训练扩散模型与强化学习

凯文-布莱克[*1] **迈克尔-詹纳**[*1] **杜一伦**[2] **伊利亚-科斯特里科夫**[1] **Sergey Levine**[(1)(1)]
加州大学伯克利分校 [2]麻省理工学院
{kvablack, janner, kostrikov, sergey.levine}@berkeley.edu yilundu@mit.edu

摘要

扩散模型是一类灵活的生成模型，其训练方法近似于对数似然目标。然而，扩散模型的大多数应用案例并不关注似然，而是关注下游目标，如人类感知的图像质量或药物效果。在本文中，我们研究了针对此类目标直接优化扩散模型的强化学习方法。我们描述了如何将去噪作为一个多步骤决策问题来实现一类策略梯度算法，我们将 其称为去噪扩散策略优化（DDPO），它比其他奖励加权似然方法更有效。根据经验，DDPO 可以调整文本到图像的扩散模型，以适应难以通过提示表达的目标（如图像压缩性）和来自人类反馈的目标（如审美质量）。最后，我们展示了 DDPO 可以利用视觉语言模型的反馈改善提示-图像配准，而无需额外的数据收集或人工注释。项目网站：http://rl-diffusion.github.io。

1 导言

扩散概率模型（Sohl-Dickstein 等人，2015 年）最近已成为连续领域生成建模的事实标准。扩散模型能灵活地表示复杂的高维分布，因此在图像和视频合成（Ramesh 等人，2021 年；Saharia 等人，2022 年；Ho 等人，2022 年）、药物和材料设计（Xu 等人，2021 年；Xie 等人，2021 年；Schneuing 等人，2022 年）和连续控制（Janner 等人，2022 年；Wang 等人，2022 年；Hansen-Estruch 等人，2023 年）等应用中得到广泛采用。扩散模型背后的关键思想是通过应用顺序去噪过程，将简单的先验分布迭代转换为目标分布。这一过程通常被视为最大似然估计问题，其目标是训练数据对数似然的变分下限。

然而，大多数扩散模型的使用案例并不直接关注似然性，而是下游目标，如人类感知的图像质量或药物疗效。在本文中，我们考虑的问题是训练扩散模型以直接满足这些目标，而不是匹配数据分布。这个问题极具挑战性，因为扩散模型的精确似然计算难以实现，这使得许多传统的强化学习（RL）算法难以应用。相反，我们建议将去噪作为一项多步骤决策任务，在每个去噪步骤中使用精确似然值，以取代完整去噪过程中产生的近似似然值。我们提出了一种策略梯度算法，并将其称为去噪扩散策略优化（DDPO），该算法只需使用黑盒奖励函数就能优化下游任务的扩散模型。

我们将算法应用于大型文本到图像扩散模型的微调。我们最初的评估重点是那些难以通过提示指定的任务，如图像压缩性，以及那些来自人类反馈的任务，如审美质量。然而，由于许多感兴趣的奖励函数难以通过编程指定，微调程序通常依赖于大规模的人工标注工作来获得奖励信号（欧阳等人，2022 年）。就文本到图像的扩散而言，我们提出了一种用视觉语言模型（VLM）的反馈来取代这种标记的方法。与语言模型的 RLAIF 微调类似（Bai 等人，2022b），由此产生的程序可使扩散模型适应奖励函数，而这些奖励函数需要在其他情况下进行调整。

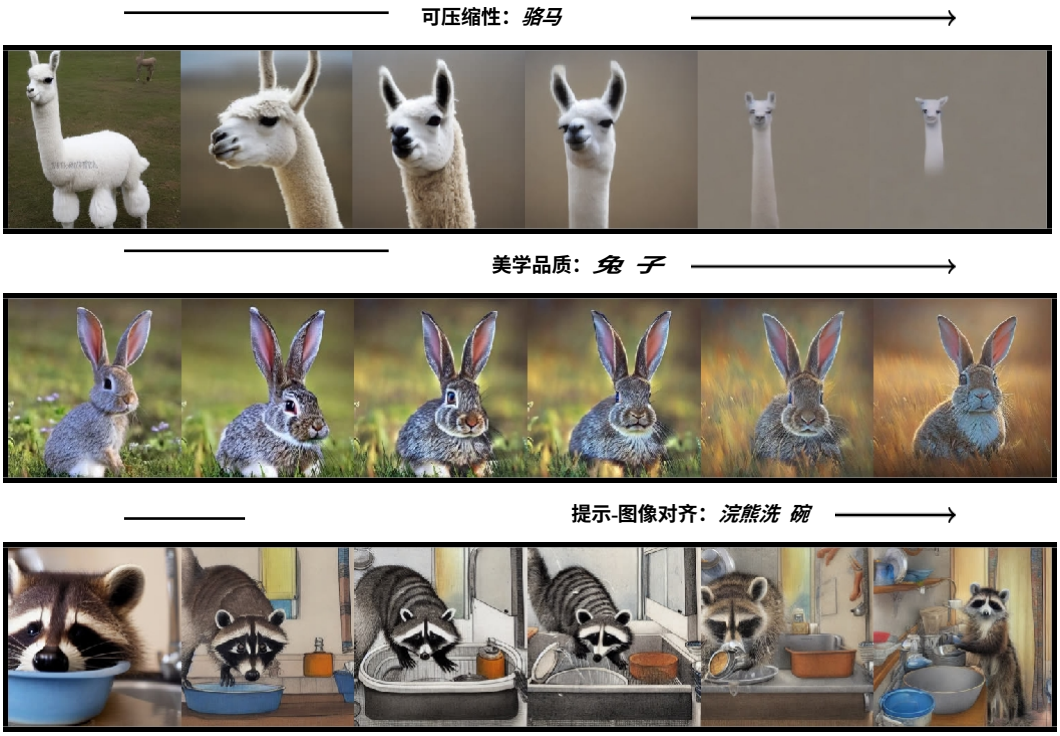

图 1（扩散模型的强化学习）我们提出了一种强化学习算法 DDPO，用于优化扩散模型的下游目标，如可压缩性、美学质量以及由视觉语言模型决定的提示-图像对齐。每行显示的是同一提示和随机种子的样本在训练过程中的进展情况。

额外的人工注释。我们使用这一程序来改进不寻常主题设置组合的提示图像配准。

我们的贡献如下。我们首先介绍了 DDPO 的推导和概念动机。然后，我们记录了用于文本到图像生成的各种奖励函数的设计，包括从简单计算到涉及大型 VLM 的工作流程，并展示了 DDPO 在这些设置中与其他奖励加权似然法相比的有效性。最后，我们展示了微调程序对未见提示的通用能力。

2 相关工作

扩散概率模型。去噪扩散模型（Sohl-Dickstein 等人，2015 年；Ho 等人，2020 年）已成为一类有效的生成模型，适用于图像（Ramesh 等人，2021 年；Saharia 等人，2022 年）、视频（Ho 等人，2022 年；Singer 等人，2022 年）、三维形状（Zhou 等人，2021 年；Zeng 等人，2022 年）等模态、2022 年）、视频（Ho 等人，2022 年；Singer 等人，2022 年）、三维形状（Zhou 等人，2021 年；Zeng 等人，2022 年）和机器人轨迹（Janner 等人，2022 年；Ajay 等人，2022 年；Chi 等人，2023 年）。虽然去噪目标通常是作为似然法的近似值推导出来的，但扩散模型的训练通常会在几个方面偏离最大似然法（Ho 等人，2020 年）。修改目标以更严格地优化似然（Nichol & Dhariwal，2021；Kingma 等人，2021）往往会导致图像质量下降，因为似然并不是视觉质量的忠实代表。在本文中，我们将展示如何直接针对下游目标优化扩散模型。

利用扩散模型进行可控生成。文本到图像扩散模型的最新进展（Ramesh 等人，2021 年；Saharia 等人，2022 年）实现了细粒度的高分辨率图像合成。为了进一步提高扩散模型的可控性和质量，最近的研究方法包括在用户提供的有限数据基础上进行微调（Ruiz 等人，2022 年）、针对新概念优化文本嵌入（Gal 等人，2022 年）、合成模型（Du 等人，2023 年；Liu 等人，2022 年）、2023；Liu 等人，2022）、针对额外输入约束的适配器（Zhang & Agrawala，2023）以及推理时技术，如分类器（Dhariwal & Nichol，2021）和无分类器（Ho & Salimans，2021）引导。

从人类反馈中强化学习。许多研究都在模拟机器人控制（Christiano 等人，2017 年）、游戏（Knox 和 Stone，2008 年）、机器翻译（Nguyen 等人，2017 年）、引文检索（Menick 等人，2022 年）、基于浏览的问题解答（Nakano 等人，2021 年）、摘要（Sciennon 等人，2020 年；Ziegg 等人，2021 年）等环境中使用人类反馈来优化模型、2022 年）、基于浏览的问题解答（Nakano 等人，2021 年）、总结（Stiennon 等人，2020 年；Ziegler 等人，2019 年）、指令跟踪（欧阳等人，2022 年）以及与规范的对齐（Bai 等人，2022a 年）。最近，Lee 等人（2023 年）使用一种基于报酬加权似然最大化的方法，研究了文本到图像扩散模型与人类偏好的配准。在我们的比较中，他们的方法相当于奖励加权回归（RWR）方法的一次迭代。我们的结果表明，DDPO 甚至明显优于加权似然最大化（RWR 式）优化的多次迭代。

作为顺序决策过程的扩散模型。虽然早于扩散模型，但 Bachman 和 Precup（2015 年）同样将数据生成作为一个顺序决策问题，并利用由此产生的框架将强化学习方法应用于图像生成。最近，Fan & Lee（2023 年）提出了一种用于训练扩散模型的策略梯度法。不过，该论文旨在改善数据分布匹配，而不是优化下游目标，因此考虑的唯一奖励函数是类似于 GAN 的判别器。在与我们同时进行的工作中，DPOK（Fan 等人，2023 年）在 Fan & Lee（2023 年）和 Lee 等人（2023 年）的基础上，使用策略梯度算法使文本到图像的扩散模型更好地与人类偏好相匹配。与 Lee 等人（2023 年）一样，DPOK 只考虑了一个基于偏好的奖励函数（Xu 等人，2023 年）；此外，他们的工作研究了 KL 正则化，并主要侧重于为每个提示训练一个不同的扩散模型。相比之下，我们一次对多个提示进行训练（多达 398 个），并展示了对训练集之外的更多提示的泛化。此外，我们还研究了 DDPO 如何应用于基于人类反馈之外的多种奖励功能，包括从 VLM 自动得出的奖励如何改善提示与图像的对齐。我们在附录 C 中提供了与 DPOK 的直接比较。

3 前言

在本节中，我们将简要介绍扩散模型和 RL 问题表述的背景。

3.1 扩散模型

在这项工作中，我们考虑的是条件扩散概率模型（Sohl-Dickstein 等人，2015 年；Ho 等人，2020 年），它表示在样本 $\mathbf{x}_0$和相应上下文 $\mathbf{c}$ 的数据集上的分布 $p(\mathbf{x}_{(0)}\,\mathbf{c})$。该分布被建模为马尔可夫正向过程 $q(\mathbf{x}_{(t)}\,x_{(t)-(1)})$ 的逆过程，该过程会迭代地向数据添加噪声。反转前向过程可以通过训练神经网络 $\boldsymbol{\mu}_{(t)}(\mathbf{x}_{(t)},\ \mathbf{c},\ t)$ 来实现，其目标如下：

$$\text{LDDPM}(\theta) = \mathsf{E}_{(\mathbf{x},\mathbf{c})\sim p(\mathbf{x},\mathbf{c}),\ t\sim\mathcal{U}\{0,T\},\ \mathbf{x}\sim q(\mathbf{x}|\ \mathbf{x})}\left\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_0,\ t) - \mu_\theta(x_t,\ \mathbf{c},\ t)\right\|^2 \tag{1}$$

其中 $\tilde{\boldsymbol{\mu}}$ 是前向过程的后验均值，即 $\mathbf{x}_0$和 $\mathbf{x}_t$的加权平均值。这一目标被认为是最大化数据对数似然的变分下限（Ho 等人，2020 年）。

从扩散模型中取样，首先随机抽取 $\mathbf{x}_T \quad (\mathbf{0},\ \mathbf{I})$，然后按照 $p_\theta(\mathbf{x}_{t-1}\mathbf{x}_{(t)},\ \mathbf{c})$ 的反向过程产生轨迹 $\mathbf{x}_{(T)},\ \mathbf{x}_{T-1},\ .,\ \mathbf{x}_0$，最后得到一个样本 $\mathbf{x}_0$。采样过程不仅取决于预测因子 $\boldsymbol{\mu}_\theta$，还取决于采样器的选择。大多数流行的采样器（Ho 等人，2020 年；Song 等人，2021 年）使用各向同性的高斯反向过程，其方差与时间步长相关：

$$p_\theta(\mathbf{x}_{t-1}|\ \mathbf{x}_{(t)},\ \mathbf{c}) = \mathcal{N}\ (\mathbf{x}_{(t)-1}|\ \boldsymbol{\mu}_{(\theta)\ (x_{(t)},\ \mathbf{c},\ t)},\ \sigma_{(t)}^2\mathbf{I}). \tag{2}$$

3.2 马尔可夫决策过程与强化学习

马尔可夫决策过程（Markov decision process，MDP）是顺序决策问题的形式化。马尔可夫决策过程由一个元组 $(\mathcal{S}, \mathcal{A}, \rho_0, P, R)$ 定义，其中 $(\mathcal{S}, \mathcal{A}, \rho_0, P, R)$是状态空间，$\mathcal{S}$ 是行动空间，$\mathcal{A}$是初始状态分布，P 是转换内核，R 是奖励函数。在每个时间步 t，代理观察一个状态 $\mathbf{s}_t \in \mathcal{S}$，采取一个行动 $\mathbf{a}_t \in \mathcal{A}$，获得奖励 $R(\mathbf{s}_{(t)},\ \mathbf{a}_{(t)})$ 并过渡到一个新状态 $\mathbf{s}_{(t)\,(+1)}\sim P\,(\mathbf{s}_{(t)\,(+1)}|\ \mathbf{s}_{(t)},\ \mathbf{a}_{(t)})$。代理根据策略 $\pi(\mathbf{a}|\ \mathbf{s})$ 行事。

代理在 MDP 中行动时，会产生轨迹，即状态和行动的序列 $\tau = (\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T)$。强化学习（RL）的目标是让代理最大化 $J_{\mathrm{RL}}(\pi)$ -- 即从其策略中采样的轨迹的预期累积奖励：

$$J_{\mathrm{RL}}(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left(\sum_{t=0}^{T} R(\mathbf{s}_t, \mathbf{a}_t) \right).$$

4 扩散模型的强化学习训练

现在我们介绍如何使用 RL 算法来训练扩散模型。我们将介绍两类方法，并说明每种方法都对应于去噪过程到 MDP 框架的不同映射。

4.1 问题陈述

我们假设有一个预先存在的扩散模型，该模型可以是预训练的，也可以是随机初始化的。假设采样器是固定的，扩散模型会产生一个样本分布 $p_\theta(\mathbf{x}_0|\mathbf{c})$。去噪扩散 RL 的目标是最大化定义在样本和上下文上的奖励信号 r：

$$J_{\mathrm{DDRL}}(\theta) = \mathbb{E}_{\mathbf{c} \sim p(\mathbf{c}),\, \mathbf{x}_0 \sim p_\theta(\mathbf{x}_0|\mathbf{c})} [r(\mathbf{x}_0, \mathbf{c})]$$

对于我们选择的某种上下文分布 $p(\mathbf{c})$。

4.2 奖励加权回归

为了在尽量不改变标准扩散模型训练的情况下优化 J_{DDRL}，我们可以使用去噪损失 L_{DDPM}（等式 1），但训练数据 $\mathbf{x}_0 \sim p_\theta(\mathbf{x}_0|\mathbf{c})$ 以及取决于奖励 $r(\mathbf{x}_0, \mathbf{c})$ 的附加加权。Lee 等人（2023 年）描述了单轮版本的扩散模型程序，但一般来说，这种方法可以进行多轮交替采样和训练，从而形成在线 RL 方法。我们将这一类算法称为奖励加权回归（RWR）（Peters & Schaal，2007 年）。

标准加权方案使用指数奖励来确保非负性、

$$w_{\mathrm{RWR}}(\mathbf{x}_0, \mathbf{c}) = \frac{1}{Z} \exp \beta r(\mathbf{x}_0, \mathbf{c}) \text{ 、}$$

其中，β 是反温度，Z 是归一化常数。我们还考虑了一种使用二进制权重的简化加权方案、

$$w_{\mathrm{sparse}}(\mathbf{x}_0, \mathbf{c}) = \mathbf{1}\, r(\mathbf{x}_0, \mathbf{c}) \geq C \text{ 、}$$

其中 C 是奖励阈值，决定哪些样本用于训练。从监督学习的角度来看，这相当于对来自模型的训练数据进行反复过滤微调。

在 RL 形式中，RWR 程序对应于以下一步 MDP：

$$\mathbf{s} \triangleq \mathbf{c} \qquad \mathbf{a} \triangleq \mathbf{x}_0 \qquad \pi(\mathbf{a}|\mathbf{s}) \triangleq p_\theta(\mathbf{x}_0|\mathbf{c}) \qquad \rho_0(\mathbf{s}) \triangleq p(\mathbf{c}) \qquad R(\mathbf{s}, \mathbf{a}) \triangleq r(\mathbf{x}_0, \mathbf{c})$$

的过渡核 P 会立即导致一个吸收终止状态。因此，在此 MDP 中，最大化 $J_{\mathrm{DDRL}}(\theta)$ 等于最大化 RL 目标 $J_{\mathrm{RL}}(\pi_\theta)$。

从 RL 文献中得知，用 w_{RWR} 加权对数似然目标可近似最大化 $J_{\mathrm{RL}}(\pi_\theta)$，但要受到 KL 分歧对 π 的约束（Nair 等人，2020 年）。然而，L_{DDPM}（等式 1）并不涉及精确的对数似然，而是作为 $\log p_\theta(\mathbf{x}_0|\mathbf{c})$ 的变异约束推导出来的。因此，将 RWR 程序应用于扩散模型训练在理论上是站不住脚的，只能非常近似地优化 J_{DDRL}。

4.3 去噪扩散策略优化

RWR 依赖于近似对数似然，因为它忽略了去噪过程的连续性，只使用最终样本 $\mathbf{x}_0$。在本节中，我们将展示如何将去噪过程重构为*多步骤* MDP，从而利用策略梯度直接优化 J_{DDRL}。

估计器。这沿用了 Fan & Lee（2023）的推导，他们证明了自己的方法与策略梯度算法之间的等价性，其中奖励是一个类似于 GAN 的判别器。我们提出了一个具有任意奖励函数的通用框架，其动机是我们希望优化任意下游目标（第 5 节）。我们将这一类算法称为去噪扩散策略优化（DDPO），并介绍了基于特定梯度估计器的两种变体。

作为多步骤 MDP 的去噪。我们将迭代去噪过程映射为以下 MDP：

$$\mathbf{s}_t \triangleq (\mathbf{c},\ t,\ \mathbf{x}_t) \qquad \pi(\mathbf{a}_t|\ \mathbf{s}_t) \triangleq p_\theta(\mathbf{x}_{t-1}|\ \mathbf{x}_t,\ \mathbf{c}) \qquad P(\mathbf{s}_{t+1}|\ \mathbf{s}_t,\ \mathbf{a}_t) \triangleq \delta_{\mathbf{c}},\ \delta_{t-1},\ \delta_{\mathbf{x}_{t-1}}$$

$$\mathbf{a}_t \triangleq \mathbf{x}_{t-1} \qquad \rho_0(\mathbf{s}_0) \triangleq p(\mathbf{c}),\ \delta_0,\ \mathcal{N}(\mathbf{0}, \mathbf{I}) \qquad R(\mathbf{s}_t, \mathbf{a}_t) \triangleq \begin{cases} r(\mathbf{x}_0, \mathbf{c}) & \text{if } t = 0 \\ 0 & \text{否则} \end{cases}$$

轨迹由 T 个时间步组成，T 个时间步之后，P 将导致一个终止状态。每个轨迹的累积奖励等于 $r(\mathbf{x}_0, \mathbf{c})$，因此在此 MDP 中，最大化 $J_{\text{DDRL}}(\theta)$等于最大化 $J_{\text{RL}}(\pi)$。

这种表述的好处在于，如果我们使用标准采样器，将 $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$ 参数化为等式 2 中的参数，那么策略 π 就会变成各向同性的高斯分布，而不是 RWR 表述中任意复杂的分布 $p_\theta(\mathbf{x}_0|\mathbf{c})$。通过这种简化，可以评估与扩散模型参数相关的精确对数似然及其梯度。

政策梯度估计。利用似然和似然梯度，我们可以直接对$\nabla_\theta J_{\text{DDRL}}$进行蒙特卡罗估计。与 RWR 类似，DDPO 交替收集去噪轨迹
$\{\mathbf{x}_T, \mathbf{x}_{T-1}, \ldots, \mathbf{x}_0\}$，并通过梯度下降更新参数。

DDPO 的第一种变体，我们称之为 DDPO_{SF}，使用得分函数政策梯度估计法，也称为似然比法或 REINFORCE（Williams，1992；Mohamed 等，2020）：

$$\nabla_\theta J_{\text{DDRL}} = \mathbb{E}\left[\sum_{t=0}^{T} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})\, r(\mathbf{x}_0, \mathbf{c})\right] \tag{DDPO_{SF}}$$

其中，期望值取自当前参数 θ 产生的去噪轨迹。

然而，由于梯度必须使用当前参数生成的数据来计算，因此这种估计器只允许在每一轮数据收集中进行一步优化。为了进行多步优化，我们可以使用重要性采样估计器（Kakade & Langford, 2002）：

$$\nabla_\theta J_{\text{DDRL}} = \mathbb{E}\left[\sum_{t=0}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\ \mathbf{x}_t,\ \mathbf{c})}{p_{\theta_{\text{old}}}(\mathbf{x}_{t-1}|\ \mathbf{x}_t,\ \mathbf{c})} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1}|\ \mathbf{x}_t,\ \mathbf{c})\, r(\mathbf{x}_0,\ \mathbf{c})\right] \tag{DDPO_{IS}}$$

其中，期望值取自参数 θ_{old}生成的去噪轨迹。如果 p_θ与 $p_{\theta_{\text{old}}}$偏差过大，这个估计值就会变得不准确，这时可以使用信任区域（Schulman 等人，2015 年）来限制更新的大小。在实践中，我们通过剪裁来实现信任区域，就像近端策略优化一样（Schulman 等人，2017 年）。

5 文本到图像扩散的奖励函数

在这项工作中，我们对文本到图像扩散的方法进行了评估。文本到图像的扩散是强化学习的一个重要测试环境，因为它可以使用大量预训练模型，还可以使用多种多样、视觉上有趣的奖励函数。在本节中，我们将概述奖励函数的选择。我们研究了一系列复杂程度不同的奖励函数，其中既有可以直接指定和评估的函数，也有能够捕捉真实世界下游任务深度的函数。

5.1 可压缩性和不可压缩性

文本到图像扩散模型的功能受到训练分布中文本和图像共现的限制。例如，图像很少标注文件大小，因此无法通过提示指定所需的文件大小。这种限制使得基于文件大小的奖励函数成为一个方便的案例研究：它们计算简单，但无法通过可能性最大化和提示工程的传统方法进行控制。

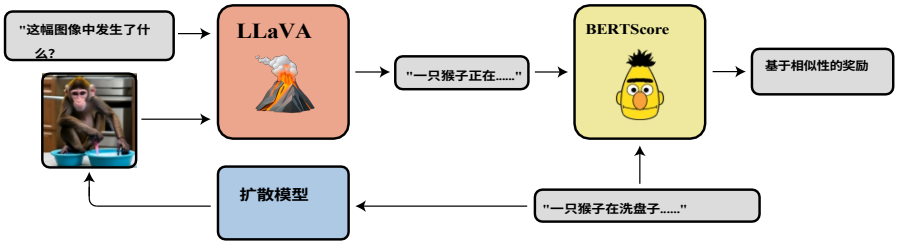

图 2（VLM 奖励函数）基于 VLM 的提示-图像配准奖励函数图解。LLaVA（Liu 等人，2023 年）提供了对生成图像的简短描述；奖励是通过 BERTScore（Zhang 等人，2020 年）测量的该描述与原始提示之间的相似度。

我们将扩散模型样本的分辨率固定为 512x512，因此文件大小完全由图像的可压缩性决定。我们根据文件大小定义了两个任务：可压缩性（JPEG 压缩后图像的文件大小最小化）和不可压缩性（同一指标最大化）。

5.2 美学质量

为了捕捉对人类用户有用的奖励函数，我们定义了一项基于感知美学质量的任务。我们使用 LAION 美学预测器（Schuhmann，2022 年），该预测器是在 176,000 次人类图像评分的基础上训练出来的。该预测器是在 CLIP 嵌入（Radford 等人，2021 年）的基础上以线性模型的形式实现的。注释范围在 1 到 10 之间，评分最高的图片大多包含艺术品。由于美学质量预测器是根据人类的判断进行训练的，因此这项任务构成了从人类反馈中进行强化学习（Ouyang 等人，2022；Christiano 等人，2017；Ziegler 等人，2019）。

5.3 利用视觉语言模型自动对齐提示信息

用于训练文本到图像模型的一个非常通用的奖励函数是提示图像对齐。然而，指定一个能捕捉通用提示配准的奖励是很困难的，通常需要大规模的人工标注工作。我们建议使用现有的 VLM 来替代额外的人工标注。这一设计受到了 RLAIF（Bai 等人，2022b）近期工作的启发，在该工作中，语言模型通过自身反馈得到改进。

我们使用最先进的 VLM LLaVA（Liu 等人，2023 年）来描述图像。微调奖励是 BERTScore（Zhang 等人，2020 年）召回度量，这是一种语义相似性度量，使用提示语作为参考句，VLM 描述作为候选句。更忠实地包含提示语所有细节的样本会获得更高的奖励，只要这些视觉细节对 VLM 来说是可读的。

在图 2 中，我们展示了一个简单的问题："这幅*图像中发生了什么？*虽然这反映的是提示-图像配准的一般任务，但原则上任何问题都可以用来为特定用例指定复杂或难以定义的奖励函数。我们甚至可以使用语言模型来自动生成候选问题，并根据提示对回答进行评估。这一框架提供了一个灵活的界面，奖励函数的复杂性仅受相关视觉和语言模型能力的限制。

6 实验评估

我们实验的目的是评估 RL 算法在微调扩散模型以符合各种用户指定目标方面的有效性。在考察了一般方法的可行性后，我们重点研究了以下问题：

1. DDPO 的变体与 RWR 相比以及相互之间的比较如何？
2. VLM 能否优化难以手动指定的奖励？
3. RL 微调的效果是否会扩展到微调过程中未出现的提示？

图 3（DDPO 样本）定性描述了 RL 微调对不同奖励函数的影响。DDPO 将自然图像转换为风格化的艺术品，以最大限度地提高审美质量；去除背景内容并应用前景平滑处理，以最大限度地提高可压缩性；添加高频噪声，以最大限度地提高不可压缩性。

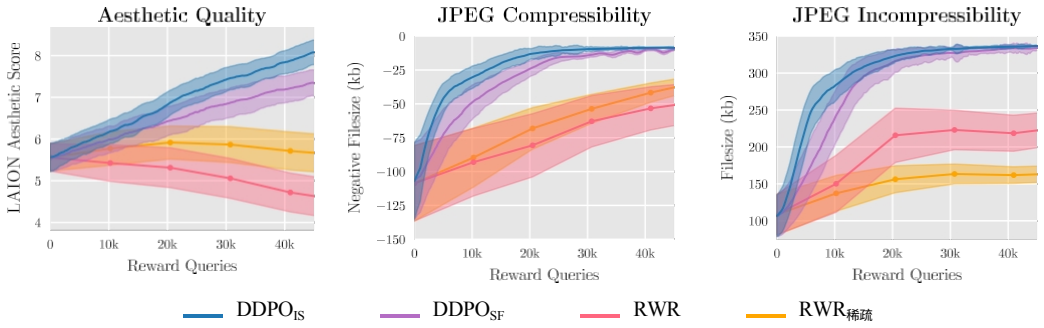

图 4（微调有效性）不同 RL 算法在三种奖励函数上的相对有效性。我们发现，与 RWR 变体相比，策略梯度变体（表示为 DDPO）是更有效的优化器。

6.1 算法比较

我们首先在可压缩性、不可压缩性和美学质量任务上对所有方法进行评估，因为这些任务将 RL 方法的有效性与 VLM 奖励函数相关的考虑因素隔离开来。我们使用 Stable Diffusion v1.4（Rombach 等人，2022 年）作为所有实验的基础模型。可压缩性和不可压缩性提示从 ImageNet-1000 类别（Deng 等人，2009 年）中的所有 398 种动物中统一采样。审美质量提示则从较小的 45 种常见动物中统一取样。

如图 3 所示，DDPO 只需指定奖励函数，无需进一步的数据整理，就能有效地调整预训练模型。优化每种奖励的策略并不复杂，例如，为了最大限度地提高 LAION 预测的美学质量，DDPO 会将生成自然图像的模型转换为生成艺术图画的模型。为了最大限度地提高可压缩性，DDPO 会去除背景并对剩余部分进行平滑处理。为了最大限度地提高不可压缩性，DDPO 会发现 JPEG 压缩算法难以编码的人工痕迹，如高频噪声和锐利边缘。附录 G 中提供了 RWR 的样本，以供比较。

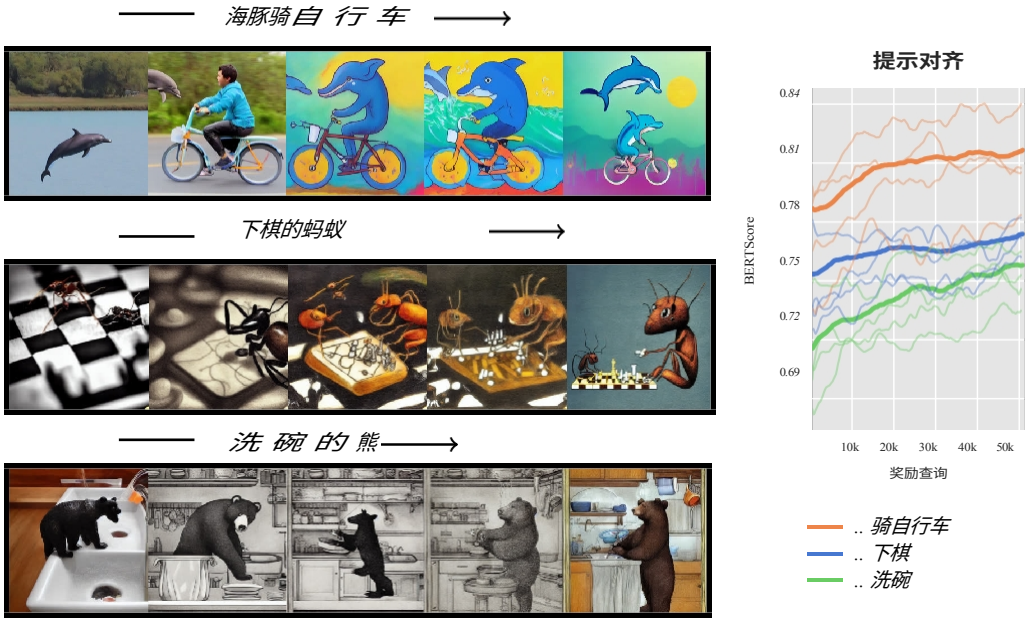

图 5（提示对齐）(L) 同一提示和随机种子的样本在训练过程中的变化。图像明显更忠实于提示。样本还采用了卡通风格，我们假设这是因为在预训练分布中，提示更可能被描绘成插图，而不是逼真的照片。**(R)** 提示对齐的定量改进。每条粗线是一项活动的平均得分，而细线则是随机选取的几个单个提示的平均得分。

我们在图 4 中对所有方法进行了定量比较。我们绘制了获得的奖励与奖励函数查询次数的函数关系图，因为在许多实际应用中，奖励评估已成为限制因素。在所有任务中，DDPO 都比 RWR 有明显优势，这表明将去噪过程表述为多步 MDP 并直接估计策略梯度比优化对数似然的奖励加权变分约束更有效。在 DDPO 类别中，重要性采样估计器的性能略优于分数函数估计器，这可能是由于优化步骤的数量增加了。在 RWR 类别中，加权方案的性能不相上下，稀疏加权方案因其简单性和减少的资源需求，在这些任务中更受欢迎。

6.2 自动提示对齐

接下来，我们将评估 VLM 与 DDPO 结合使用的能力，以自动改进预训练模型的图像提示对齐，而无需额外的人工标注。在本实验中，我们将重点关注 DDPO_{IS}，因为我们在第 6.1 节中发现它是最有效的算法。这项任务的提示语均为 "*a(n) [animal] [activity]* "，其中动物来自第 6.1 节中使用的 45 种常见动物列表，活动则从 3 种活动列表中选择："*骑自行车*"、"*下棋*"和 "*洗碗*"。

微调的过程如图 5 所示。从质量上看，在整个训练过程中，样本会更加忠实地描述提示。这一趋势在数量上也有所体现，但由于 BERTScore 的微小变化可能对应着相关性的巨大差异（Zhang et al.）值得注意的是，微调集中的一些提示，如 "*海豚骑自行车*"，在预训练模型中的成功率为零；如果单独训练，由于没有奖励信号，这种提示不可能得到改进。只有通过跨提示的可迁移学习，这些困难的提示才能得到改善。

在微调过程中，几乎所有样本都变得更加卡通或艺术化。这并不是直接优化的结果。我们推测，这可能是训练前分布的一个函数（人们会认为对动物日常活动的描述更常见的是卡通风格而不是逼真风格），也可能是奖励函数的一个函数（也许 LLaVA 更容易识别简单的卡通风格图像的内容）。

图 6（泛化）对有限动物集的微调可泛化到新动物和非动物日常物品。最右边两列的提示分别是 "*水豚洗碗*"和 "*鸭子考试*"。附录 F 提供了定量分析，附录 G 提供了更多样本。

6.3 泛化

在大型语言模型上进行 RL 微调已被证明能产生有趣的泛化特性；例如，几乎完全以英语进行的指令微调已被证明能提高其他语言的能力（欧阳等人，2022 年）。这种现象很难与我们目前对泛化的理解相协调；*先验*地看，微调似乎更有可能只对微调提示集或分布产生影响。为了用扩散模型研究同样的现象，图 6 显示了一组与微调过程中没有出现的提示相对应的 DDPO 微调模型样本。与语言建模中的 "指令-跟随-迁移 "相一致，我们发现即使提示语的分布范围很窄，只有 45 种动物和 3 种活动，微调的效果也能普遍化。我们发现有证据表明，微调效果可以推广到训练分布以外的动物、非动物的日常物品，在提示-图像对齐的情况下，甚至可以推广到 "*考试*"等新颖的活动。

7 讨论与局限性

我们提出了一个基于 RL 的框架，用于训练去噪扩散模型，以直接优化各种奖励函数。通过将迭代去噪过程视为一个多步骤决策问题，我们设计出了一类能高效训练扩散模型的策略梯度算法。我们发现，对于难以通过提示指定的任务（如图像压缩）和难以通过编程评估的任务（如与提示语义一致），DDPO 是一种有效的优化器。为了提供一种自动获取奖励的方法，我们还提出了一种使用 VLM 对生成图像的质量提供反馈的方法。虽然我们的评估考虑了各种提示，但我们实验中的全部图像都受到了限制（*例如*，正在进行活动的动物）。未来的迭代可以扩大向 VLM 提出的问题（可能使用语言模型根据提示提出相关问题）以及提示分布的多样性。我们还选择不研究过度优化问题，这是 RL 微调的一个常见问题，即模型偏离原始分布太远而无法发挥作用（见附录 A）；我们强调这是未来工作的一个重要领域。我们希望这项工作能为大型生成模型提供更有针对性的训练，通过 RL 优化生成的模型能有效实现用户指定的目标，而不是简单地匹配整个数据分布。

更广泛的影响。生成模型可以成为有价值的生产力辅助工具，但也可能在用于虚假信息、冒充或网络钓鱼时造成危害。我们的工作旨在使扩散模型能够优化用户指定的目标，从而使其更加有用。这种调整具有有益的应用价值，例如生成更易理解的教育材料，但也可能被恶意使用，我们在此不作概述。要减轻生成模型的这种危害，可靠检测合成内容的工作仍然非常重要。

8 致谢

这项工作得到了美国海军研究办公室（Office of Naval Research）的部分支持，以及谷歌通过 TPU 研究云（TRC）捐赠的计算资源。迈克尔-詹纳（Michael Janner）得到了开放慈善项目（Open Philanthropy Project）的奖学金支持。杜一伦和凯文-布莱克获得了美国国家科学基金会的奖学金支持。

代码参考

我们在这项工作中使用了以下开源库：NumPy (Harris et al., 2020)、JAX (Bradbury et al., 2018)、Flax (Heek et al., 2023)、optax (Babuschkin et al., 2020)、h5py (Collette, 2013)、transformers (Wolf et al., 2020) 和 diffusers (von Platen et al., 2022)。

参考文献

Anurag Ajay、Yilun Du、Abhi Gupta、Joshua Tenenbaum、Tommi Jaakkola 和 Pulkit Agrawal。决策只需要传统生成模型吗?

Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev、 Michael King、 Markus Kunesch、Lena Martens、Hamza Merzic、Vladimir Mikulik、Tamara Norman、George Papamakarios、John Quan、Roman Ring、 Francisco Ruiz、 Alvaro Sanchez、 Rosalia Schneider、 Eren Sezener、 Stephen Spencer、 Srivatsan Srinivasan、 Wojciech Stokowiec、 王鲁豫、周光耀和 Fabio Viola。DeepMind JAX 生态系统，2020 年。网址 http://github.com/deepmind。

Philip Bachman 和 Doina Precup。作为顺序决策的数据生成。*神经信息处理系统进展*》，2015 年第 28 期。

白云涛、安迪-琼斯、卡迈勒-恩杜塞、阿曼达-阿斯凯尔、安娜-陈、诺瓦-达斯萨玛、道恩-德赖恩、斯坦尼斯拉夫-福特、迪普-甘古利、汤姆-汉尼根、尼古拉斯-约瑟夫、绍拉夫-卡达瓦斯、杰克逊-克尼恩、汤姆-康纳利、希尔-埃尔-肖克、尼尔森-埃尔哈格、Zac Hatfield-Dodds、Danny Hernandez、Tristan Hume、Scott Johnston、Shauna Kravec、Liane Lovitt、Neel Nanda、Catherine Olsson、Dario Amodei、Tom Brown、Jack Clark、Sam McCandlish、Chris Olah、Ben Mann 和 Jared Kaplan。从人类反馈中强化学习，训练一个乐于助人且无害的助手。*arXiv 预印本 arXiv:2204.05862*, 2022a.

白云涛、Saurav Kadavath、Sandipan Kundu、Amanda Askell、Jackson Kernion、Andy Jones、Anna Chen、Anna Goldie、 Azalia Mirhoseini、 Cameron McKinnon、 Carol Chen、 Catherine Olsson、 Christopher Olah、 Danny Hernandez、Dawn Drain、Deep Ganguli、Dustin Li、Eli Tran-Johnson、Ethan Perez、Jamie Kerr、Jared Mueller、Jeffrey Ladish、约书亚-兰道、卡迈勒-恩杜塞、卡米勒-卢科苏伊特、莉安-洛维特、迈克尔-塞利托、尼尔森-埃尔哈格、尼古拉斯-谢弗、诺埃米-梅尔卡多、诺瓦-达斯萨玛、罗伯特-拉森比、罗宾-拉森、萨姆-林格、斯科特-约翰斯顿、肖娜-克拉维茨、希尔-埃尔-肖克、斯坦尼斯拉夫-福特、塔梅拉-拉纳姆、蒂莫西-特林-劳顿、汤姆-康纳利、汤姆-亨尼根、特里斯坦-休姆、塞缪尔-R.Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan.*ArXiv preprint arXiv:2212.08073*, 2022b.

Arpit Bansal、Hong-Min Chu、Avi Schwarzschild、Soumyadip Sengupta、Micah Goldblum、Jonas Geiping 和 Tom Goldstein。扩散模型的通用指导。*IEEE/CVF 计算机视觉与模式识别会议论文集*》，第 843-852 页，2023 年。

James Bradbury、Roy Frostig、Peter Hawkins、Matthew James Johnson、Chris Leary、Dougal Maclaurin、George Necula、Adam Paszke、Jake VanderPlas、Skye Wanderman-Milne 和 Qiao Zhang。JAX：Python+NumPy 程序的

可组合转换，2018 年。URL http://github.com/google/jax.