ImageReward：学习和评估人类对文本到图像生成的偏好

摘要

作者提出了一种全面的解决方案，旨在通过人类偏好反馈来学习和改进文本到图像的生成模型。首先，作者构建了 ImageReward——这是一个通用的文本到图像的人类偏好奖励模型，能够高效地编码人类偏好。该模型基于作者设计的一套系统化的标注流程进行训练，此流程涵盖了评级与排序两种形式，并且至今已经收集到了 137,000 次由专家做出的比较结果。实验结果显示，在人工评估中，ImageReward 的表现优于现有的评分模型及评价指标，显示出其作为自动评估工具在文本到图像合成领域中的巨大潜力。在此基础上，为进一步提升模型性能，作者引入了奖励反馈学习 (ReFL) 方法，这是一种直接优化针对评分者扩散模型的有效算法。无论是通过自动化测试还是人工评测，ReFL 均展示了相对于其他对比方案的优势。为了便于学术界与工业界的进一步研究与应用，相关的所有代码及数据集已公开发布于 https://github.com/THUDM/ImageReward。

关键词：扩散模型；人类偏好学习

1 引言

近年来，文本到图像生成模型，包括自回归和基于扩散的方法，取得了显著进展。给定适当的文本描述（即提示），这些模型能够生成高保真度且语义相关的图像，涵盖了广泛的主题，引起了公众对其潜在应用和影响的极大兴趣。尽管取得了显著进展，现有的自监督预训练生成器仍存在诸多不足，主要挑战在于使模型与人类偏好保持一致。因为预训练分布往往较为嘈杂，与实际用户提示分布存在差异，这种固有的差异导致生成的图像中出现了一些已知的问题，包括但不限于文本图像对齐度，身体结构，人类审美，毒性内容和偏见等问题。这些问题难以仅通过改进模型架构和预训练数据来解决。在自然语言处理（NLP）领域，研究人员采用了从人类反馈中强化学习（RLHF）的方法，以引导大型语言模型向人类偏好和价值观靠拢。该方法依赖于学习奖励模型（RM），从大量专家注释的模型输出比较中捕捉人类偏好。虽然这种方法非常有效，但注释过程可能成本高昂且具有挑战性，需要数月的努力来建立标注标准、招募和培训专家、验证响应并最终生成奖励模型。认识到在生成模型中解决这些挑战的重要性，本文作者提出并发布了第一个通用的文本到图像人类偏好奖励模型——ImageReward。该模型基于现实世界的用户提示和相应的模型输出，在总共 137,000 对专家比较上进行了训练和评估。在此基础上，作者进一步研究了直接优化方法——奖励反馈学习（ReFL），以改进扩散生成模型。实验结果表明，ReFL 在自动和人工评估中均优于其他对比方法，展示了其在提升文本到图像生成质量方面的潜力。

该工作不仅提出了一个通用的文本到图像人类偏好数据集及其对应的奖励模型，还提出了第一个能够使用奖励模型直接微调扩散模型的方法。先前使用奖励模型微调扩散模型的方法通常是使用奖励模型过滤数据或对训练样本进行加权，作者提出的 ReFL 方法在扩散模型的人类偏好学习中具有开创性，值得学习和借鉴。

2 相关工作

该部分对 ImageReward 的相关工作进行简要介绍，主要分为扩散模型，学习人类偏好，使用监督学习以及使用强化学习进行奖励微调这四个部分。

2.1 扩散模型

去噪扩散概率模型（DDPMs）[8,17] 是一类被广泛使用的生成模型，已经成为大多数连续数据模态的范例，包括图像，视频，音频和 3D 模型等。文本到图像扩散模型（根据文本提示生成图像）随着 Latent Diffusion [16]，DALLE-2 [14] 等模型的出现，已经成为重要的工具。

2.2 学习人类偏好

人类偏好学习训练模型判断人们喜欢哪些行为，而不是直接从人类的示例中学习。这种训练通常是通过学习反映人类偏好的奖励模型，然后学习最大化奖励的策略来完成的 [3]。偏好学习方法（例如从人类反馈中强化学习 (RLHF)）已广泛用于微调大型语言模型，以便它们能够提高总结能力 [20]，指令遵循能力 [13] 或一般对话能力 [1]。现有的学习人类偏好的工作也将视角放在了扩散模型上，例如 PickScore [9] 和 Human Preference Score v2 [23]，这些模型是根据扩散模型针对同一提示词生成的图像对之间的人类判断进行训练的。

2.3 使用监督学习进行奖励微调

Lee 等人 [10] 和 Wu 等人 [24] 使用监督方法对奖励的扩散模型进行微调。这些方法使用预训练模型生成图像，然后在图像上进行微调，同时根据奖励函数对训练样本进行加权或丢弃低奖励样本。与 RL 方法不同，该模型不是在当前策略生成的样本上进行在线训练的。然而，Dong 等人 [5] 使用了这种方法的在线版本，其中样本在多轮训练中重新生成，这可以看作是一种简单的强化学习。

2.4 使用强化学习进行奖励微调

Fan [6] 将去噪过程解释为多步骤决策任务，并使用策略梯度算法来微调扩散采样器。在此基础上，Black 等人 [2] 使用策略梯度算法来微调任意黑盒目标的扩散模型。Hao 等人 [7] 不是优化模型参数，而是应用 RL 来改进输入提示词。虽然 RL 方法很灵活，它们不需要可微的奖励，但是在实践中，许多奖励函数是可微分的，或者可以被可微分的方式重新实现，因此通常可以使用或分析奖励函数的梯度。在这种情况下，使用强化学习会丢弃有用的信息，导致优化效率低下。

3 本文方法

3.1 本文方法概述

论文提出了两个主要方法来改进文本到图像生成模型的效果：ImageReward 和 Reward Feedback Learning (ReFL)。作者了设计一个通用的奖励模型来捕捉人类对文本到图像生成任务的偏好。文章作者设计了一个系统化的标注流程以对生成的图像进行评分，最后从每组图像得到从最好到最差的偏好排序。然后使用标注好的数据对奖励模型进行偏好训练，得到一个能对输入图像输出其奖励分数的奖励模型。训练好奖励模型后，作者使用了自己提出的直接调优方法 ReFL 以使用奖励模型监督微调扩散生成模型。因奖励模型在扩散过程的早期步骤不能准确反映奖励分数，因此作者在扩散过程中的某些中间步骤使用奖励模型监督微调扩散模型。

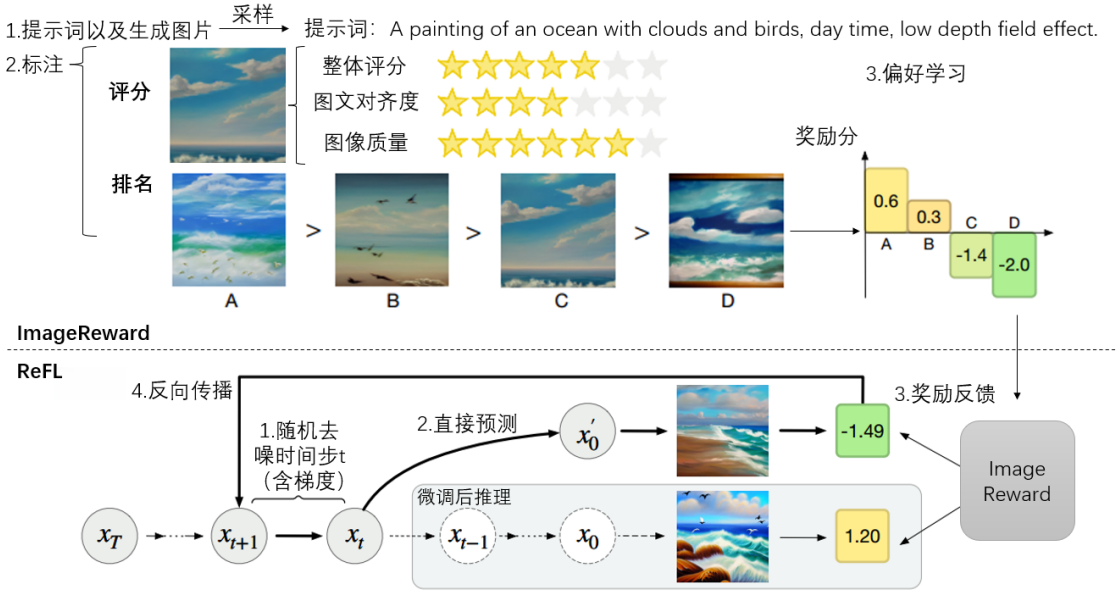

图 1. 方法示意图

3.2 奖励模型

训练奖励模型首先需要反映人类偏好的比较对样本。为了收集足量的比较对样本，作者挑选了生成模型的提示词以及使用生成模型对每一个提示词生成了多张图像。提示词数据集利用了来自开源数据集 DiffusionDB [22] 的多样化真实用户提示选择。为了确保所选提示的多样性，作者采用了一种基于图的算法，该算法利用了基于语言模型的提示相似性 [15,18,21]，最终产生了 10,000 个候选提示。每个提示词都附有 4 到 9 张生成图像，从而产生了 177,304 个用于标注的候选比较对。

有了生成图像和对应的提示词之后，作者设计了一个专门的标注流程来对每张图像进行评分。标注流程涉及提示词注释阶段，其中包括对提示词进行分类和识别有问题的提示词，以及文本图像评级阶段，其中根据对齐，保真度和无害性对图像进行评级。随后，标注者按偏好顺序对图像进行排名。为了解决排名中的潜在矛盾，作者在标注文档中提供了权衡规则。作者的标注系统由三个阶段组成：提示词标注、文本图像评级和图像排名。标注者通过专门的

培训后再进行标注，并且由专门的人员对标注质量进行检查，最后对标注质量不合格的样本进行重新标注。最终作者收集了 8,878 个提示的有效注释，得到了 136,892 个比较对。

与先前研究 [13, 20] 中对语言模型的 RM 训练类似，作者将偏好注释制定为排名。我们有 $k \in [4, 9]$ 个图像对同一提示 T 进行排名（从最好到最差的表示为 $x_1, x_2, ..., x_k$），如果两幅图像之间没有联系，则最多获得 C_k^2 个比较对。对于每次比较，如果 x_i 更好而 x_j 更差，则损失函数可以表示为：

$$loss(\theta) = -\mathbb{E}_{(T,x_i,x_j)\sim\mathcal{D}}[\log(\sigma(f_\theta(T, x_i) - f_\theta(T, x_j)))]$$

其中，$f_\theta(T, x)$ 是一个表示提示词 T 和生成图像 x 对应偏好分数的标量。

在训练环节中，作者使用 BLIP [11] 作为 ImageReward 的主干，因为它在作者的初步实验中表现优于传统的奖励模型所使用的 CLIP。通过 BLIP 提取图像和文本特征，将它们与交叉注意相结合，并在最后使用一个 MLP 生成标量以进行偏好比较。作者观察到训练奖励模型时出现的快速收敛和随之而来的过拟合。为了解决这个问题，作者冻结了一些 BLIP 层的参数，发现适当数量的固定层可以提高 ImageReward 的性能。ImageReward 还表现出对训练超参数的敏感性，例如学习率和批量大小，因此作者根据验证集执行仔细的网格搜索以确定最佳超参数的值。

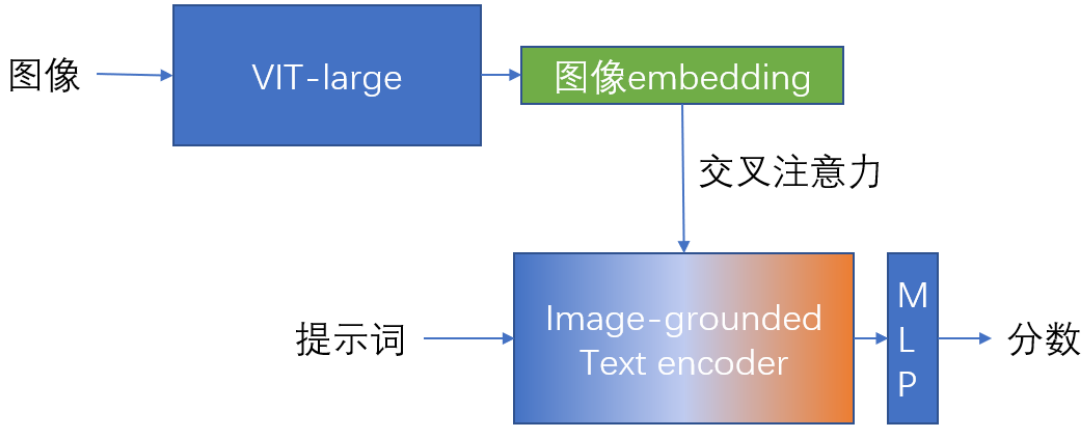

图 2. 奖励模型示意图

3.3 ReLF 微调方法

尽管 ImageReward 可以从提示词的生成的多个图像中挑选出人类高度偏爱的图像，但生成图像然后过滤图像的范式在实际应用中可能成本高昂且效率低下。因此，作者寻求改进文本到图像的生成模型，特别是当下广泛使用的潜在扩散模型，以便在一次或极少数的尝试中实现高质量的生成。

在 NLP 中，研究人员使用强化学习算法来引导语言模型与人类偏好保持一致 [12, 13, 20]，这依赖于整个生成过程中的似然概率。然而，与语言模型不同，潜在扩散模型 (LDM) 的多步去噪生成无法为其生成提供似然概率，因此无法采用相同的 RLHF 方法。一种潜在的类似方法是 LDM 推理期间的分类器引导 [4, 19] 技术。尽管如此，它仅用于推理，并使用必须在嘈杂的中间潜在上训练的分类器，这自然与 RM 的输出分数相矛盾，在 RM 的输出分数中，图

像需要完全去噪才能输出正确的人类偏好。一些同期的工作提出了替代的间接解决方案，例如使用 RM 过滤数据集以进行微调 [5,24]，或根据训练样本的质量重新调整其权重损失 [10]。然而，这些面向数据的方法实际上是间接的。它们可能严重依赖微调数据分布，最终只能轻微改善 LDM。

为了解决当前方法不能直接微调扩散模型的缺点，作者开发了一种直接优化方法，以根据 RM（例如 ImageReward）改进 LDM。通过研究去噪步骤中的 ImageReward 分数，作者得出了一个有趣的见解，即当在步骤 t 处直接预测最终去噪图像时，不同的 t 对于 ImageReward 对生成图像的预测分数的性能上会有所不同。当 $t \leq 15$ 时，ImageReward 对所有生成图像的预测分数都很低。当 $15 \leq t \leq 30$ 时，高质量生成图像的分数开始脱颖而出，但总体而言，仍然无法根据当前的 ImageReward 分数清楚地判断所有生成图像的最终质量。当 $t \geq 30$ 时，不同生成图像的 ImageReward 分数能够清楚地区分开来。总体而言，就是在较低的去噪步数直接预测最终图像时，ImageReward 无法对生成图像判别出有意义的分数。

根据观察，作者得出结论，经过 30 步去噪（不一定是最后一步）后直接预测的生成图像的 ImageReward 分数可以作为改进 LDM 的可靠反馈。

因此，作者提出了一种算法，通过将 RM 的分数视为人类偏好损失来直接微调 LDM，以反向传播梯度（到去噪过程中随机选择的后一步 t。随机选择 t 而不是使用最后一步的原因是，如果只保留最后一步去噪的梯度，训练就会非常不稳定，结果也会很糟糕。在实践中，为了避免快速过度拟合和稳定微调，作者重新加权 ReFL 损失并用预训练损失进行正则化。最终的损失可写为：

$$\mathcal{L}_{reward} = \lambda \mathbb{E}_{(y_i)\sim\mathcal{Y}}(\phi(r(y_i, g_\theta(y_i))))$$

$$\mathcal{L}_{pre} = \mathbb{E}_{(y_i,x_i)\sim\mathcal{D}}(\mathbb{E}_{\varepsilon(\curvearrowright\supset),\curvearrowright\supset,\epsilon\sim\mathcal{N}(\nvdash,\nVdash),\approx}[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y_i))\|_2^2])$$

其中，θ 表示 LDM 的参数，$g_\theta(y_i)$ 表示使用提示词 y_i 且由 LDM 生成的图像。$\mathcal{L}_{pre}$ 则是预训练损失，与 LDM 的训练损失相同。

4 复现细节

4.1 与已有开源代码对比

本文复现的参考代码为 https://github.com/THUDM/ImageReward。在使用源代码的数据集加载工具时，发现数据集加载失败，原因是作者没有完整实现数据集构建的代码，因此额外补全了相关代码。为了可视化训练过程，也实现了训练过程中的可视化代码（主要观察损失的变化）。

4.2 实验环境搭建

此次复现的工作为深度学习领域的工作，所使用的编程语言为 python，主要使用 anaconda 来搭建实验环境，操作系统为 linux 系统。

首先在 anaconda 官网（https://repo.anaconda.com/archive/）上查找所需版本安装包，然后在控制台输入：wget https://repo.anaconda.com/archive/[所需版本]，例如：wget https://repo.anaconda.com/archive/Anaconda3-2022.10-Linux-x86_64.sh，下载完毕后依次输入 chmod +x Anaconda3-5.3.0-Linux-x86_64.sh 以及./Anaconda3-5.3.0-Linux-x86_64.sh（出现需要确定的

选项点击 Enter 或输入 yes 按照默认配置即可）。打开新的终端后，输入 conda -V，若显示 conda 版本则表示安装成功。

接下来使用 anaconda 创建虚拟环境。输入：conda create -n ImageReward python=3.11，将该环境命名为 ImageReward 以及 python 版本指定为 3.11。创建虚拟环境后，使用命令激活环境：conda activate ImageReward。之后，根据自身硬件设备（可输入 nvidia-smi 查看），在 pytorch 官网（https://pytorch.org/）中选择合适的版本安装。在本文中为：conda install pytorch torchvision torchaudio pytorch-cuda=12.4 -c pytorch -c nvidia。

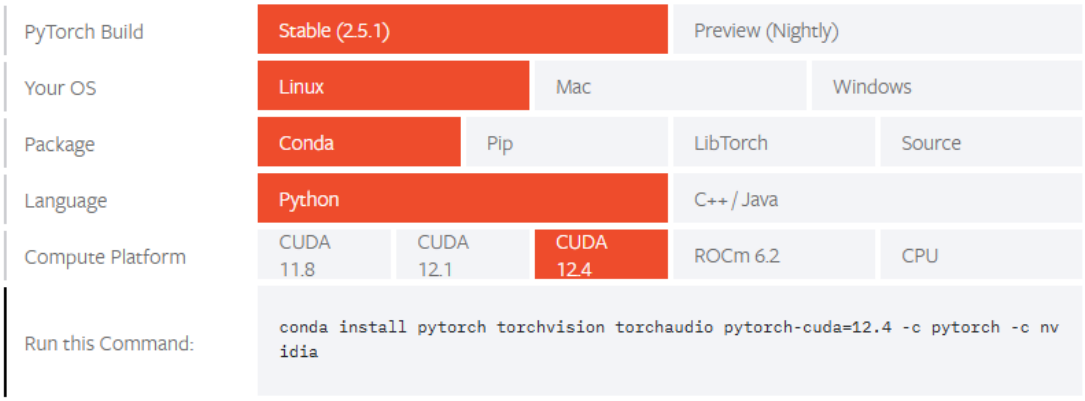

图 3. 在 pytorch 官网选择合适的版本

最后，在源码官网中拉取项目文件到本地之后，让控制台的工作目录处在项目文件中，输入 pip install -r requirements.txt 安装所需依赖包即可完成实验环境的搭建。

5 实验结果分析

由于实验室硬件条件的限制，batchsize 以及 epoch 等参数不能与原文一致，其他超参数等实验细节设置则尽量与原文保持一致。本次实验在训练 ImageReward 时使用了 2 张 32GB 的 V100，批量大小为 16，epoch 为 10，梯度累积设置为 4 步，同时冻结了 BLIP 中 Image-text-encoder 的前 70% 的参数，学习率设置为 1e-5。可以从图 4 中看到，尽管训练损失还没有完全收敛就结束训练，但是原文中提到训练 RM 时容易遇到过拟合的问题，因此在观察到 loss 有较为明显的下降时就停止训练，而不是等待 loss 完全收敛。表 1 和表 2 也证明了这样训练出的 RM 在反映人类偏好的性能上与原文的 RM 一致，尽管在分数的数值上有一定的偏移（RM 是以排序对训练的，因此单一的具体分数没有太大意义）。图 6 则展示了 RM 对同一个提示词生成的不同图像的评价，可以看到评价最低的图像是最不符合提示词且最偏离人类审美的，评价最高的图像则符合提示词以及人类审美，但仍有一定的缺陷，因此分数不高。

在微调扩散模型时，使用的扩散模型为预训练的 Stable Diffusion1.4，使用 1 张 32GB 的 V100 进行训练，为了与原文训练的样本数量一致，设置的批量大小为 2，不进行梯度累积，最大步数为 3200。但由此微调出的模型生成效果不理想，在训练过程中已经观察到损失已经稳定，但无论是使用训练中途的参数还是训练完成的参数，生成的图像质量均不如微调之前的质量。推测是这种偏好训练模型微调扩散模型时需要较大的批量大小，才能在反向传播时有效表征人类的整体偏好，因为一张图像通常是在多个维度上进行评价的，较少的样本很容易造成某一维度上缺陷的放大。

表 1. 不同模型的偏好准确率比较

模型	**偏好准确率**
CLIP Score	54.82
Aesthetic Score	57.35
BLIP Score	57.76
ImageReward（原文）	65.14
ImageReward（复现）	**65.53**

表 2. 文本转图像模型由人工和自动指标（ImageReward、CLIP 和 FID）进行的排名。

数据集 & 模型	真实用户提示词					
	人类评价		ImageReward（原文）		ImageReward（复现）	
	排名	# 获胜	排名	得分	排名	得分
Openjourney	1	507	1	0.2614	1	0.4480
Stable Diffusion 2.1-base	2	463	2	0.2458	2	0.4379
DALL-E 2	3	390	3	0.2114	3	0.3952
Stable Diffusion 1.4	4	362	4	0.1344	4	0.3553
Versatile Diffusion	5	340	5	0.2470	5	0.0223
CogView 2	6	74	6	-1.2376	6	-0.8808
与人类评价的相关系数 ρ	-		1.00		1.00	

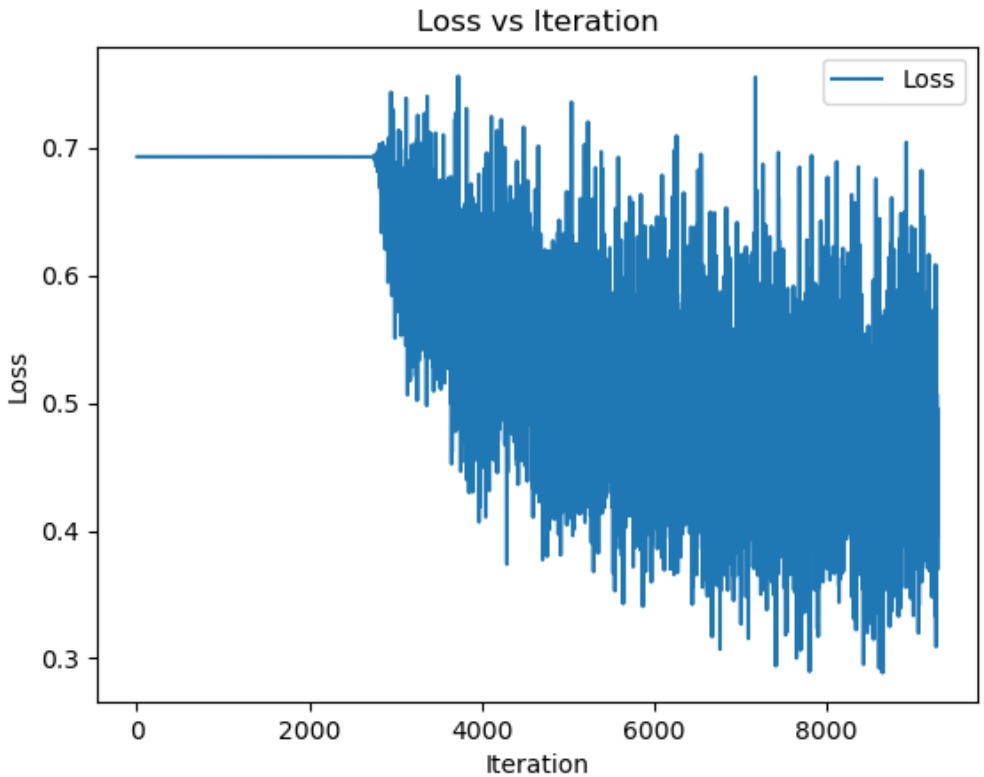

图 4. 训练 RM 时的损失随迭代次数的变化

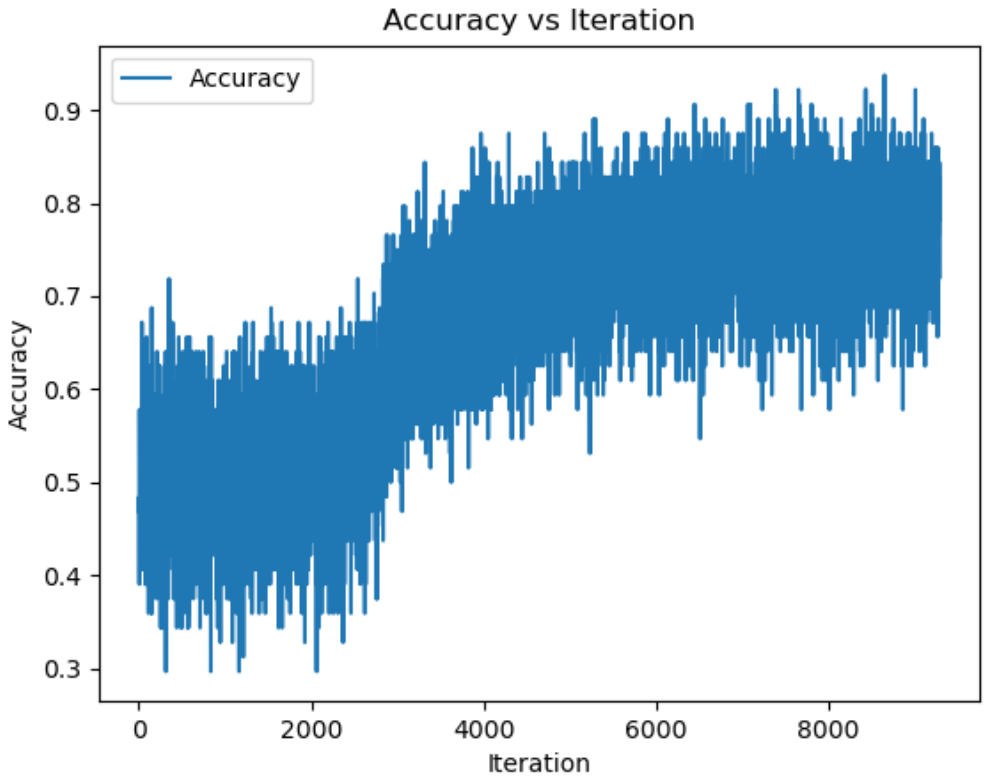

图 5. 训练 RM 时的准确率随迭代次数的变化

图 6. ImageReward 对模型生成图片的评价分数

6 总结与展望

本文介绍了 ImageReward 和 ReFL 的数据集准备，主要框架和训练方法。同时介绍了扩散模型，人类偏好学习，奖励微调扩散模型的相关工作。并讲解了实验的复现细节，环境搭

建以及实验结果的分析。通过本次实验，成功复现了 ImageReward 的代码以及实验结果，但是 ReFL 部分的实验结果不尽人意，生成图像的质量很不理想，同时也有复现者出现了类似的结果。未来可进一步研究微调效果不理想的原因并对其进行改进，希望能研究出在硬件资源较少的情况下更加稳定有效的微调方法。

参考文献

[1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[3] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[5] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

[6] Ying Fan and Kangwook Lee. Optimizing ddpm sampling with shortcut fine-tuning. *arXiv preprint arXiv:2301.13362*, 2023.

[7] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[9] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

[10] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[12] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback, 2021. *URL https://arxiv.org/abs/2112.09332*, 2021.

[13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[14] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[15] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[17] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[18] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

[19] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[20] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[21] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*, 2022.

[22] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

[23] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[24] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3), 2023.