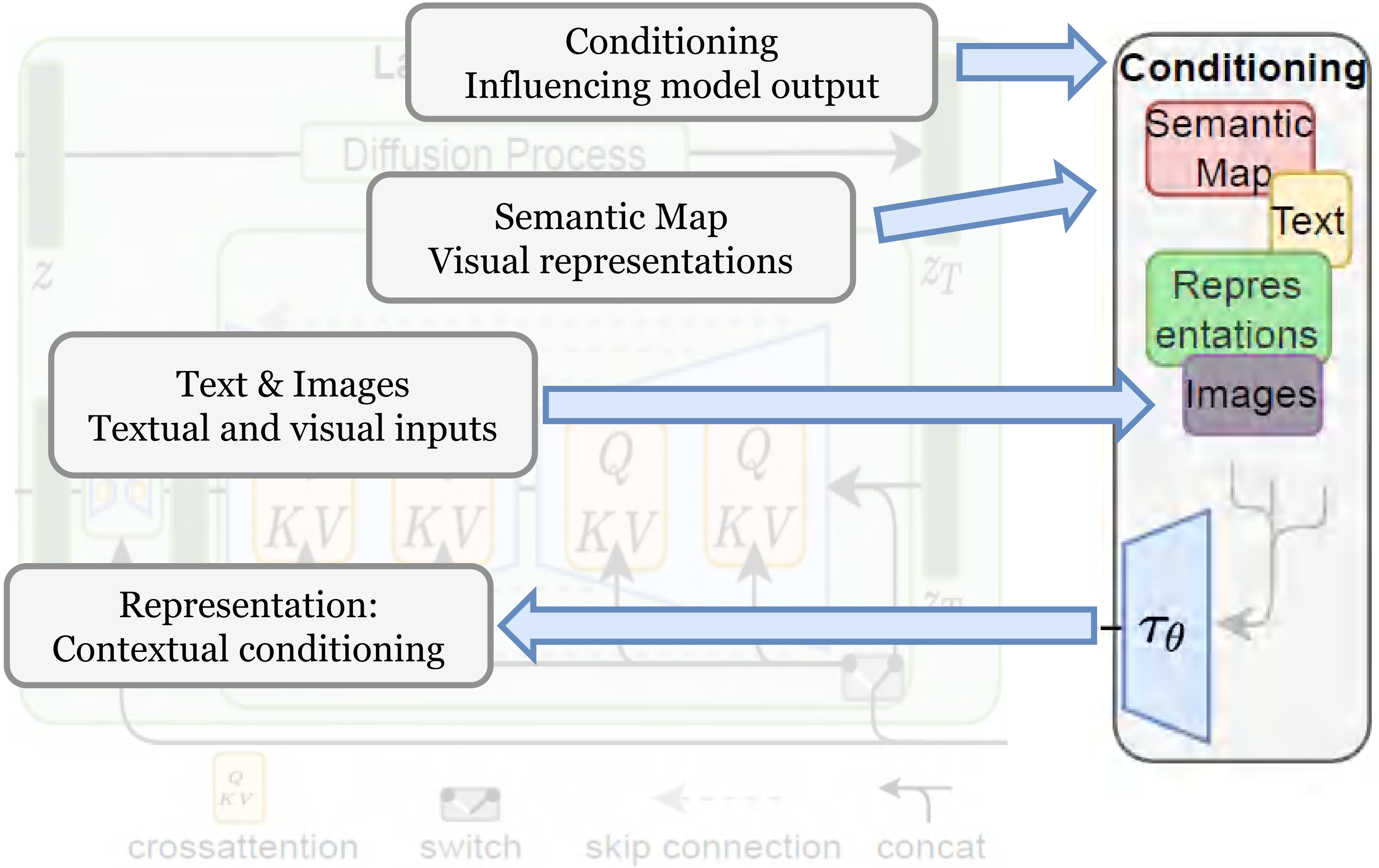
Latent Space
Representation space for diffusion



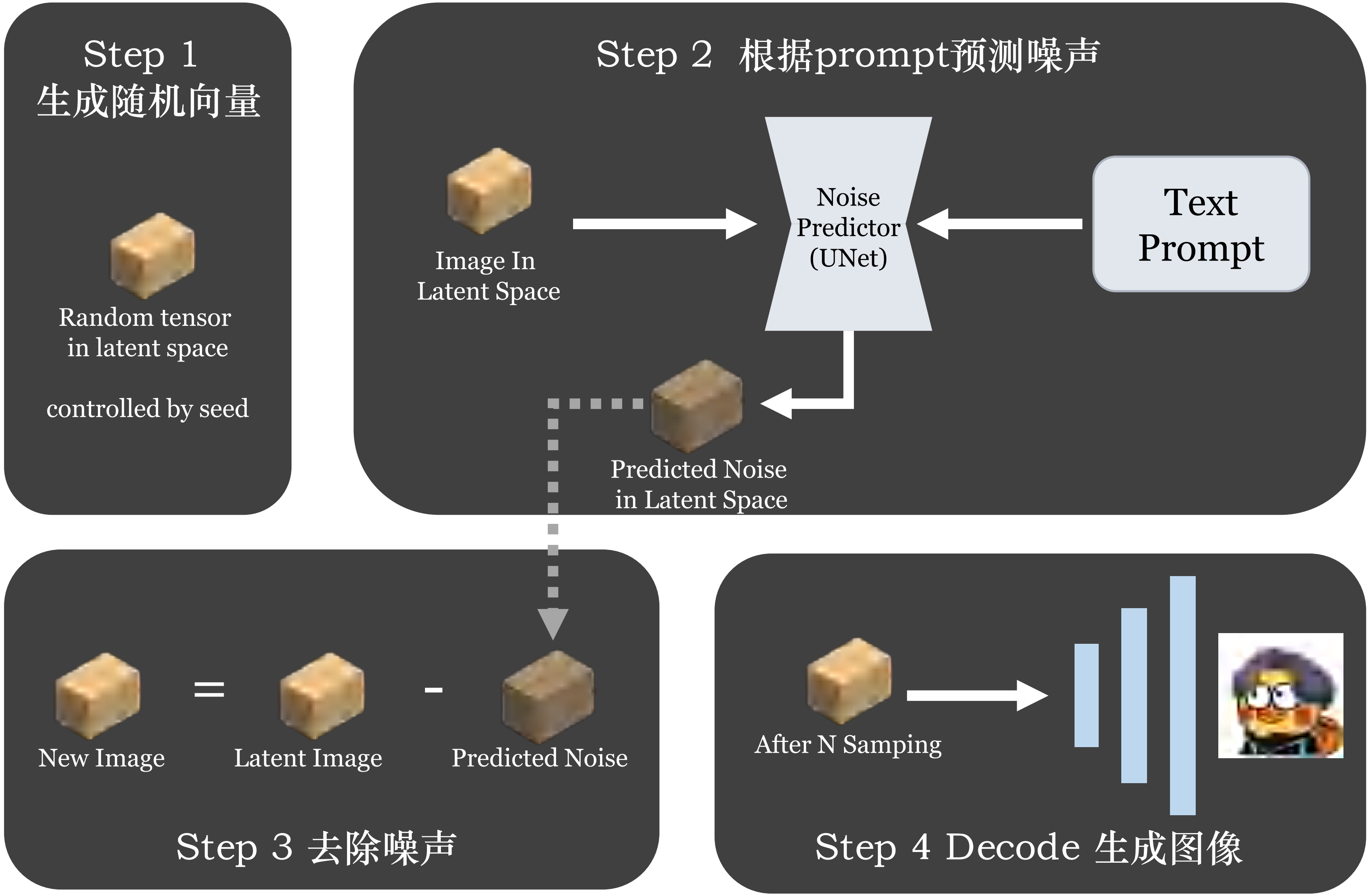
Denoising U-Net:
Denoising architecture

Attention Mechanism
Cross-attention with QKV

Application 01 —— 图像生成 LDM



Application 01 —— LDM 原理



Application 01 —— 图像生成 DALLE2

Have a try! API supported

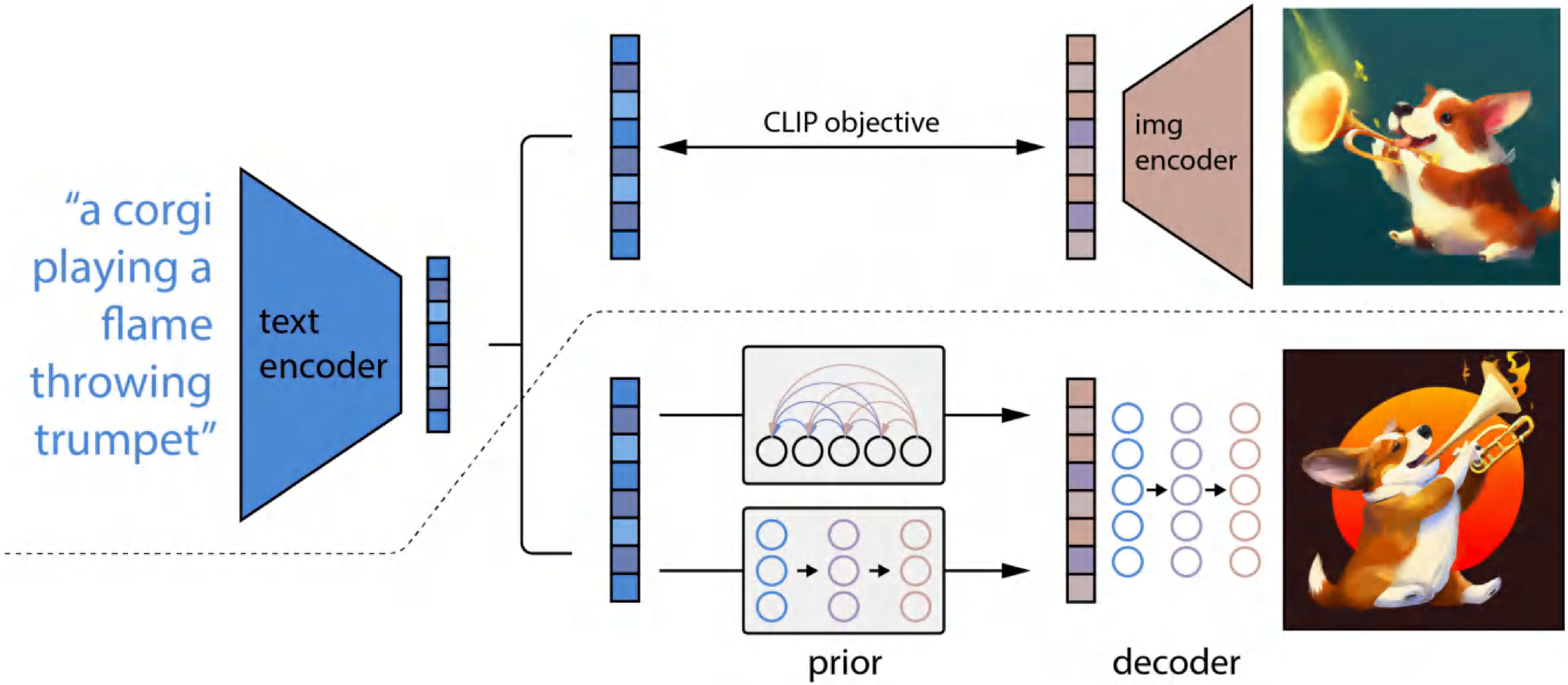
https://platform.openai.com/docs/overview

A photo of a white fur monster standing in a purple room

panda mad scientist mixing sparkling chemicals, artstation

vibrant portrait painting of Salvador Dalí with a robotic half face

Application 01 —— 图像生成 DALLE2



Hierarchical Text-Conditional Image Generation with CLIP Latents

Application 02 —— LLaDA

扩散模型也能玩转大语言模型？

Large Language Diffusion Models

What is LLaDA？ **L**arge **La**nguage **D**iffusion with m**A**sking

A text generation method different from the traditional **left-to-right** approach

Application 02 —— LLaDA

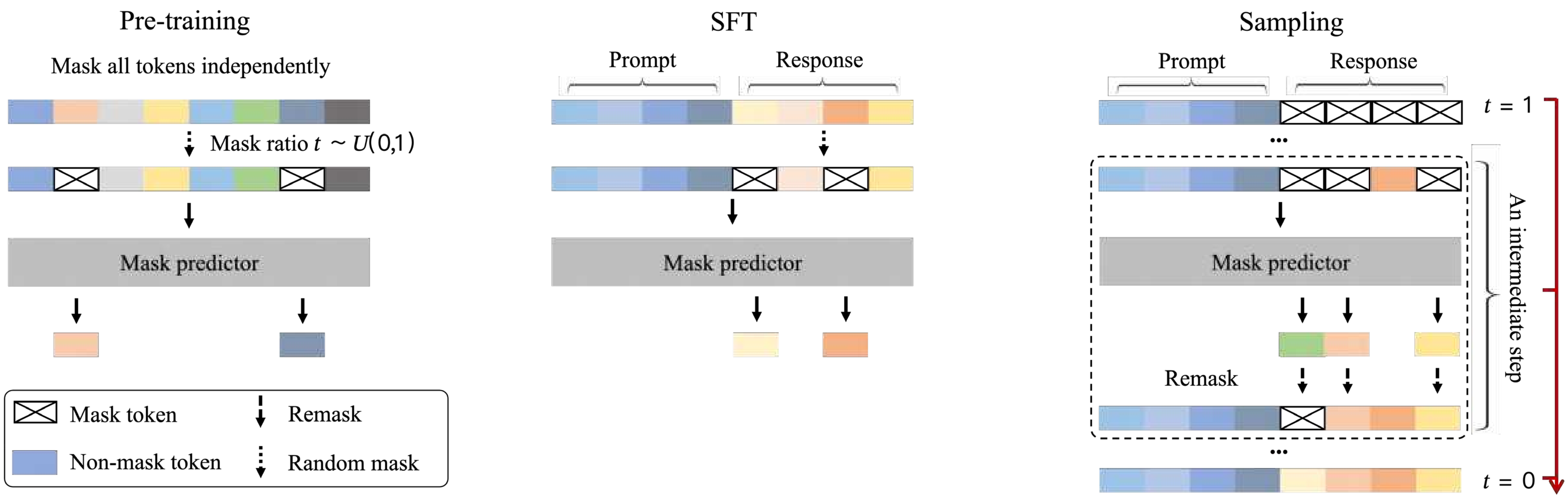
传统路径：Auto Regression

每次生成一个token，新生成的token会拼到序列末尾
每个token的生成依赖于之前所有已生成的tokens

Talk is Cheap，Show me the Code!

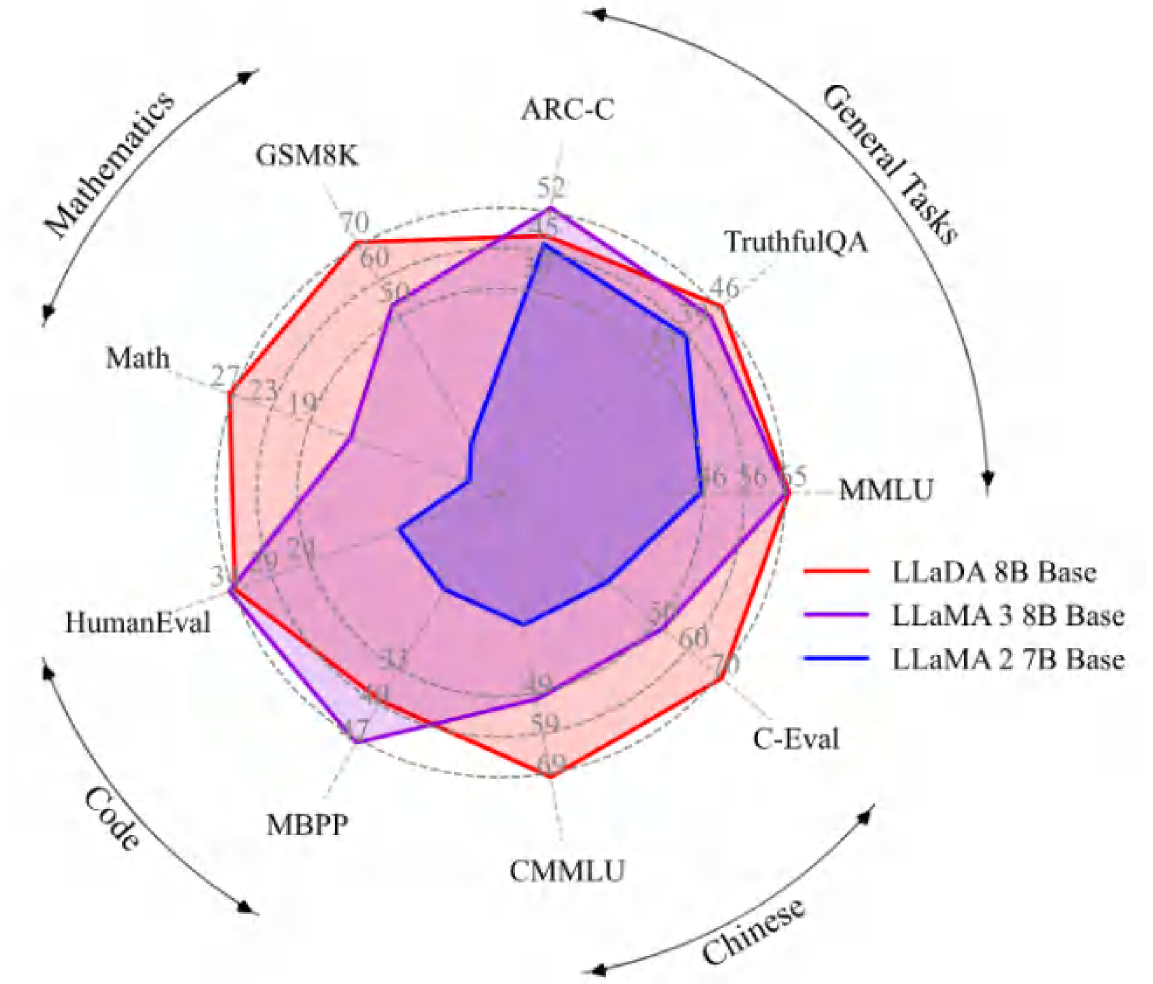
LLaDA

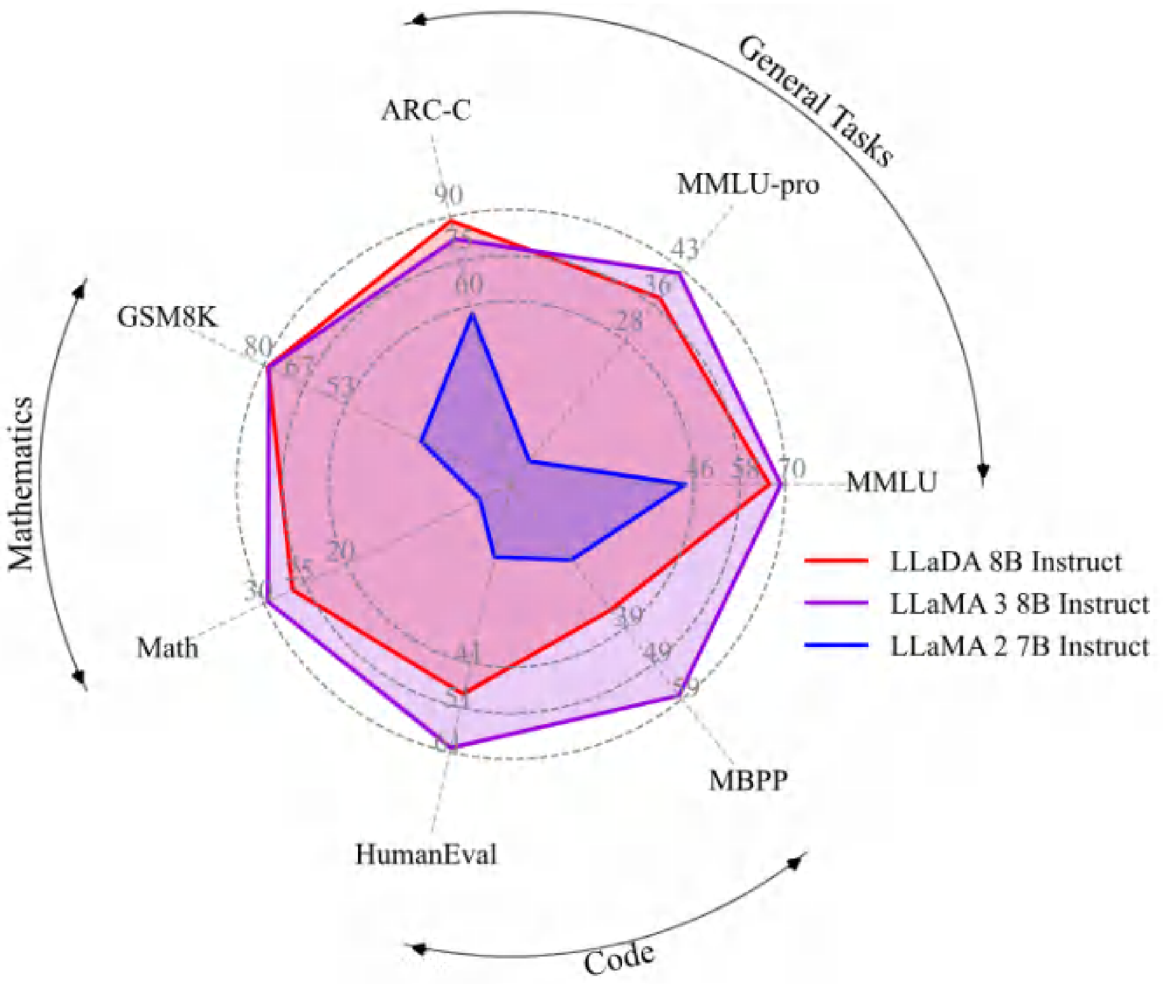
Mask的过程体现了diffusion加噪去噪的思想



Application 02 —— LLaDA

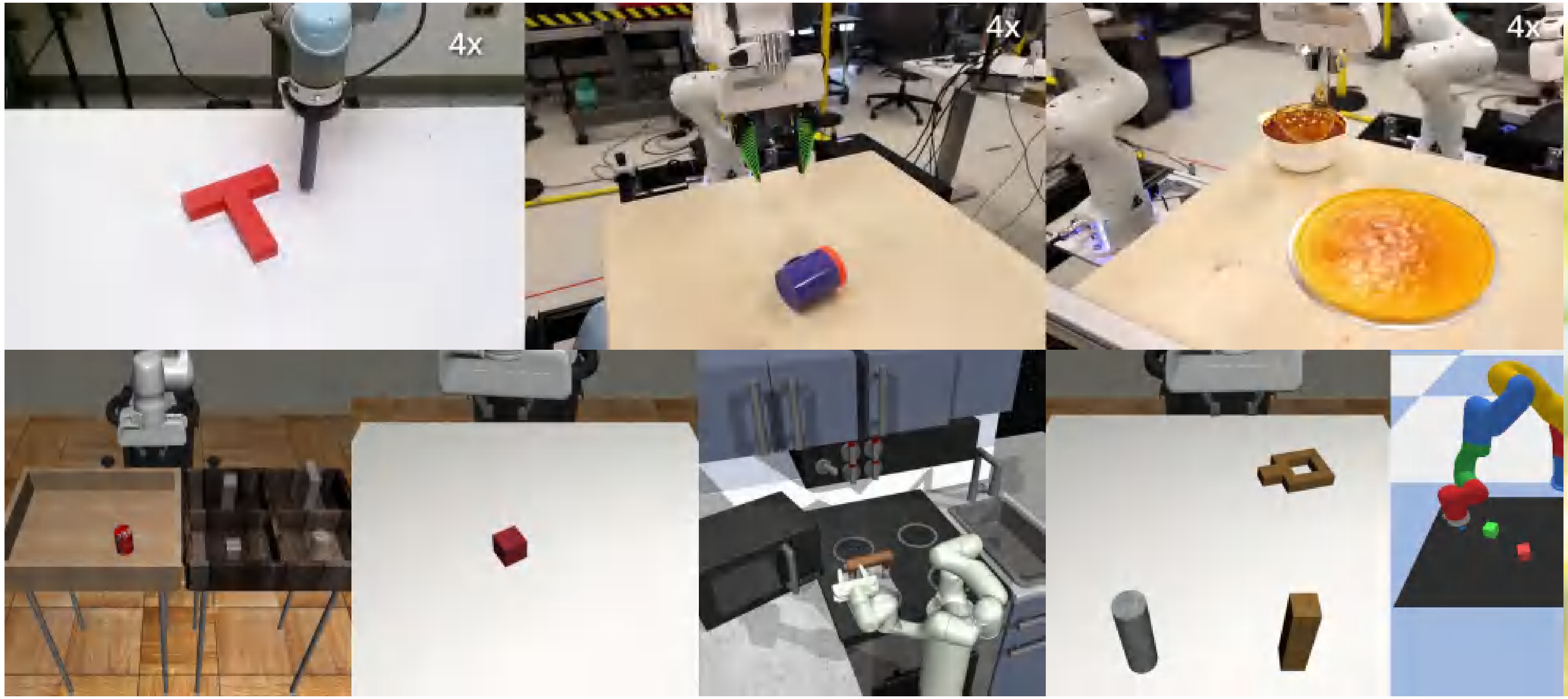
Performance





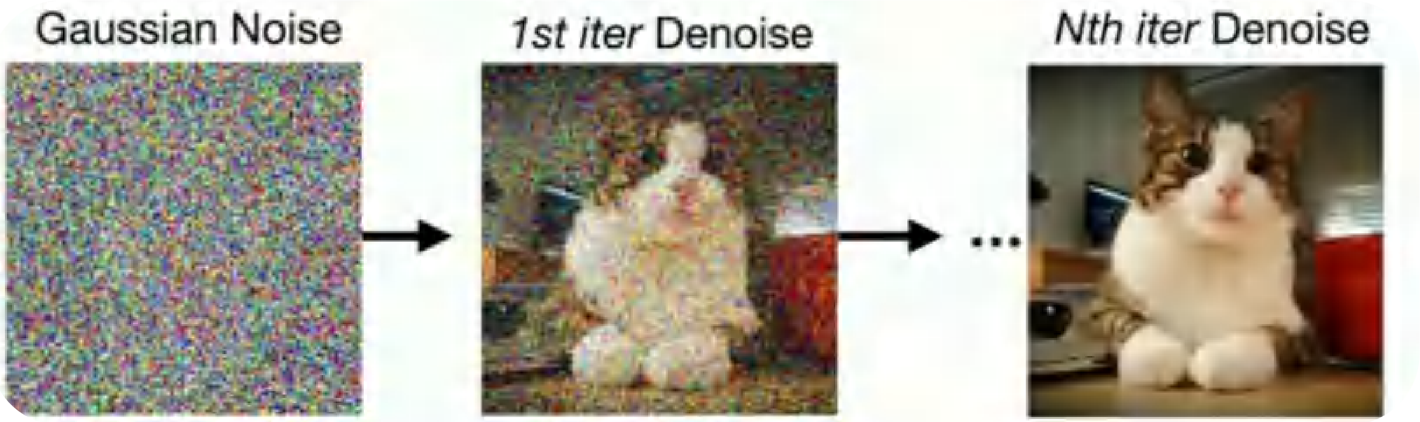
Application 03 —— Robotics

Diffusion Policy

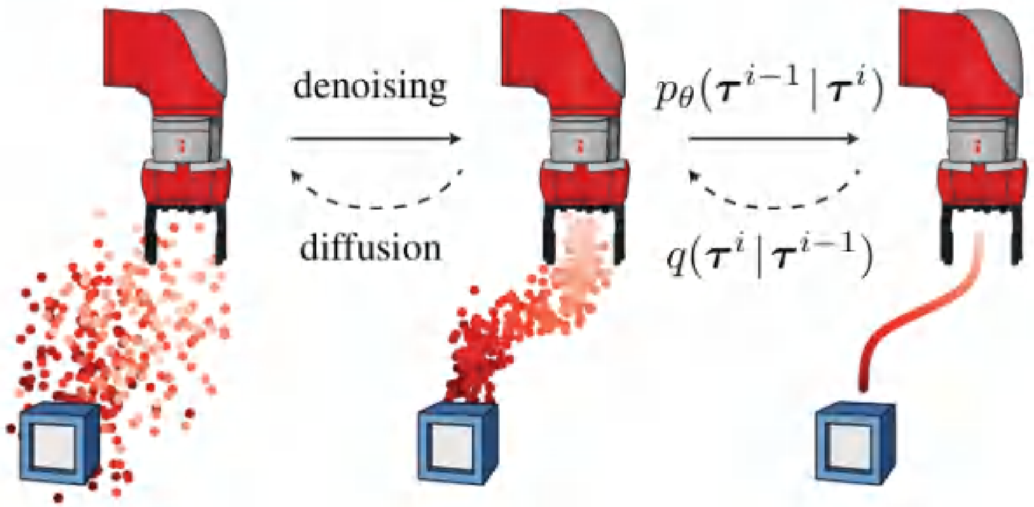


Application 03 —— Robotics

Image Diffusion



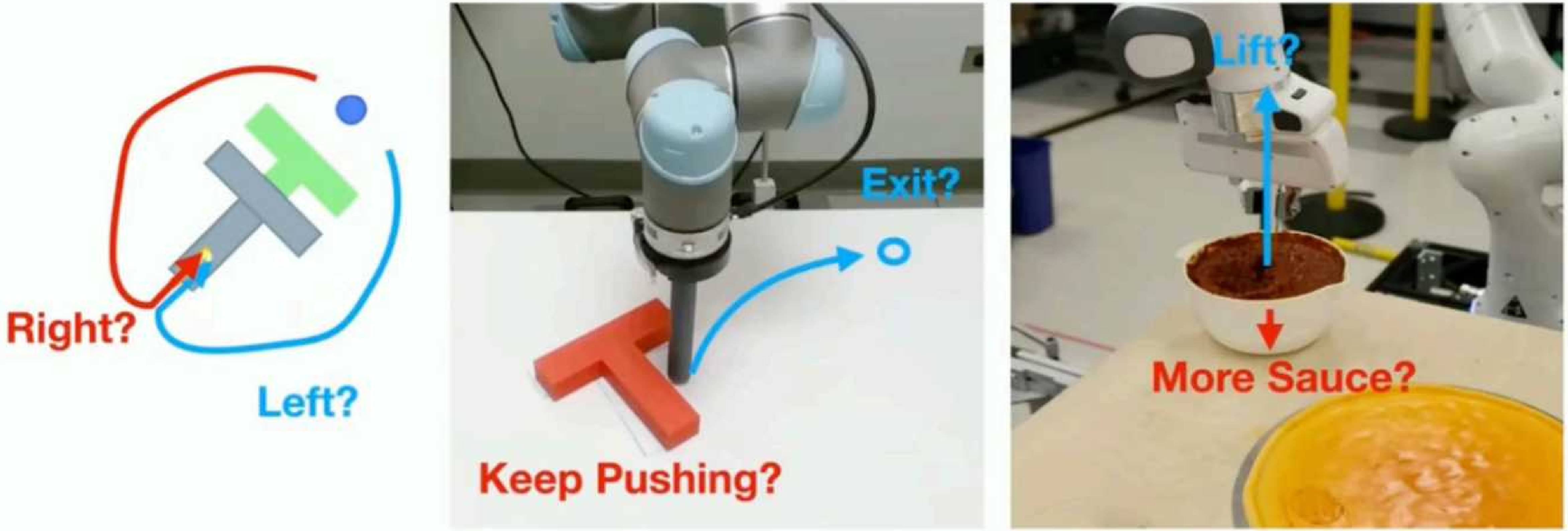
Action Diffusion



Application 03 —— Diffusion Policy

Action Multimodality

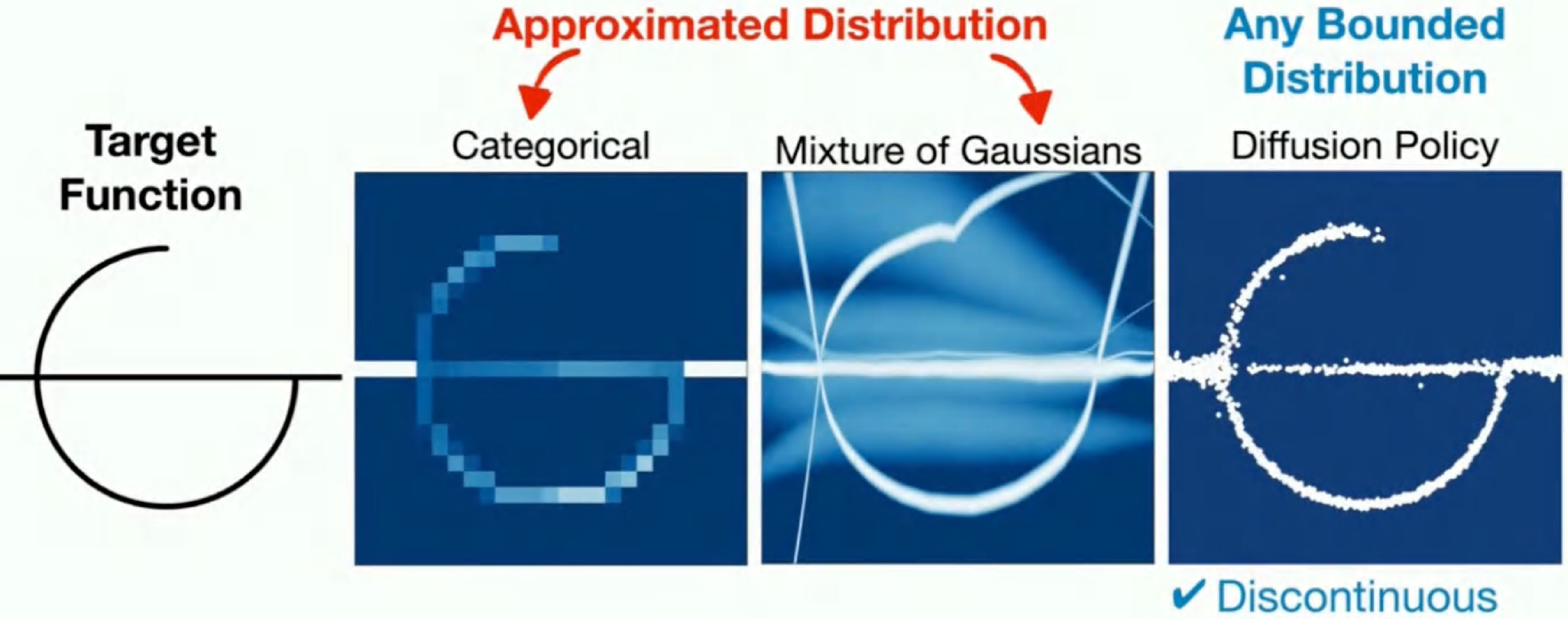
Multiple valid actions for the same observation



Surprisingly Common

Application 03 —— Diffusion Policy

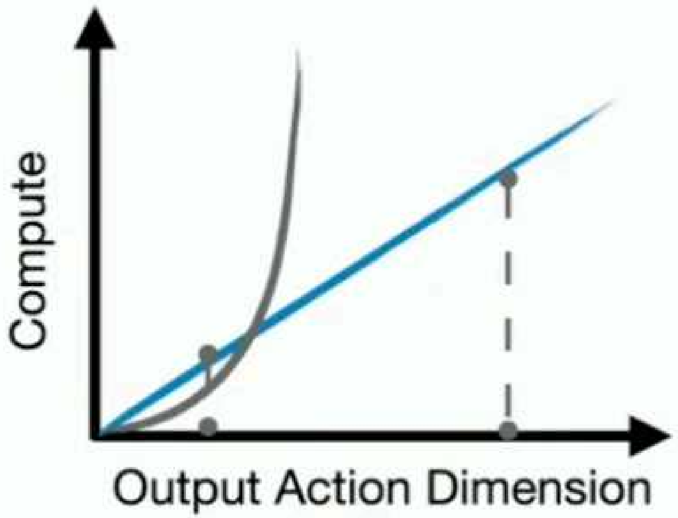
Action Multimodality

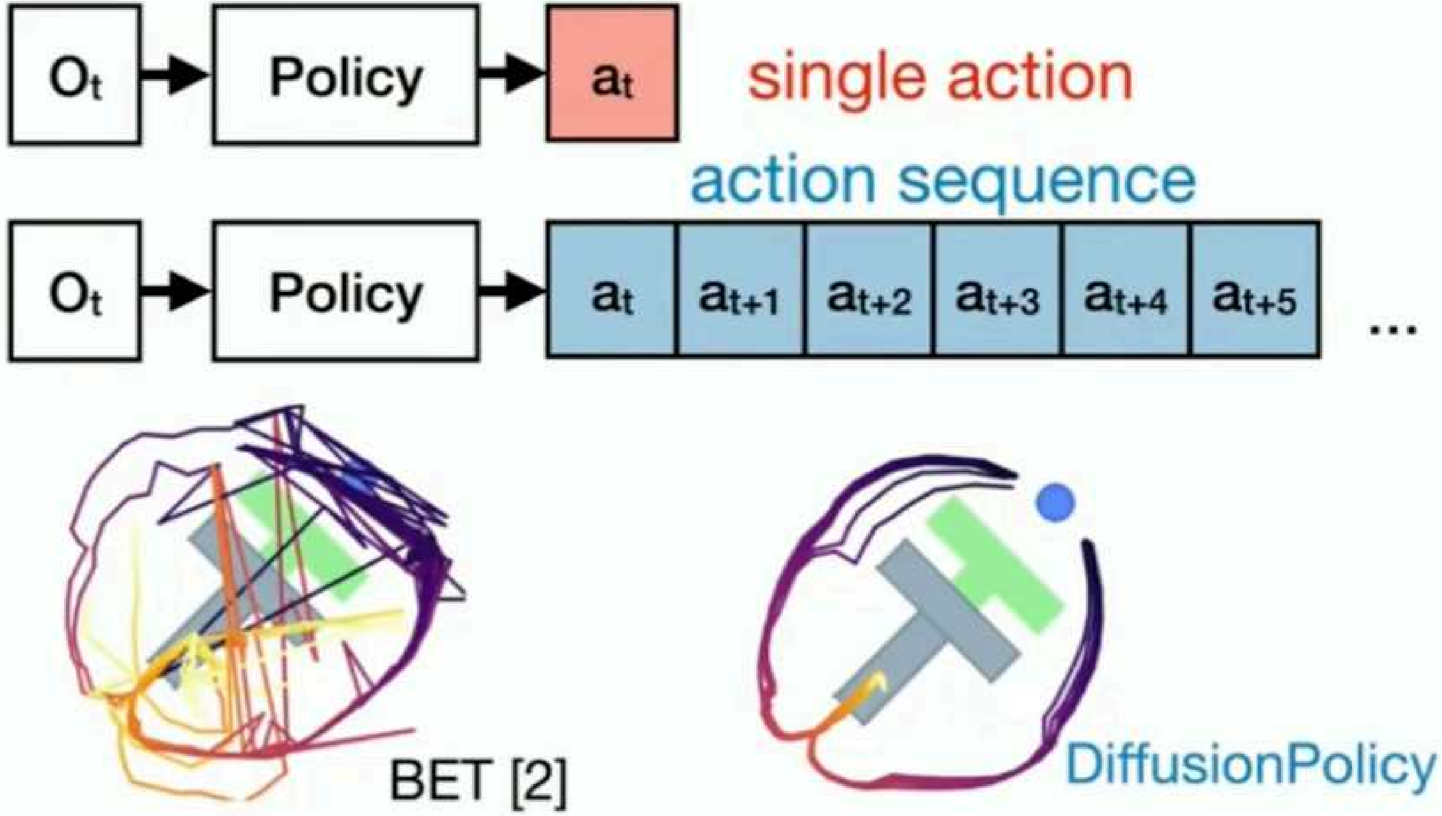


Application 03 —— Diffusion Policy

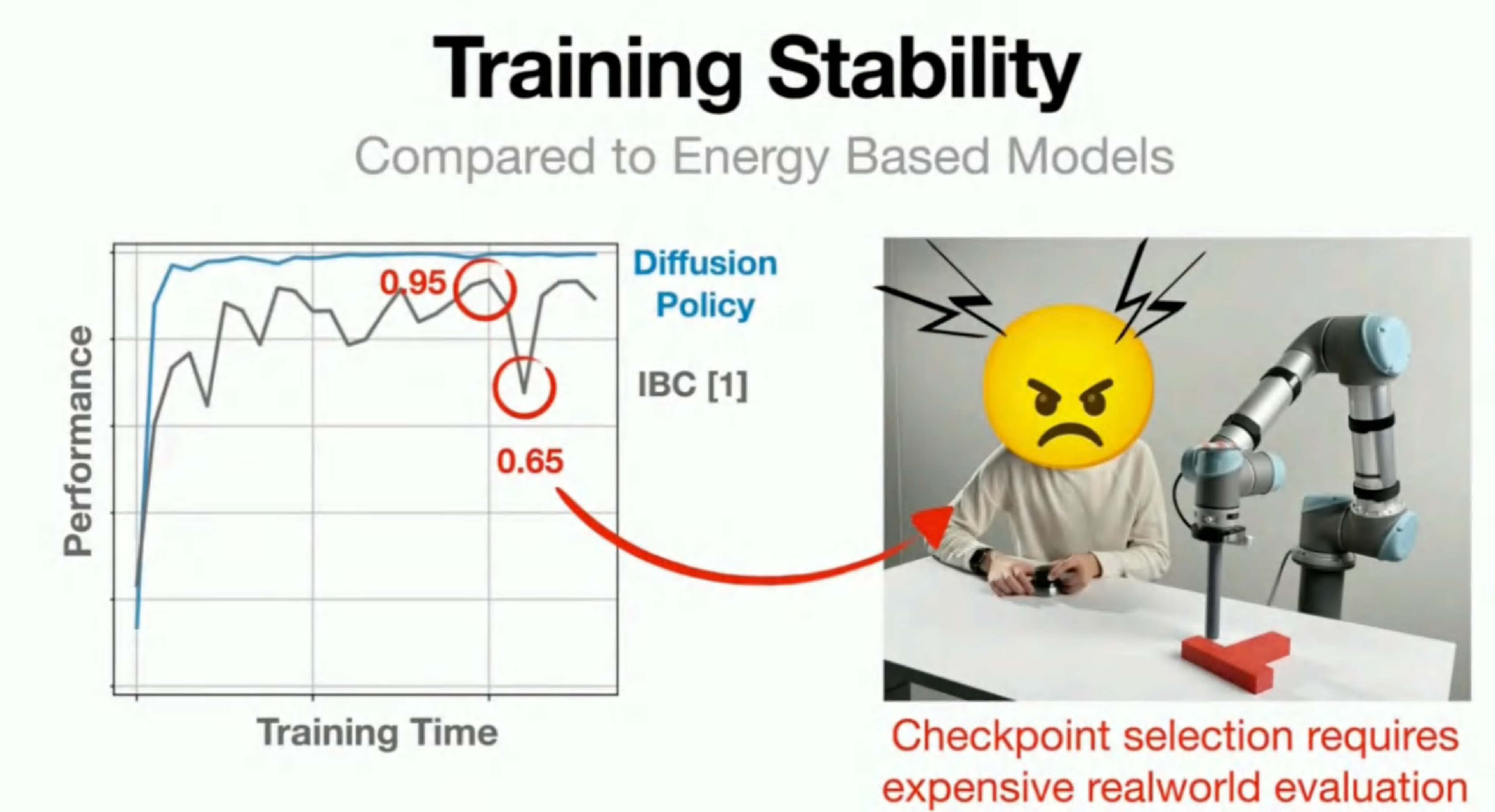
Action Space Scalability

Easily afford action sequence prediction





Application 03 —— Diffusion Policy



Application 03 —— Diffusion Policy

Approach

Step1 训练阶段：学习“去噪”过程　经过多次迭代去噪，模型生成出符合当前观察条件的“干净”动作x0

$$x_{k-1} = \alpha(x_k - \gamma\epsilon_\theta(x_k, k) + N(0, \sigma^2 I))$$

去噪步长的缩放系数　噪声预测函数

$$L = MSE(\epsilon_k, \epsilon_\theta(x_0 + \epsilon_k, k))$$

噪声预测函数

Step2 得分匹配来优化动作的能量

$$\nabla_a \log p(a|o) = -\nabla_a E_\theta(a, o)$$

沿着降低能量的方向调整动作

Step3 InfoNCE损失：区分“好”动作与“坏”动作

$$L_{infoNCE} = -\log\left(\frac{e^{-E_\theta(o,a)}}{e^{-E_\theta(o,a)} + \sum_{j=1}^{N_{neg}} e^{-E_\theta(o,\tilde{a}_j)}}\right)$$

“好”动作

“坏”动作

Step4 推理阶段：实时生成动作序列

Reference

If you are interested in this topic, check these papers uploaded in DingTalk Group

Thanks