使用强化学习技术微调扩散模型教程与综述

赵雨来
普林斯顿大学
2024/08/07

关于

- 赵雨来
- 普林斯顿大学，即将PhD三年级。在清华获得学士学位。
- 主要研究方向：强化学习，生成模型，AI for Science。
- 曾在华盛顿大学，苏黎世联邦理工，基因泰克实习。

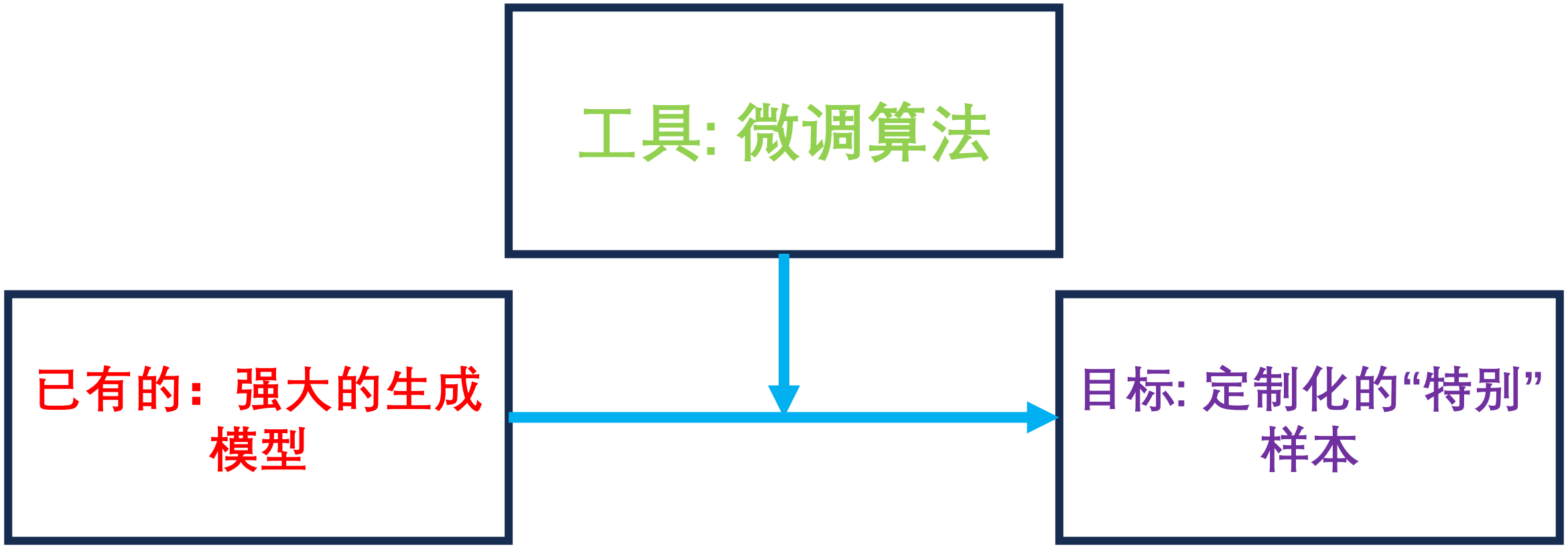
- More about me: https://yulaizhao.com/

关于

- 预印论文： https://arxiv.org/pdf/2407.13734

- 作者: Masatoshi Uehara*, Yulai Zhao*, Tommaso Biancalani, Sergey Levine

总体路线：AI辅助的设计



例子 —— 图像

预训练的生成模型

微调

根据艺术效果，版面布局或是文件大小定制的图像

例子：图像——以“艺术效果”为目标

预训练的生成模型

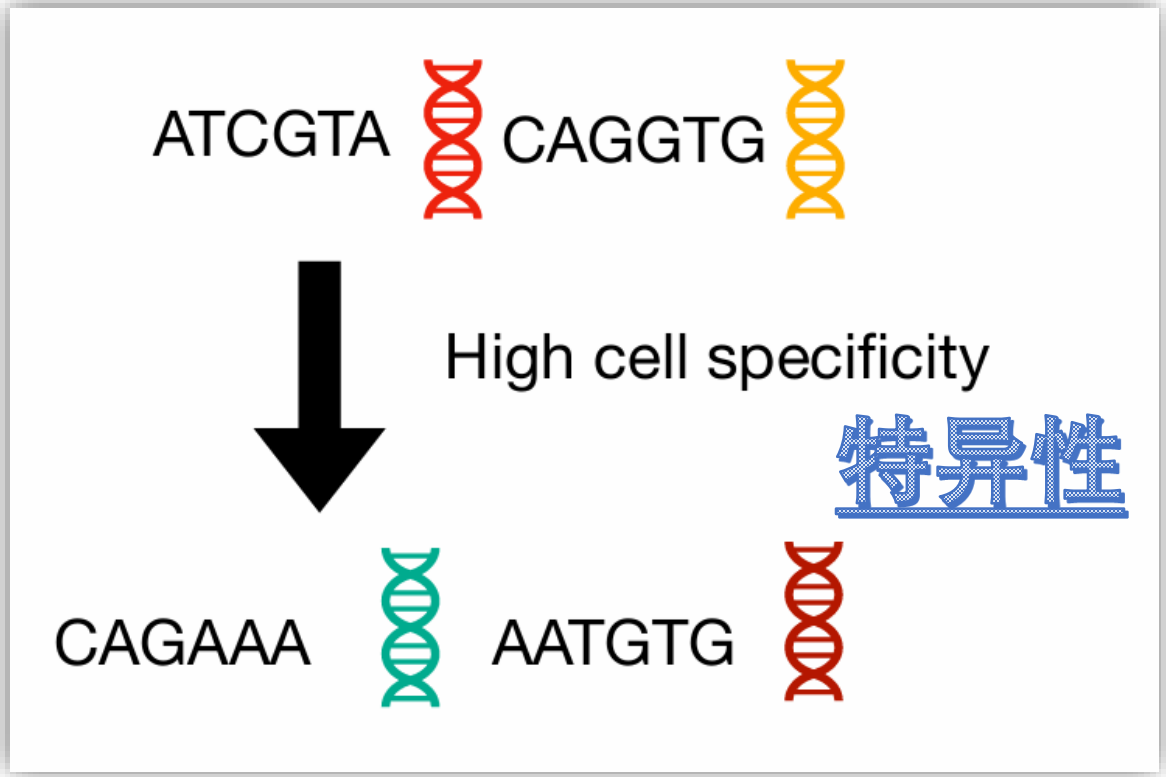
艺术效果得到提高

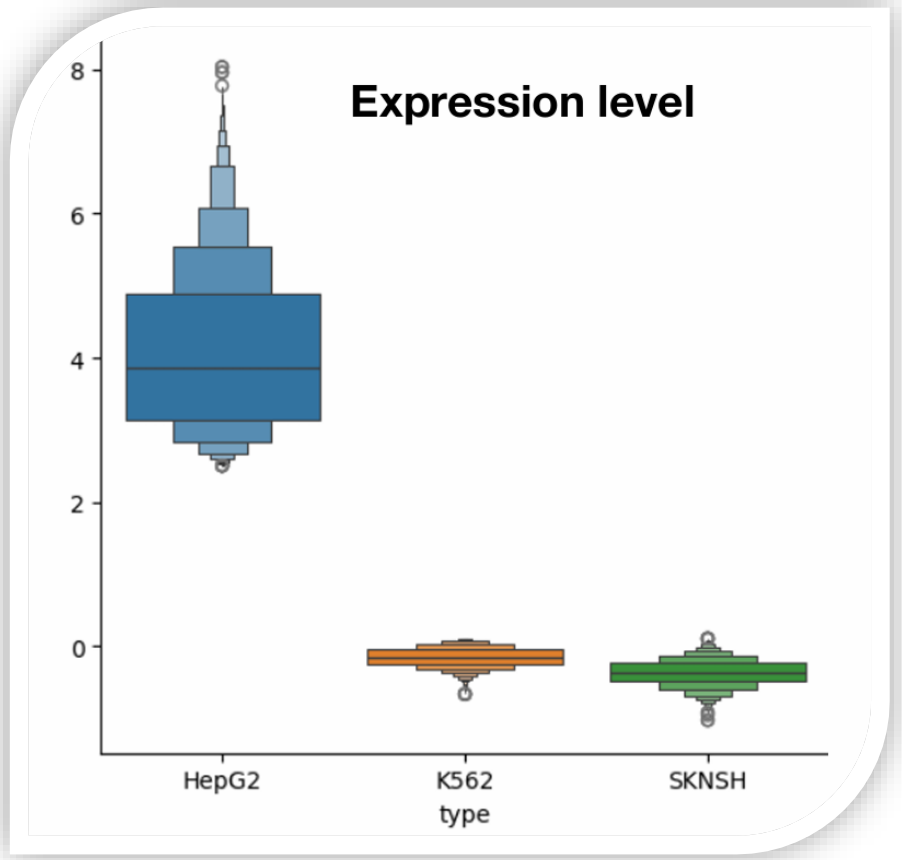
针对“艺术效果”进行微调

例子 —— 生物序列设计

DNA 增强子

（DNA上一小段区域，可促进基因的转录作用）





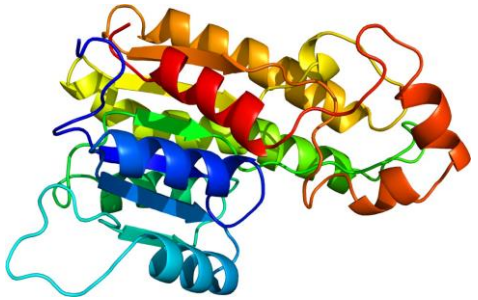
设计目标：仅在单一细胞系内达到高活性！

为何重点研究扩散模型?

- 在图像生成中极为成功

为何重点研究扩散模型?

- 在科学领域愈加广泛



DiscDiff: Latent Diffusion Model for DNA Sequence Generation

Zehui Li [1] Yuhao Ni [2] William A V Beardall [1] Guoxuan Xia [2] Akashaditya Das [1] Guy-Bart Stan [1] Yiren Zhao [2]

Dirichlet Diffusion Score Model for Biological Sequence Generation

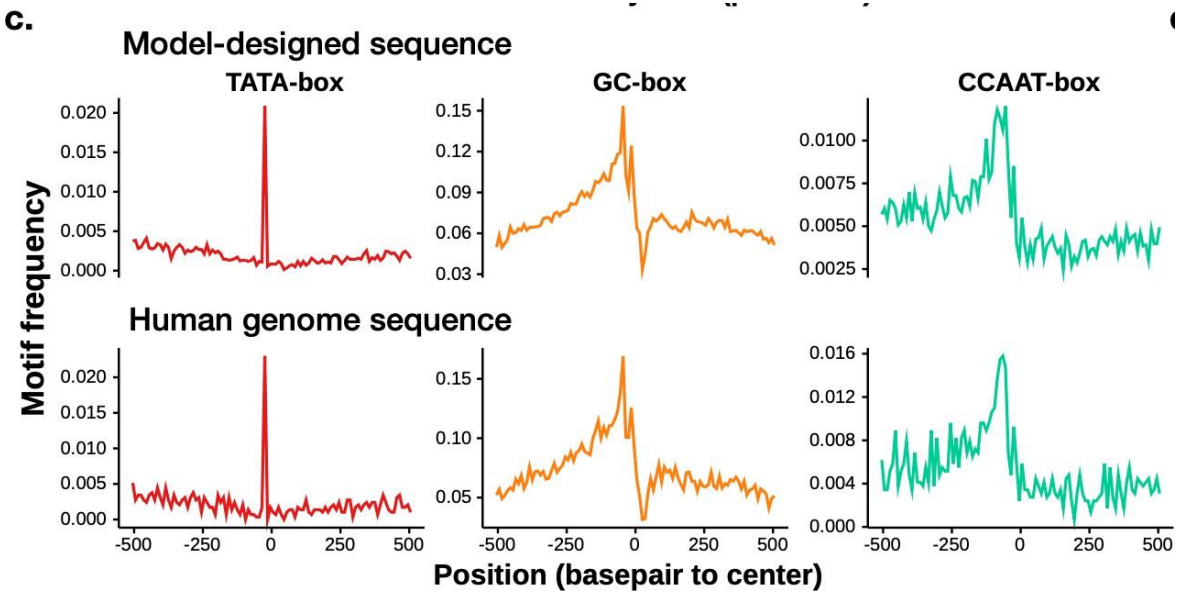
Pavel Avdeyev [1] Chenlai Shi [1] Yuhao Tan [1] Kseniia Dudnyk [1] Jian Zhou [1]

Generative Flows on Discrete State-Spaces: Enabling Multimodal Flows with Applications to Protein Co-Design

Andrew Campbell [*1] Jason Yim [*2] Regina Barzilay [2] Tom Rainforth [1] Tommi Jaakkola [2]

Dirichlet Flow Matching with Applications to DNA Sequence Design

Hannes Stark [*1] Bowen Jing [*1] Chenyu Wang [1] Gabriele Corso [1] Bonnie Berger [1 2] Regina Barzilay [1] Tommi Jaakkola [1]



内容

1. 背景介绍
    1. 扩散模型
    2. 奖励函数
    3. 扩散模型的微调

2. 基于强化学习的微调算法
    1. 奖励函数(reward function)可访问
    2. 奖励函数未知
        1. 奖励函数需要从给定的离线数据中学习 (offline data)
        2. 需要从与环境的在线交互中获得数据以学习奖励函数 (lab-in-the-loop)

3. 通过微调实现条件生成

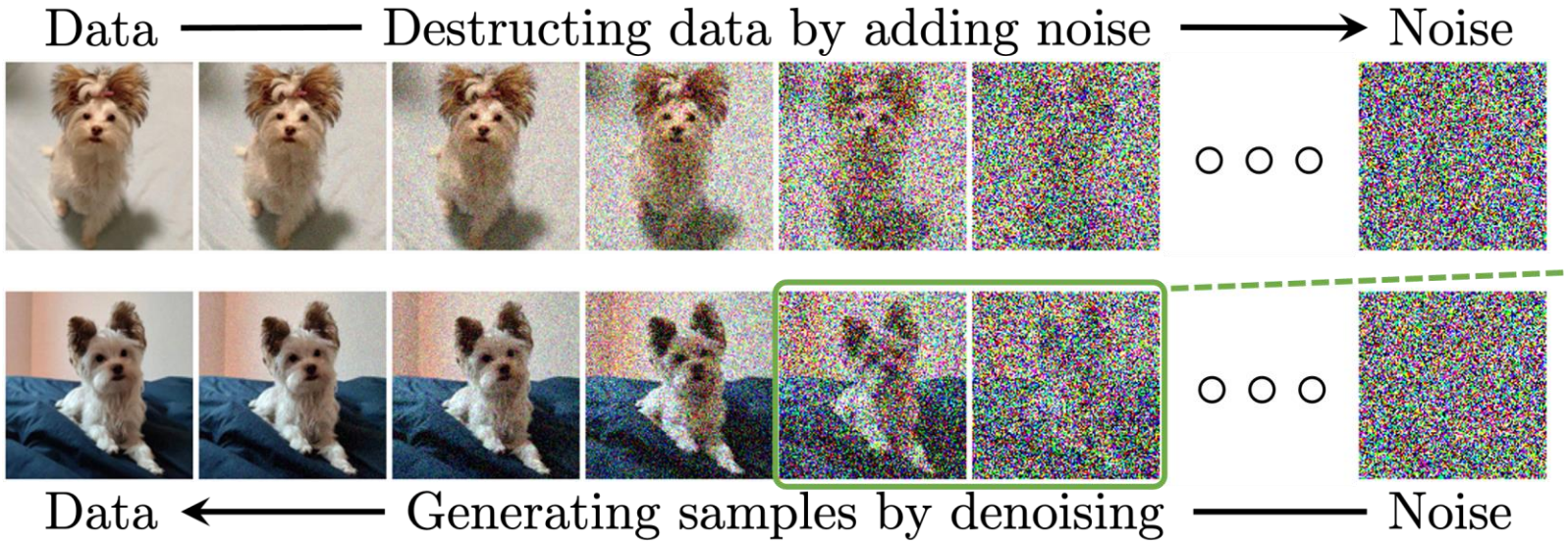
4. 总结

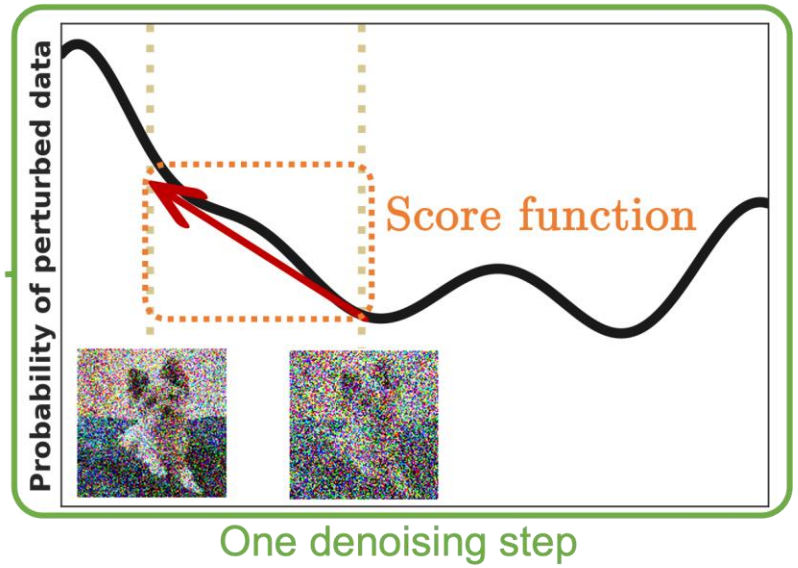
1.背景介绍

扩散模型基础

前置 —— 扩散模型

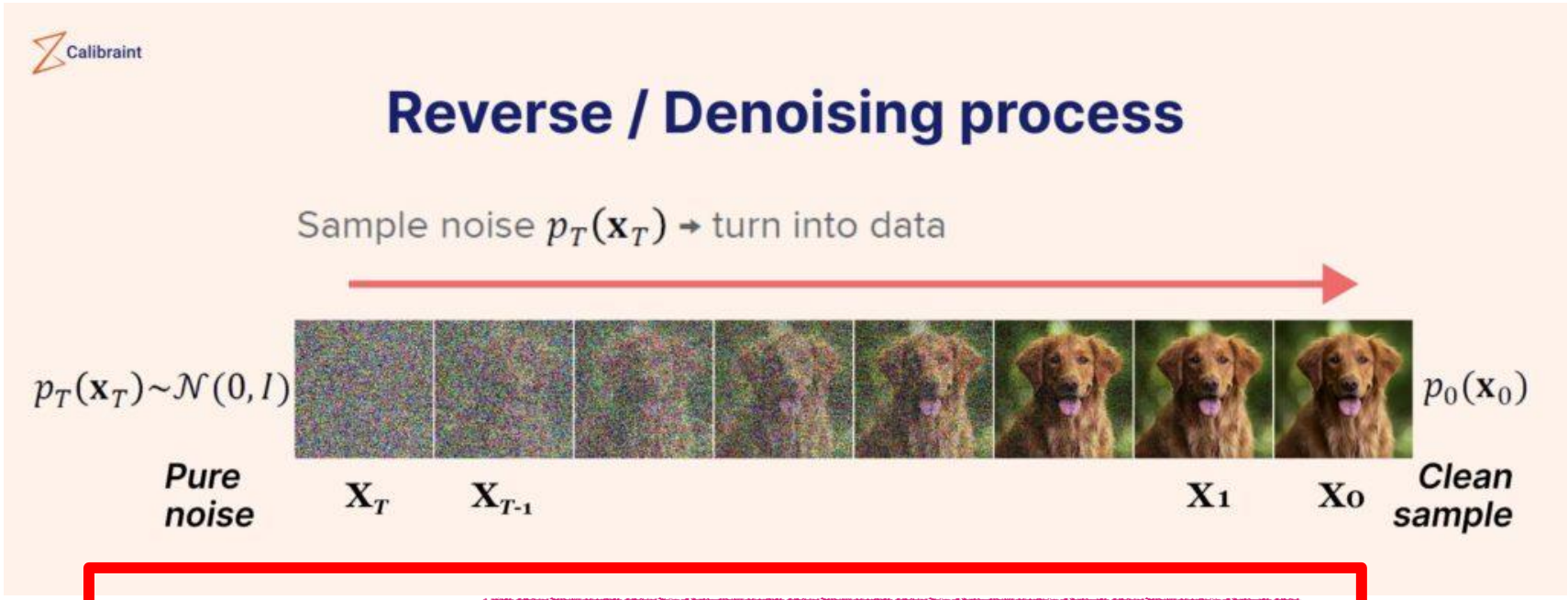
前向过程：加噪 （样本➡白噪声）





反向过程：去噪（白噪声➡样本）

如何训练扩散模型?



数据:
从数据分布 $p^{pre}(x)$
抽样的许多 x_i

$$x_{t+1} = x_t + \Delta t \times f(t, x_t; \theta) + \sqrt{\Delta t} \times \sigma(t)\epsilon$$

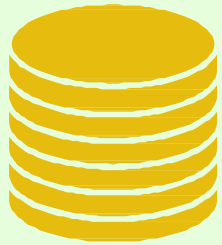
x_t → x_{t+1}

建模为连续SDE的去噪过程:

$$d\,x_t = f(t, x_t, \theta) + \sigma\, dw_t$$

目标:
从数据中学习参数 θ, 从而 $x_T \sim p^{pre}$

最终的优化目标：奖励函数



数据：$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$

标签：$r(x) = \mathrm{E}[y|x]$

奖励函数 $\hat{r}(x)$

艺术效果分数
(Aesthetic Score)

6.4

奖励函数 $\hat{r}(x)$

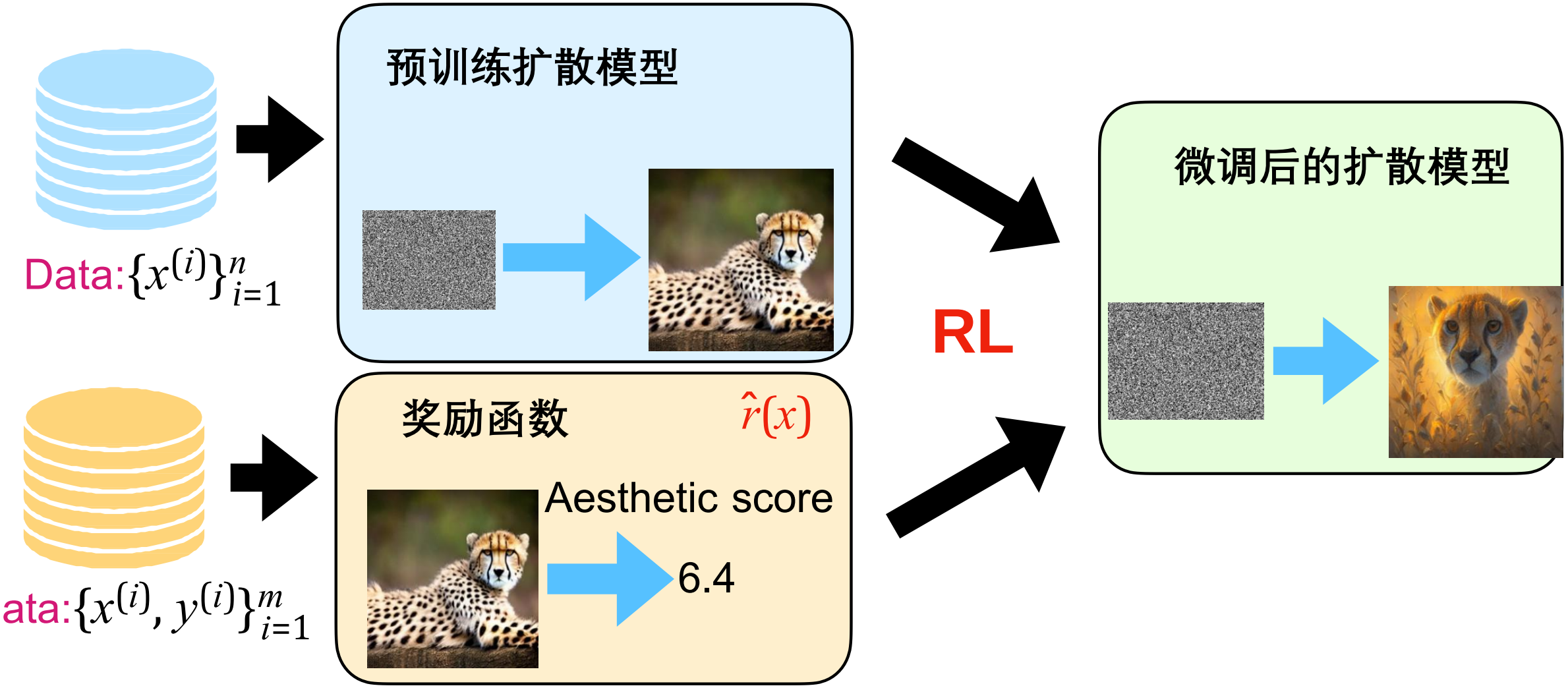
ENFORMER

BORZAI

RNA-seq

ATAC-seq

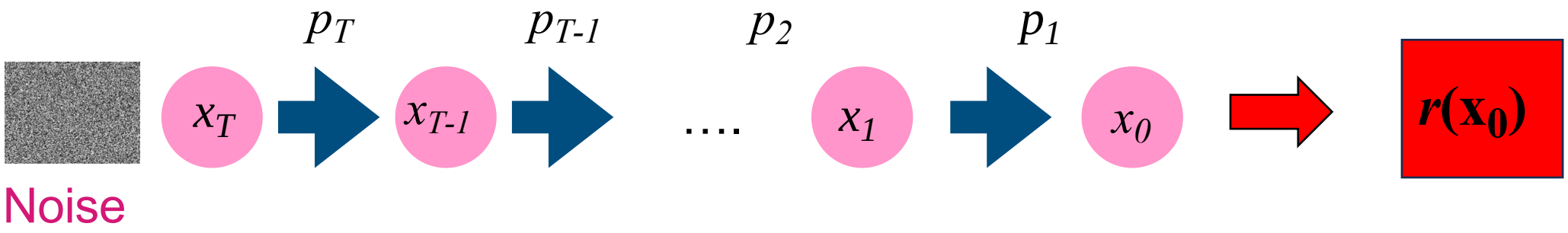
微调扩散模型 —— 整体框架


预训练扩散模型
Data:$\{x^{(i)}\}_{i=1}^{n}$
奖励函数
$\hat{r}(x)$
Aesthetic score
6.4
Data:$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$
RL
微调后的扩散模型

2.基于强化学习的微调算法

算法介绍

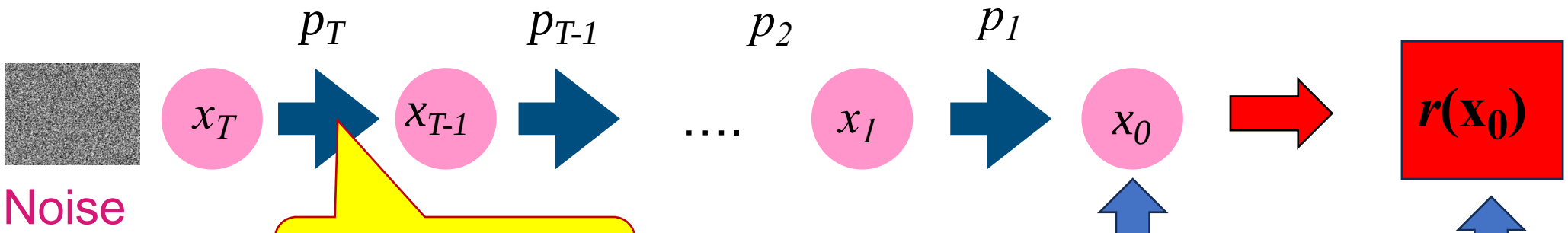
扩散模型的微调 —— 强化学习(RL)视角



如何通过马尔科夫决策过程(MDP)建模去噪过程?

1. MDP中的状态空间S 与动作空间A 都对应于样本空间X。
2. 将MDP中的状态转移概率建模为一个恒等映射: $\mathrm{P}_t(s_{t+1}|s_t, a_t) = \delta(s_{t+1} = a_t)$.
3. 奖励 r_t 只会在最后一步获得（即$r(x_0)$），其他步均为0.

扩散模型的微调 —— 强化学习(RL)视角



Solve RL problem:

$$\max_{\theta} J(\theta) = E_{P^\theta}[r(x_0)]$$

其中，P^θ: $x_{T-1} \sim p_T(x_T), x_{T-2} \sim p_{T-1}(x_{T-1}) \ldots,$
$x_0 \sim p_1(x_1)$

去噪的每一步均以θ(待学的参数)参数化

Images

比如：艺术效果分数
(Aesthetic Score)

6.4

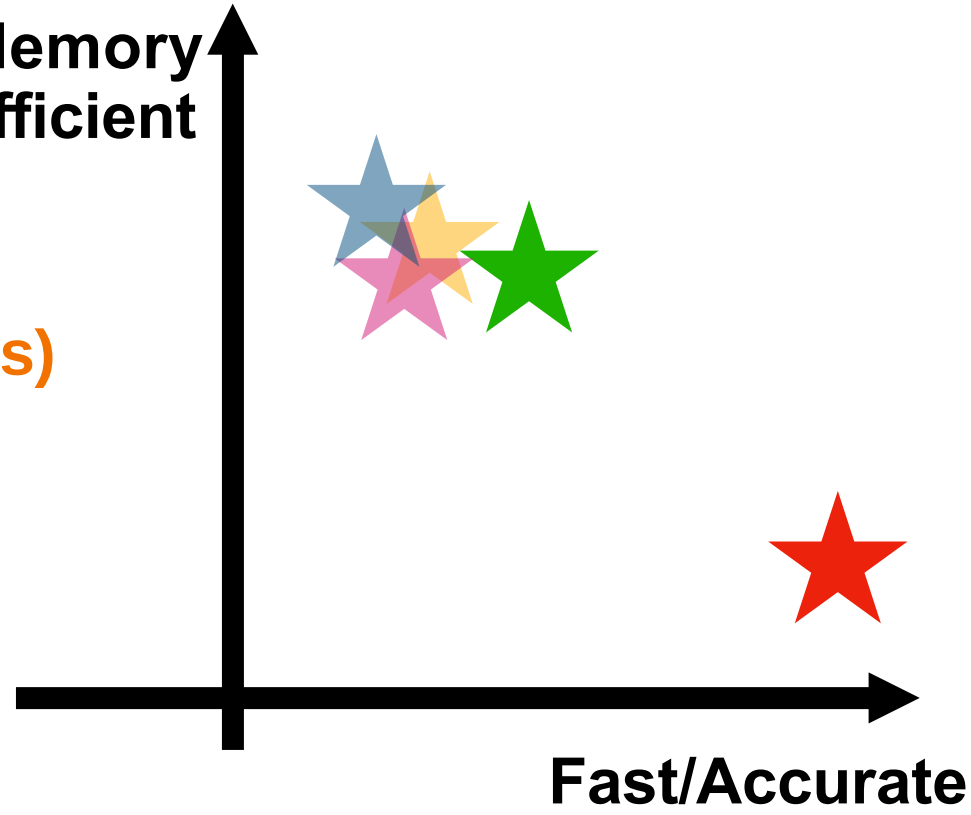
2.1 奖励函数可访问

包含奖励函数可导或不可导的情形

如何求解RL?

我们可以应用任何off-the-shelf RL算法来进行优化!

- **Neural SDE** [Clark et al. 23, Uehara et al. 24]
- **Policy gradient (PPO)** [Black et al. 23; Fan et al. 23]
- **Soft-Q-learning (Training loss in Gflownets)**
- **Advantage weighted learning**
- **Classifier-based guidance (plug-in-play)**



2.1.1 PPO (近端策略优化)

- 优化目标：$\max_{\theta} J(\theta) = E_{P^{\theta}}[r(x_0)]$。经典的策略梯度(policy gradient)为

$$\nabla_{\theta} J(\theta) = E\left[\sum_{t=0}^{T} \nabla_{\theta} \log p_{\theta}(x_{t-1}|x_t) \cdot r(x_0)\right]$$

- 若使用旧参数(θ_{old})收集的轨迹时(Importance sampling)：

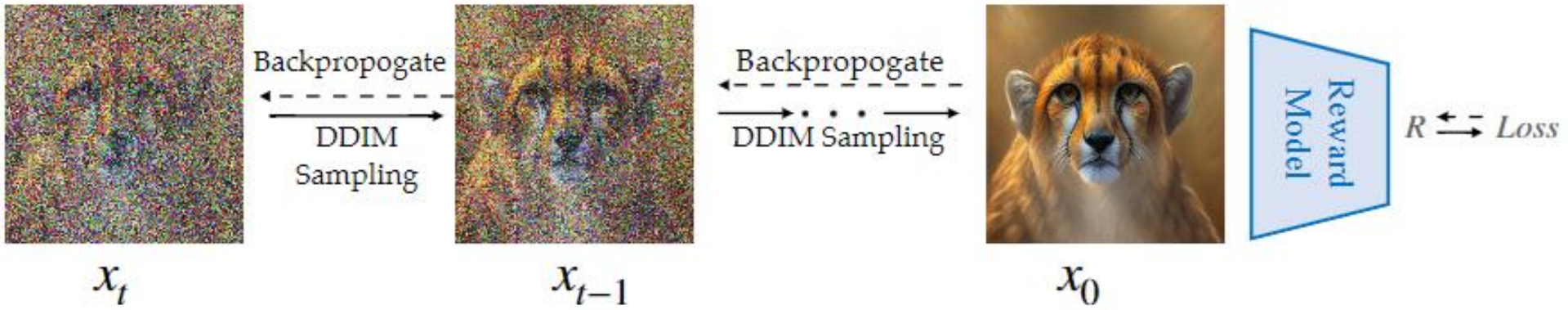
$$\nabla_{\theta} J(\theta) = E\left[\sum_{t=0}^{T} \frac{\mathrm{p}_{\theta}(x_{t-1}|x_t)}{p_{\theta_{old}}(x_{t-1}|x_t)} \nabla_{\theta} \log p_{\theta}(x_{t-1}|x_t) \cdot r(x_0)\right]$$

PPO使用 $(1-\epsilon, 1+\epsilon)$ 来clip这一项，以保证conservative updates

Training Diffusion Models with Reinforcement Learning, 2023

2.1.2 Direct backpropagation (直接优化)

- 直接使用backpropagation优化 $\max_{\theta} J(\theta) = E_{p^{\theta}}[r(x_0)]$，每一个去噪步均得到更新
- 优点：直接，高效。
- 缺点： 奖励函数必须可微
- 可引入KL熵以提升算法表现，防止reward overoptimization



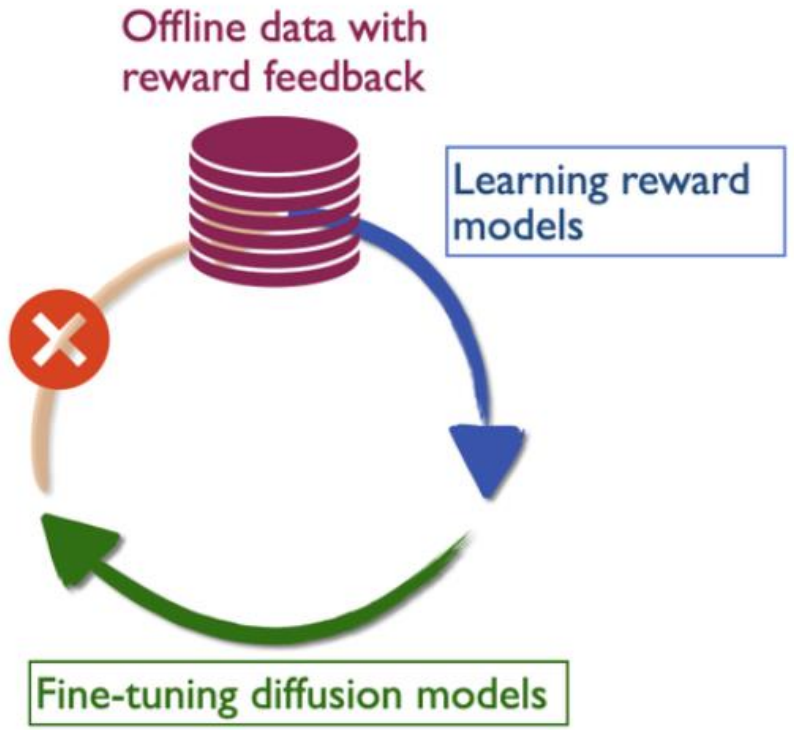
Aligning Text-To-Image Diffusion Models With Reward Backpropagation, 2023

2.2 奖励函数未知

包含使用离线数据或在线与环境交互以获得数据的情形

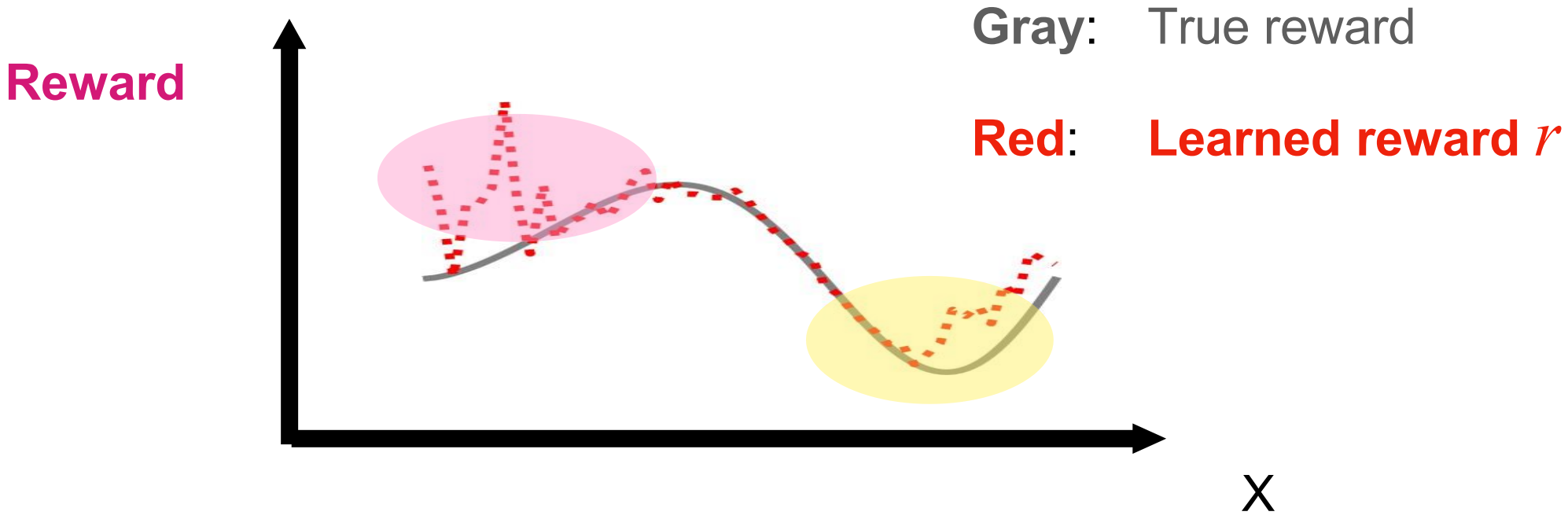
2.2.1 使用离线数据训练奖励函数

- 需要从有限甚至不够代表性的数据中训练奖励函数
- 不允许获得新的反馈数据



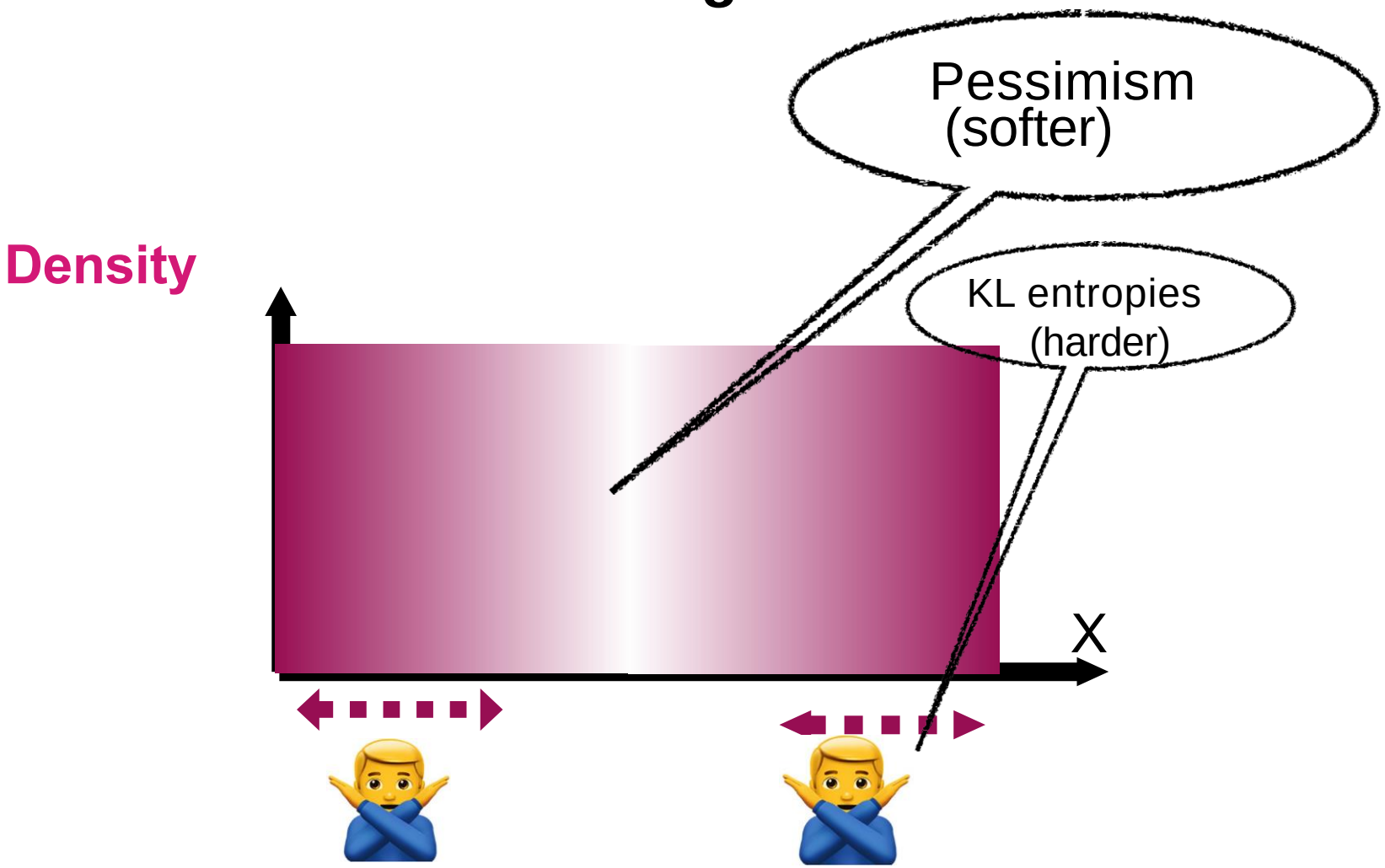
2.2.1 使用离线数据训练奖励函数

- **挑战：通过有限数据训练的奖励函数并不准确**



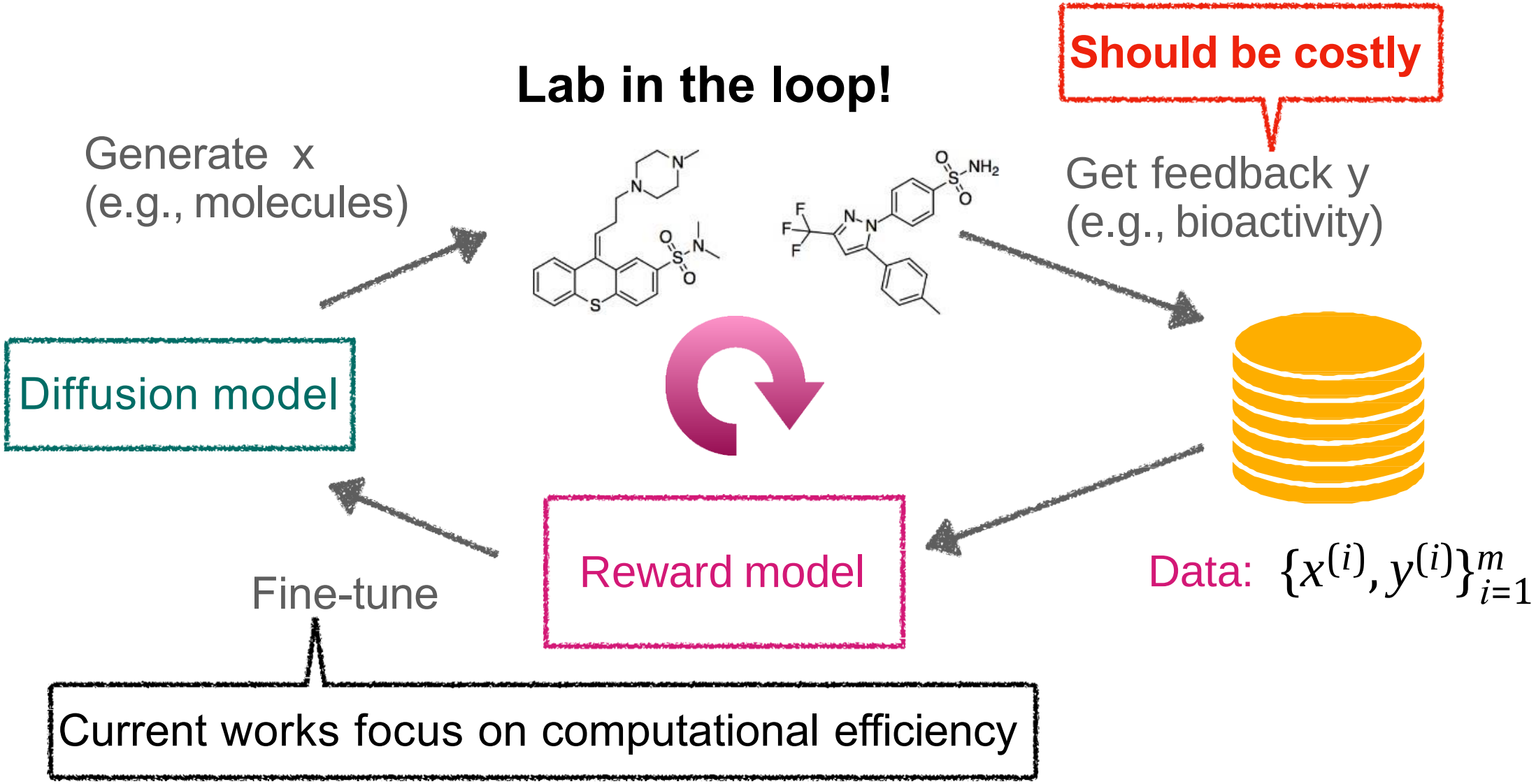
Outliers使得一些预测非常不准确!

BRAID: douBly conseRvAtive fIne-tuning Diffusion models



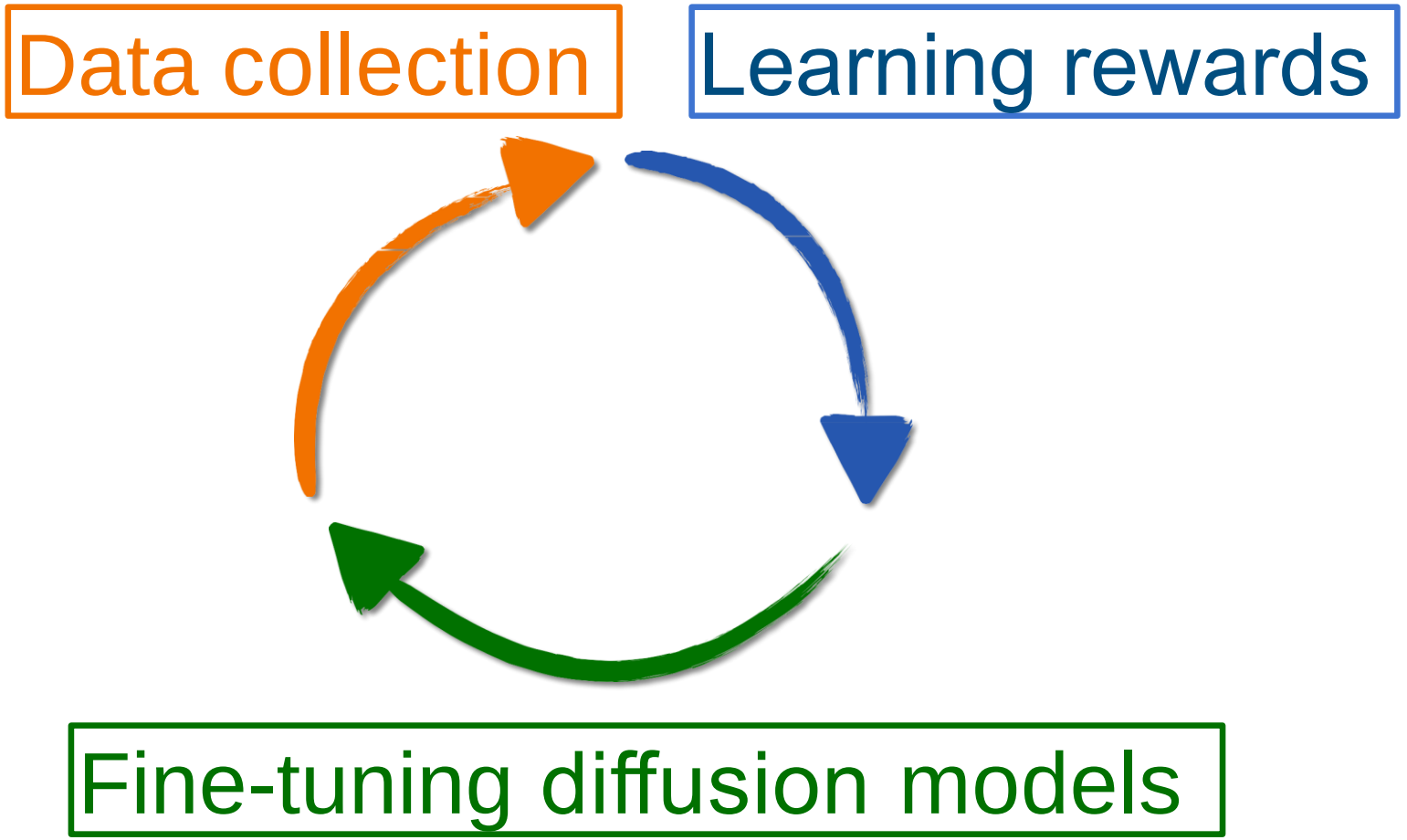
Bridging Model-Based Optimization and Generative Modeling via Conservative Fine-Tuning of Diffusion Models, 2024

2.2.1 在线获取数据并更新模型(online)



SEIKO (OptimiStic finE-tuning of dIffusion with KL cOnstraint)

可以动态结合三个环节的全新框架！



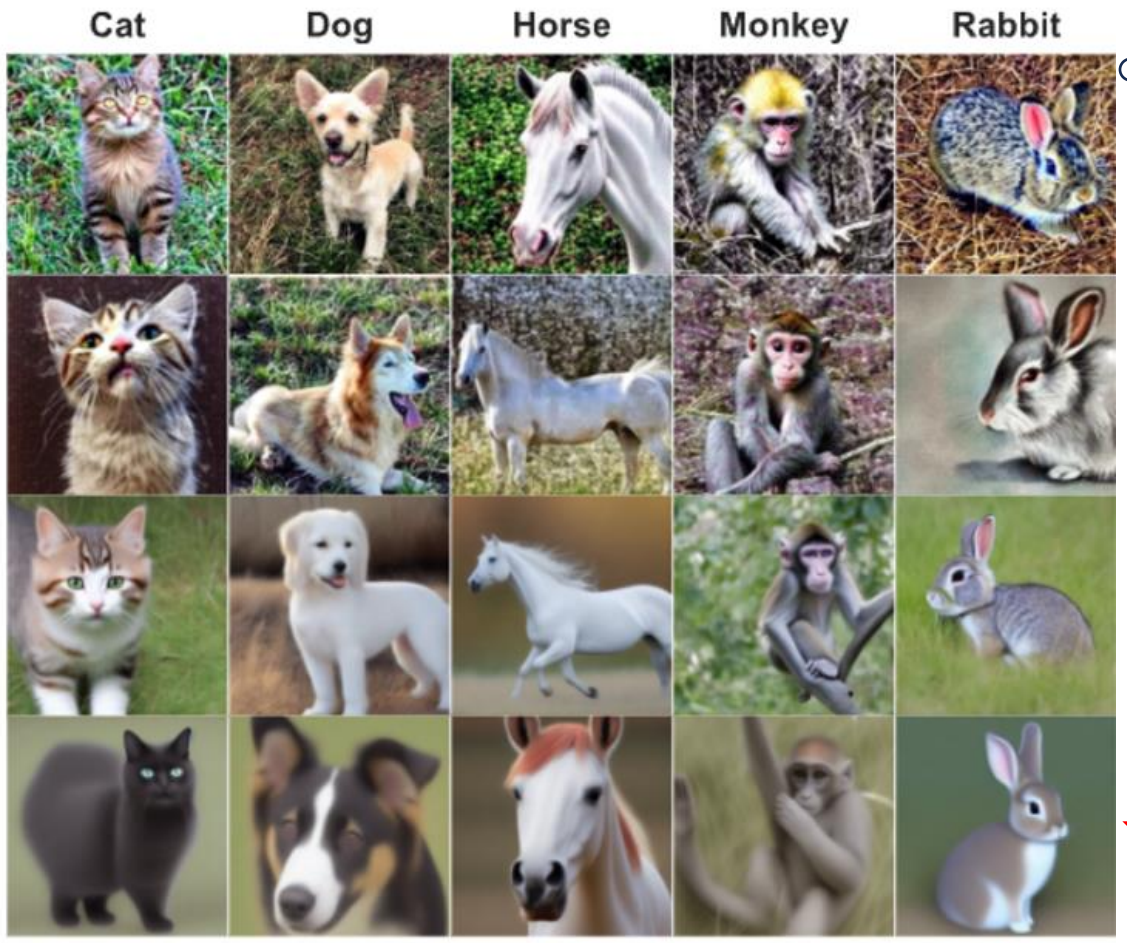
Feedback Efficient Online Fine-Tuning of Diffusion Models, 2024

3.通过微调实现条件生成

相比常见的classifier-based guidance或classifier-free guidance，更加节省样本

微调：控制条件生成

增加一个新的条件(condition)



均由一个模型生成

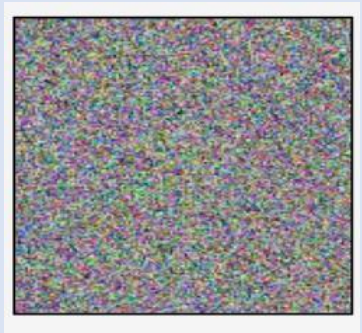
任意的文件大小！

Adding Conditional Control to Diffusion Models with Reinforcement Learning, 2024

方法

Pretrained conditional diffusion: $p(x|c)$

c: prompt



Data: $\{x_i, c_i, y_i\}$

y: compressibility

Train a classifier

$p^{\diamond}(y|x,c)$

RL

Intuition: 使生成的样本 x 对应“正确的” y

Objective func:

$\gamma \log p(y^{\diamond}|x_T, c) - KL(p||p^{pre})$

Guidance “strength”

理论解释

微调

目标函数:

$$\gamma \log p(y|x_T, c) - KL(p||p^{pre})$$

微调后的
条件概率$p(x|y,c)$

$$p(x|y,c) \propto \left(p^{\diamond}(y|x,c)\right)^{\gamma} \cdot p^{pre}(x|c)$$

同时也是classifiered-based guidance的目标分布!

4.总结

结语

- 文章详细解释了如何将微调扩散模型以最大化下游奖励函数形式化为马尔可夫决策过程（MDPs）中的强化学习问题
- 我们详细阐述了各种基于强化学习的算法，如PPO。
- 根据奖励反馈的获取方式对不同场景进行分类并推荐合适的算法。
- 文章深入讨论了与其他相关内容的联系，包括classifier guidance、Gflownets、以及MCMC。请查看原文。

Thank you!