文本到图像生成的丰富人工反馈

Youwei Liang*†[1], Junfeng He*‡[2], Gang Li*(‡[2]) , Peizhao Li†[5], Arseniy Klimovskiy[2], Nicholas Carolan[2]、 Jiao Sun†§[3], Jordi Pont-Tuset[2], Sarah Young[(2)], Feng Yang[(2)], Junjie Ke[(2)], Krishnamurthy Dj Dvijotham[2], Katherine M.Collins†[4], Yiwen Luo[2], Yang Li[(2)], Kai J Kohlhoff[(2)], Deepak Ramachandran[(2)], and Vidhya Navalpakkam[(2)]

[1]加州大学圣地亚哥分校
[2]谷歌研究院
[3]南加州大学 [4]剑桥大学 [5]布兰戴斯大学

摘要

最近的文本到图像（T2I）生成模型，如稳定扩散和 Imagen，在根据文本描述生成高分辨率图像方面取得了重大进展。然而，许多生成的图像仍然存在假象/似是而非、与文本描述不对齐以及美学质量低等问题。受针对大型语言模型的"人反馈强化学习"（RLHF）的成功启发，之前的研究收集了人类提供的分数作为对生成图像的反馈，并训练了一个奖励模型来改进 T2I 的生成。在本文中，我们通过以下方式丰富了反馈信号：(i) 标记与文本不符或错位的图像区域；(ii) 标注文本提示中哪些单词在图像上被错误呈现或缺失。我们收集了 18K 张生成图像上的丰富人类反馈（RichHF- 18K），并训练了一个多模态转换器来自动预测丰富的反馈。我们的研究表明，可以利用预测到的丰富人类反馈来改进图像生成，例如，通过选择高质量的训练数据来微调和改进生成模型，或者通过创建具有预测热图的遮罩来涂抹有问题的区域。值得注意的是，除了用于生成采集人体反馈数据的图像的模型（Muse）之外，这些改进还能推广到其他模型（稳定差异融合变体）。RichHF-18K 数据集将发布在我们的 GitHub 存储库中：https://github.com/google-research/google-research/tree/master/richhf 18k。

*共同第一作者，技术贡献相同
†工作是在谷歌实习期间完成的。
电子邮件： youwei@ucsd.edu
‡通讯作者，主要贡献相同。
电子邮件：{Junfenghe,leebird}@google.com
§目前隶属于谷歌双子座团队

1. 简介

文本到图像（T2I）生成模型[12, 17, 42, 43, 57, 59, 60]正迅速成为娱乐、艺术、设计和广告等各个领域内容创建的关键，同时也被推广到图像编辑[4, 28, 45, 51]、视频生成[23, 36, 54]等许多其他应用中。尽管最近取得了重大进展，但其输出结果通常仍存在一些问题，如伪造/似是而非、与文本描述不匹配以及审美质量低等[31, 53, 55]。例如，在主要由稳定扩散模型变体生成的图像组成的 Pick-a-Pic 数据集[31]中，许多图像（*如*图 1）包含扭曲的人体/动物体（*如*有五个以上手指的人手）、扭曲的图像（*如*图 2）、扭曲的图像（*如*图 3）、扭曲的图像（*如*图 4）、扭曲的图像（*如*图 5）和扭曲的图像（*如*图 6）。

物体和不可靠性问题，如漂浮的灯。我们的人工评估实验发现，在数据集中生成的图像中，只有~ 10% 不存在伪影和不可信问题。

假象。同样，文本与图像不对齐的问题也很常见，*例如*，提示是 "*一个人跳入河中*"，但生成的图像却显示这个人是站着的。

然而，现有的生成图像自动评估指标，包括著名的 IS [44] 和 FID [20]，都是根据图像的分布计算的，可能无法反映单个图像的细微差别。最近的重新搜索收集了人类的偏好/评分来评价生成图像的质量，并训练了评价模式来预测这些评分[31, 53, 55]，特别是 ImageRe-。

ward [55] 或 Pick-a-Pic [31]。这些指标虽然更有针对性，但仍然是将一幅图像的质量概括为一个单一的数字分数。在提示-图像配准方面，还有一些开创性的单分指标，如 CLIP- Score [19] 和最近的问题生成和回答管道 [8, 10, 25, 58]。这些模型虽然经过校准，可操作性更强，但价格昂贵且复杂，仍无法定位图像中的错位区域。

在本文中，我们提出了一个数据集和一个细粒度多方面评价模型，这些评价是可解释和可归因的（*例如*，对有人工痕迹/似是而非或图像-文本错位的区域），与单个标量分数相比，它能提供对图像质量更丰富的理解。作为第一个贡献，我们收集了 18K 幅图像的 Rich Hu- man Feedback 数据集（RichHF-18K），其中包括：(i) 图像上的点注释，这些点注释突出了不可信/伪造以及文本-图像错位的区域；(ii) 提示语上的标注词，指明生成图像中缺失或错误表述的概念；以及 (iii) 四种类型的细粒度评分，分别针对图像的可塑性、文本与图像的一致性、美学和总体评分。借助 RichHF-18K，我们设计了一个多模态转换器模型，并将其命名为 Rich Automatic Hu- man Feedback (RAHF)，用于学习预测生成图像及其相关文本提示上这些丰富的人类注释。因此，我们的模型可以预测不靠谱和错位区域、错位关键词以及细粒度评分。这不仅能提供可靠的评分，还能提供有关生成图像质量的更详细、更可解释的见解。据我们所知，这是首个针对最先进的文本到图像生成模型的丰富反馈数据集和模型，为评估 T2I 生成提供了自动、可解释的管道。

本文的主要贡献概述如下：

1. 第一个关于生成图像的丰富人类反馈数据集（RichHF-18K）（由细粒度分数、不可信度（假象）/错位图像区域和错位关键字组成），涉及 18K 张 "Pick-a-Pic "图像。
2. 多模态变换器模型（RAHF）可预测对生成图像的丰富反馈，我们证明该模型与测试集上的人类注释高度相关。
3. 我们进一步证明了 RAHF 预测的丰富人类反馈对改进图像生成的有用性：(i) 将预测的热图用作遮罩，以涂抹有问题的图像区域；(ii) 利用预测分数帮助微调图像生成模式（如 Muse [6]），*例如*，通过选择/过滤微调数据或作为奖励指导。我们的研究表明，在这两种情况下，我们都能获得比原始模型更好的图像。
4. Muse 模型与我们训练集中生成图像的模型不同，它的改进表明我们的 RAHF 模型具有良好的泛化能力。

RAHF 模型的良好泛化能力。

2. 相关作品

文本到图像生成 文本到图像（T2I）生成模型在深度学习时代经历了几种流行模型架构的演变和迭代。早期的工作是生成对抗网络（GAN）[3, 16, 27]，它并行地训练图像生成器和判别器，以区分真实图像和生成的图像（另见[33, 39, 48, 56, 61, 63]等）。另一类生成模型来自变异自动编码器（VAE）[21, 30, 49]，它可以优化图像数据可能性的证据下限（ELBO）。

最近，扩散模型（DMs）[22, 37, 42, 47]成为图像生成领域最先进的（SOTA）技术[13]。DMs 经过训练可从随机噪声中逐步生成图像，与 GANs 相比，DMs 能够捕捉更多的多样性，并获得良好的样本质量[13]。潜在扩散模型[42]是一种进一步的改进，它在一个紧凑的潜在空间中执行扩散过程，以提高效率。

文本到图像的评估和奖励模型 最近有很多关于文本到图像模式评估的研究[9, 26, 31, 32, 38, 52, 53, 55]。Xu 等人[55]收集了一个人类偏好数据集，要求用户对多幅图像进行排序，并根据图像质量进行评分。他们训练了一个用于人类偏好学习的奖励模型 ImageReward，并提出了奖励反馈学习（Reward Feedback Learning，ReFL），用于利用 Im- ageReward 模型调整扩散模型。Kirstain 等人[31]建立了一个网络应用程序，通过让用户从一对生成的图像中选择更好的图像来收集人类偏好，从而产生了一个名为 "Pick-a-Pic "的数据集，其中包含由稳定扩散 2.1、Dreamlike Photoreal 2.05 和稳定扩散 XL vari- ants 等 T2I 模型生成的 50 多万张图片。他们利用人类偏好数据集来训练基于 CLIP [40] 的评分函数，即 PickScore，以预测人类偏好。Huang 等人[26]提出了一个名为T2I-CompBench的基准，用于评估文本到图像模型，该基准由6000个文本提示组成，描述了属性绑定、对象关系和复杂组合。他们利用多种预训练的视觉语言模型（如 CLIP [40] 和 BLIP [35]）来计算多个评估指标。Wu 等人[52, 53]收集了人类对生成图像选择的大规模数据集，并利用该数据集训练了一个分类器，该分类器可输出人类偏好分数（HPS）。他们利用 HPS 对稳定扩散进行了调整，结果显示图像生成效果有所改善。最近，Lee[32]提出了一种采用多个细粒度指标对 T2I 模型进行整体评估的方法。

尽管做出了这些宝贵的贡献，但大多数现有作品都只使用二进制人类评级或偏好排序，而没有对关键词进行优化。

图 1.**注释用户界面示意图**。注释者在图像上标注点，表示与文本提示不一致的伪造/似是而非区域（红点）或错位区域（蓝点）。然后，他们点击单词来标记错位的关键词（下划线和阴影），并选择可信度、文本-图像对齐度、美观度和整体质量（下划线）的分数。

但是，我们的研究还无法提供详细的可执行反馈，如图像中的可植入区域、对齐错误的区域或生成图像上对齐错误的关键字。与我们的工作相关的一篇最新论文是 Zhang 等人的论文[62]，他们收集了一个用于图像合成任务的人工制品区域数据集，训练了一个基于分割的模型来预测人工制品区域，并提出了一种针对这些区域的区域内绘方法。然而，他们的工作重点仅在于伪影区域，而在本文中，我们为 T2I 生成收集了丰富的反馈信息，其中不仅包含伪影区域，还包括错位区域、错位关键词以及来自多个方面的四个细粒度评分。据我们所知，这是第一项针对文本到图像模型的异构丰富人类反馈的工作。

3. 收集丰富的人类反馈

3.1. 数据收集过程

本节将讨论我们收集 RichHF-18K 数据集的过程，其中包括两个热图（伪造/似是而非和错位）、四个细粒度分数（似是而非、对齐、美学和总分）和一个文本序列（错位关键词）。

对于生成的每张图像，首先要求注释者检查图像并阅读生成图像的文本提示。然后，他们会在图像上标出一些点，以标明与文本提示不符/伪造或错位的位置。注释者被告知，每个标记点都有一个 "有效半径"（图像高度的 1/20），以标记点为中心形成一个假想的圆盘。这样，我们就可以用相对较少的点来覆盖有缺陷的图像区域。最后，标注者会对错位的关键词进行标注，并以 5 分制的李克特量表分别对可信度、图像-文本对齐度、美观度和整体质量进行四种类型的评分。去掉关于图像不可信性/伪影和配准错误的详细定义，请参见补充材料。我们设计了一个网络用户界面（如图 1 所示），以方便数据收集。有关数据收集过程的更多详情，请参阅补充材料。

3.2. 人工反馈整合

为了提高收集到的人类对生成图像的反馈的可靠性，每对图像-文本都由三位注释者进行注释。因此，我们需要整合每个样本的多个注释。在评分方面，我们只需将多个注释者对图像的评分平均，即可得到最终评分。对于不对齐的关键字注释，我们使用最常见的关键字标签进行多数票表决，以获得对齐/不对齐指标的最终序列。对于点注释，我们首先将其转换为每个注释的热图，其中每个点被转换为热图上的一个圆盘区域（如上一小节所述），然后我们计算不同注释者的平均热图。明显不可信的区域很可能被所有注释者注释，并在最终的平均热图上具有较高的值。

3.3. RichHF-18K：丰富的人类反馈数据集

我们从 "Pick-a-Pic "数据集中选择了一个图像-文本对子集进行数据标注。虽然我们的方法是通用的，适用于任何生成的图像，但由于其重要性和更广泛的应用，我们选择了大部分数据集为照片写实图像。此外，我们还希望在所有图像中实现均衡分类。为了确保平衡，我们利用 PaLI 视觉问题解答（VQA）模型[7] 从 Pick-a-Pic 数据样本中提取了一些基本特征。具体来说，我们对 Pick-a-Pic 中的每对图像-文本提出了以下问题。1) 图像是否逼真？2) 哪个类别最能描述图像？请在 "人物"、"动物"、"物体"、"室内场景 "和 "室外场景 "中选择一个。在我们的人工检查下，PaLI 对这两个问题的回答基本可靠。我们利用这些答案从 Pick-a-Pic 中抽取了不同的子集，得到了 17K 个图像-文本对。我们将 17K 个样本随机分成两个子集，一个是包含 16K 个样本的训练集，另一个是包含 1K 个样本的验证集。16K 个训练样本的属性分布见补充材料。此外，我们还从 "Pick-a-Pic "测试集中收集了关于独特提示及其相应图片的丰富的人类反馈，作为我们的测试集。我们总共收集了来自 Pick-a-Pic 的 18K 对图像-文本的丰富人类反馈。我们的 RichHF- 18K 数据集包括 16K 训练样本、1K 验证样本和 1K 测试样本。

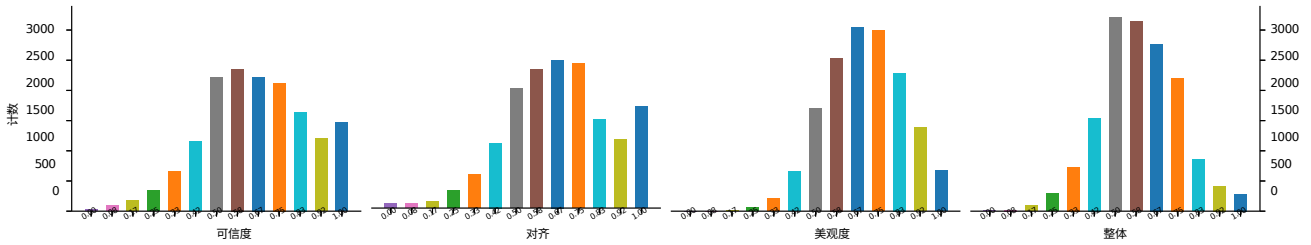


图 2.训练集中图像-文本对**平均得分**的直方图。

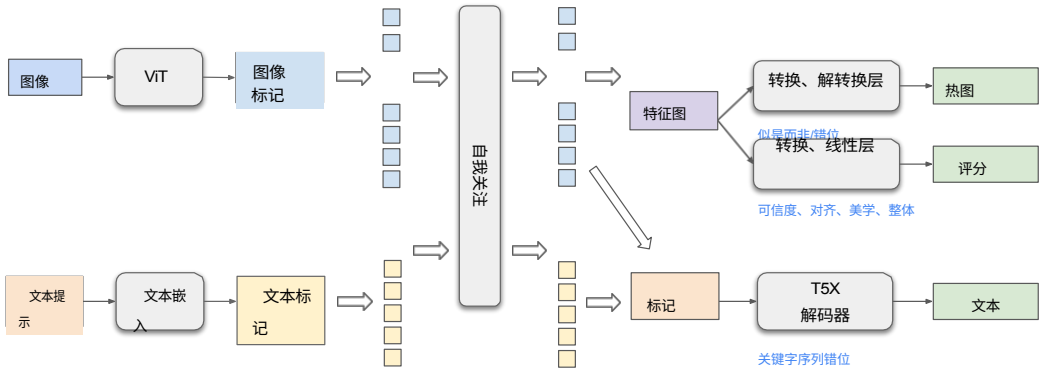


图 3.丰富反馈模型的架构我们的模型由两个计算流组成：一个视觉流和一个文本流。我们对 ViT 输出的图像标记和 Text-embed 模块输出的文本标记进行自我关注，以融合图像和文本信息。视觉标记被重塑为特征图，并映射为热图和分数。视觉标记和文本标记被发送到 Transformer 解码器，以生成文本序列。

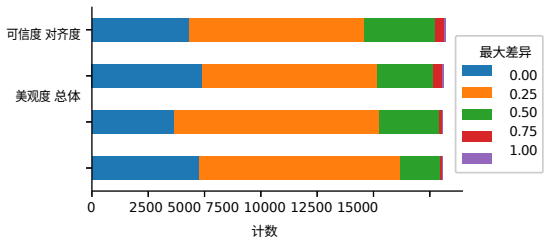


图 4.训练集中分数**差异最大的**样本计数。

3.4.RichHF-18K 的数据统计

在本节中，我们将总结分数的统计数据，并对分数进行注释者一致性分析。我们用公式（$s - s_{(min)}$）/（$s_{(max)}$）对分数 s 进行标准化。 s_{min}）（s_{max}= 5，$s_{(min)}$= 1），从而使分数位于 0，1] 范围内。

得分的直方图如图 2 所示。得分的分布类似于高斯分布，而可信度和文本-图像配准得分 1.0 的比例略高。收集到的分数分布确保了我们有合理数量的负样本和正样本来训练一个良好的奖励模型。

为了分析注释者对图像和文本的评分一致性，我们计算了分数之间的最大差异： max_{diff}= max(scores)− min(scores)
其中 scores 是图像-文本对的三个分数标签。
对的三个分数标签。我们在图 4 中绘制了 max_{diff}的直方图。我们可以

可以看出，约 25% 的样本与注释者完全一致，约 85% 的样本与注释者有良好的一致（标准化后的 $max_{(diff)}$小于或等于 0.25，或 5 点李克特量表中的 1）。

4. 预测丰富的人类反馈

4.1. 模型

4.1.1 结构

我们的模型架构如图 3 所示。我们采用了基于 ViT [14] 和 T5X [41] 模型的视觉语言模型，其灵感来自 Spotlight 模型架构 [34]，但对模型和预训练数据集进行了修改，以更好地适应我们的任务。由于我们的任务需要双向信息传播，因此我们在串联的图像标记和文本标记之间使用了自关注模块[50]，类似于 PaLI[7]。文本信息传播到图像标记，用于文本错位评分和热图预处理，而视觉信息传播到文本标记，用于更好的视觉感知文本编码，以解码文本错位序列。为了在更多样化的图像上对模型进行预训练，我们在预训练任务混合物中添加了 WebLI 数据集[7]上的自然图像标题任务。

具体来说，ViT 将生成的图像作为输入，并输出图像标记作为高级表示。文本提示标记被嵌入到密集向量中。图像标记和嵌入的文本标记被连接起来，并由 T5X 中的 Transformer 自注意编码器进行编码。在已编码的融合文本和图像标记之上、

我们使用三种预测器来预测不同的输出结果。在热图预测中，图像标记会被重塑为特征图，并通过卷积层、解压缩层和sigmoid激活层发送，输出误差热图和错位热图。对于分数预测，特征图会经过卷积层、线性层和西格码激活，产生作为细粒度分数的标量。

为了预测关键词错位序列，用于生成图像的原始提示被用作模型的文本输入。修改后的提示符被用作 T5X 解码器的预测目标。*例如，如果*生成的图像包含一只黑猫，而单词 "*yellow* "与图像错位，则*生成 "yellow 0 cat"*。在评估过程中，我们可以使用特殊后缀提取错位关键词。

4.1.2 模型变体

我们探讨了热图和分数预测头的两种模型变体。

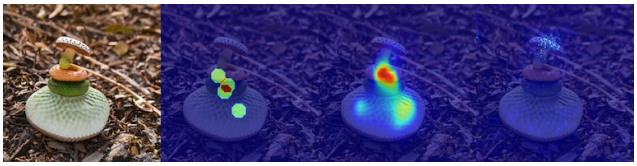
多头 预测多个热图和分数的直接方法是使用多个预测头，每个分数和热图类型使用一个预测头。这总共需要七个预测头。

增强提示 另一种方法是为每种预测类型使用一个**提示**头，*即*总共使用三个提示头，分别用于热图、得分和错配序列。为了让模型了解细粒度的热图或分数类型，我们用输出类型来增强提示。更具体地说，我们会在一个示例的每个特定任务的提示中添加一个任务字符串（*例如* "不可信度热力图"），并使用相应的标签作为训练数据。在推理过程中，通过在提示中添加相应的任务字符串，单个热图（分数）头可以预测不同的热图（分数）。正如我们在实验中展示的那样，这种增强提示方法可以创建特定任务的视觉特征图和文本编码，在某些任务中的表现明显更好。

4.1.3 模型优化

对于热图预测，我们使用像素均方误差（MSE）损失来训练模型；对于分数预测，我们使用 MSE 损失来训练模型。对于错位序列预测，模型采用教师强迫交叉熵损失进行训练。最终的损失函数是热图 MSE 损失、分数 MSE 损失和序列教师强迫交叉熵损失的加权组合。

(a) 图像 (b) GT (c) 我们的模型 (d) ResNet-50 图 5.不可信度热图示例。提示：*从尼康 D5 后方拍摄的一位身材苗条、长发披肩、身穿白色紧身衣的亚洲小女孩在海滩上奔跑的照片*



(a) 图像 (b) GT (c) 我们的模型 (d) CLIP 梯度 图 6. 错位热图示例。提示：*蘑菇上的一条蛇*

4.2. 实验

4.2.1 实验设置

我们的模型在 16K RichHF-18K 训练集上进行训练，并根据模型在 1K RichHF-18K 验证集上的表现调整超参数。超参数设置见补充材料。

评估指标 对于分数预测任务，我们重新使 用 了皮尔逊线性相关系数（PLCC）和斯皮尔曼秩相关系数（SRCC），它们是分数预测的典型评估指标[29]。对于热图预测任务，评估结果的直接方法是借用标准的显著性热图评估指标，如 NSS/KLD [5]。但是，这些指标不能直接应用于我们的情况，因为所有这些指标都假定地面实况热图不是空的；但在我们的情况下，地面实况可能是空的（*例如*，对于伪事实/似是而非热图，这意味着图像没有任何伪事实/似是而非）。因此，我们报告了所有样本的 MSE 值，以及空地面实况样本的 MSE 值，并报告了非空地面实况样本的显著性热图评估指标，如 NSS/KLD/AUC-Judd/SIM/CC[5]。对于错位关键词序列预测，我们采用标记级精度、再调用和 F1 分数。具体来说，精确度/重新调用/F1 分数是针对所有样本的错位关键词计算的。

基线 为了进行比较，我们对两个 ResNet-50 模型[18]进行了微调，分别使用多个全连接层和去卷积头来预测分数和热图。我们还使用了现成的 PickScore 模型[31]来计算 PickScores，并与我们的四个地面实况分数分别计算指标。我们使用现成的 CLIP 模型[40]作为基线来计算

	可信度		美学		文本图像对齐		总体	
	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑
ResNet-50	0.495	0.487	0.370	0.363	0.108	0.119	0.337	0.308
PickScore (现成的)	0.0098	0.0280	0.131	0.140	0.346	0.340	0.202	0.226
CLIP（现成的）	-	-	-	-	0.185	0.130	-	-
CLIP（微调）	0.390	0.378	0.357	0.360	0.398	0.390	0.353	0.352
我们的模型（多头）	0.666	0.654	**0.605**	**0.591**	**0.487**	**0.500**	**0.582**	0.561
我们的模型（增强提示）	**0.693**	**0.681**	0.600	0.589	0.474	0.496	0.580	**0.562**

表 1.测试集的分数预测结果。

	所有数据	$GT=0$	$GT>0$				
	MSE ↓	MSE ↓	CC ↑	KLD ↓	SIM ↑	NSS ↑	AUC-Judd ↑
ResNet-50	0.00996	**0.00093**	0.506	1.669	0.338	2.924	0.909
我们的（多头）	0.01216	0.00141	0.425	1.971	0.302	2.330	0.877
我们的（增强提示）	**0.00920**	0.00095	**0.556**	**1.652**	**0.409**	**3.085**	**0.913**

表 2.测试集上的似真热图预测结果。$GT=0$ 指空似然性热图，*即*没有伪差/似然性（995 个测试样本中有 69 个是空的），为地面实况。$GT>0$ 指有伪差/似是而非的热图，为地面实况。

(a) 提示：*游戏玩家在玩《英雄联盟》* (b) 提示：*一望无际的波浪状海洋* (c) 提示：机械蜜蜂飞入 (d) 提示：动漫堡垒*之夜角色*：*机械蜜蜂飞来* (d) 提示：*动漫堡垒之夜角色*。

夜晚的传说

可信度得分。

GT: 0.333, 我们的模型：0.410

总分

GT：0.417，我们的模型：0.4570.457

在五彩缤纷的夜空下 艺术性 电子产品 电机 电线 但是

粉彩绘画。可信度得分。

GT：1.0，我们的模型：

0.9790.979 总分。

GT 1.0, 我们的模型：0.848

*吨*液晶显示器。

文本图像比对得分。GT：0.583

，我们的模型：0.4080.408 美

学得分。

GT：0.75，我们的模型：0.7220.722

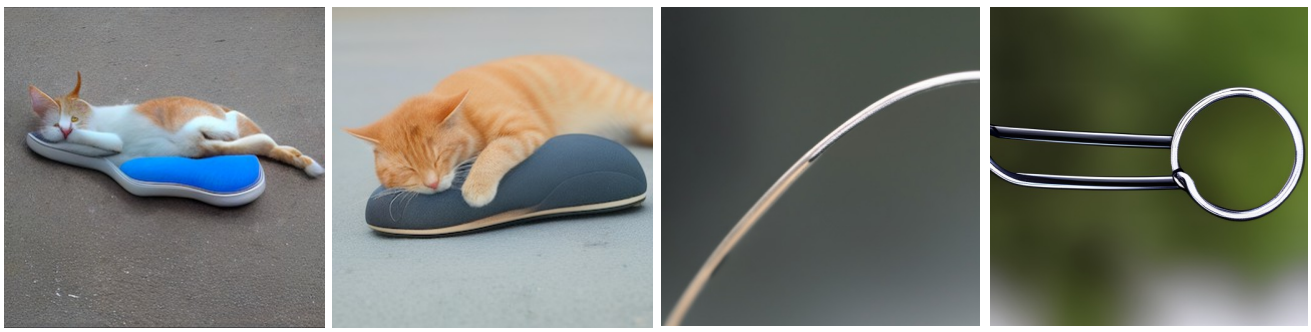
文本-图像配准得分。GT：1.0，

我们的模型：0.8970.897 美学得

分。

GT： 0.75，我们的模型： 0.7130.713

图 7.评分示例。"GT "是地面实况分数（三位注释者的平均分数）。

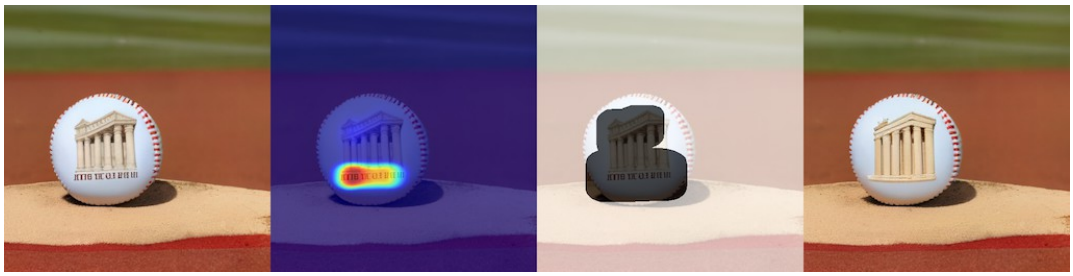


(a) 微调前的 Muse [6] (b) 微调后的 Muse [6] (c) 没有指导的 LD [42] (d) 美学指导后的 LD [42]

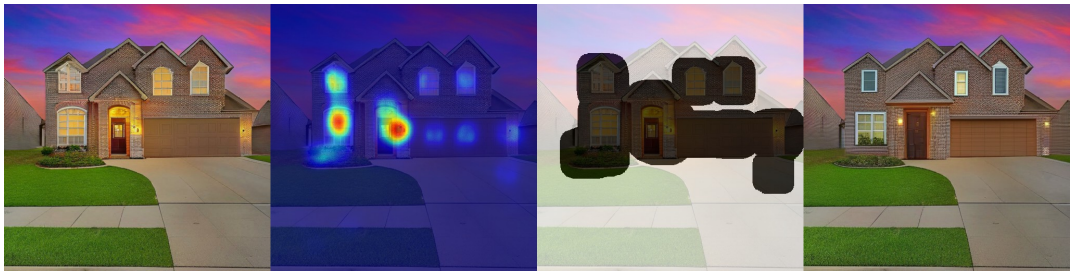
图 8.示例说明 RAHF 对生成模型的影响。(a-b）：Muse [6]生成的图像在微调之前和之后，用可信度分数过滤的例子，提示：*一只猫用一只鞋当枕头睡在地上*。(c-d）：未使用和使用美学分数作为分类器指导[2]对潜在扩散（LD）[42]进行分类的结果，提示：*一个回形针的微距镜头特写。*

	所有数据	*GT*= 0	*GT*> 0				
	MSE ↓	MSE ↓	CC ↑	KLD ↓	SIM ↑	NSS ↑	AUC-Judd ↑
CLIP 梯度	0.00817	0.00551	0.015	3.844	0.041	0.143	0.643
我们的模型（多头）	**0.00303**	0.00015	0.206	**2.932**	0.093	1.335	0.838
我们的模型（增强提示）	0.00304	**0.00006**	**0.212**	2.933	**0.106**	**1.411**	**0.841**

表 3.测试集上的文本错位热图预测结果。*GT*= 0 指空错位热图，*即*没有错位（995 个测试样本中有 144 个是空的），为地面实况。*GT*> 0 指有错位的热图，为地面实况。



(a) 提示：*一个封面上印有帕台农神庙的棒球，坐在投手丘上*



(b) 提示：*一张位于宁静街区的漂亮现代房屋的照片。房子由砖砌成，前廊很大。草坪修剪整齐，后院宽敞。*

图 9.使用 Muse [6] 生成模型进行区域内绘。这 4 张图从左到右依次为：Muse 提供的带有人工痕迹的原始图像、我们的模型预测出的不可信度热图、对热图进行处理（阈值化、扩张）后得到的遮罩，以及 Muse 使用遮罩进行区域内绘后得到的新图像。

	精确度	回收率	F1 分数
多头	**62.9**	33.0	43.3
增强提示	61.3	**34.1**	**43.9**

表 4.测试集的文本错位预测结果。

偏好	≫	>	≈	<	≪
百分比	21.5%	30.33%	31.33%	12.67%	4.17%

表 5.**人工评估结果：微调后的 Muse ��原始 Muse 模型偏好对比**：微调后的 Muse 明显更好 (≫)，稍好 (>)，差不多 ( )，稍差 ( )，明显更差 ( ) 的例子百分比。
相同 (≈)、稍差 (<)、明显差 (≪) 的例子所占百分比。
比原来的 Muse 差很多。数据是从 6 个人进行的一次随机调查中收集的。
的调查中收集了 6 个人的数据。

由于 CLIP 余弦相似度旨在反映图像和提示之间的对齐情况，因此我们利用图像和文本嵌入的余弦相似度来计算文本-图像对齐度量。此外，我们还利用训练数据集对 CLIP 模型进行了微调，以预测四种类型的分数。对于错位热图预测，我们使用 CLIP 梯度图 [46] 作为基准。

4.2.2 RichHF-18K 测试集上的预测结果

定量分析 我们的模型在 RichHF-18K 测试集上对四种细粒度评分、可信度热力图、错位热力图和错配关键词序列进行预测的实验结果见表 1、表 2 和表 3。1, Tab.2, Tab.3 和 Tab.4 中分别列出。

在 Tab.在表 1 和表 3 中，我们提出的模型的两个变体都明显优于 ResNet-50（或文本图像配准得分 CLIP）。在表 1 和表 3 中，我们提出的模型的两个变体都明显优于 ResNet-50（或文本-图像配准得分 CLIP）。然而，在表 2 中，我们的多头模型的表现却不如 ResNet-50。然而，在表 2 中，我们模型的多头版本表现不如 ResNet-50，但我们的增强提示版本却优于 ResNet-50。主要原因可能是，在多头版本中，没有在提示中增强预测任务，所有七个预测任务都使用相同的提示，因此所有任务的特征图和文本标记都是相同的。要在这些任务中找到一个很好的平衡点可能并不容易，因此一些任务（如伪造/似是而非热图）的性能会变差。不过，在将预测任务增强为提示任务后，特征图和文本标记就能适应每个特定任务，并取得更好的效果。

每个特定任务，从而获得更好的结果。此外，我们还注意到，错位热图预测的结果通常比工件/似真性热图预测的结果更差，这可能是因为错位区域的定义不够清晰，因此注释可能会更嘈杂。

定性示例 我们展示了一些模型预判的示例，如似是而非热力图（图 5），我们的模型可以识别出具有伪事实/似是而非的区域；以及错位热力图（图 6），我们的模型可以识别出与提示不符的物体。图 7 显示了一些示例图像及其地面实况和预测得分。更多示例见补充材料。

5. 从丰富的人类反馈中学习

在本节中，我们将研究预测的丰富人类反馈（*如*分数和热图）是否可用于改进图像生成。为了确保 RAHF 模型的收益在生成模型家族中具有普遍性，我们主要使用 Muse [6] 作为目标模型进行仿真，该模型基于屏蔽变换器架构，因此不同于 RichHF-18K 数据集中的稳定扩散模型变体。

利用预测分数对**生成模型进行微调** 我们首先说明利用 RAHF 分数对**生成模型**进行微调可以改进 Muse。首先，我们使用预先训练好的 Muse 模型为 12,564 个提示（提示集是通过 PaLM 2 [1, 11] 和一些种子提示创建的）中的每个提示生成八幅图像。我们会预测每张图片的 RAHF 分数，如果每张提示图片的最高分高于固定阈值，就会被选中作为微调数据集的一部分。然后再利用该数据集对 Muse 模型进行微调。这种方法可以看作是简化版的直接偏好优化 [15]。

图 8(a)-(b) 展示了利用我们预测的可信度得分（threshold=0.8）对 Muse 进行微调的一个示例。为了量化 Muse 微调的收益，我们使用了 100 个新提示来生成图像，并要求
6 名注释者分别对原始缪斯和微调缪斯的两幅图像进行并排比较（以确定可信度）。注释者从五种可能的回答中进行选择（图像 A 明显/略好于图像 B、基本相同、图像 B 略好于/明显好于图像 A），但不知道图像 A/B 是由哪个模型生成的。表 5 中的结果表明，经过微调的图像 A/B表 5 中的结果表明，使用 RAHF 似真性分数的微调缪斯图像的伪影/似真性明显少于原始缪斯图像。

此外，在图 8(c)-(d)中，我们展示了一个使用 RAHF 美学分数作为 Latent Diffusion 模型[42]的分类指导的例子，这与 Bansal 等 人[2]的方法类似，表明每个细粒度分数都能在不同方面改善图像的美学效果。这表明，每一个细粒度分数都能改善通用模型/结果的不同方面。

利用预测的热图和分数进行**区域涂抹** 我们证明，模型预测的热图和分数可用于进行区域涂抹，以提高生成图像的质量。对于每幅图像，我们首先预测不可信度热图，然后通过处理热图（使用阈值和扩张）创建遮罩。在屏蔽区域内应用 Muse Inpainting [6]，生成与文本提示相匹配的新图像。生成多张图片后，根据 RAHF 预测的最高可信度得分选择最终图片。

在图 9 中，我们展示了几种内绘结果，以及预设的可信度热图和可信度分数。如图所示，内绘后生成的图像可信度更高，伪影更少。这再次表明，我们的 RAHF 能够很好地通用于生成模型与用于训练 RAHF 的图像截然不同的图像。更多详情和示例请参阅补充材料。

6. 结论和局限性

在这项工作中，我们贡献了 RichHF-18K，这是第一个用于图像生成的丰富的人类反馈数据集。我们设计并训练了一个多模态变换器来预测丰富的人类反馈，并演示了一些利用丰富的人类反馈生成图像的实例。

虽然我们的一些结果令人兴奋和充满希望，但我们的工作还存在一些局限性。首先，模型在不对齐热图上的表现比在不可信度热图上的表现要差，这可能是由于不对齐热图中的噪声造成的。如何标注某些不对齐情况（如图像上不存在的物体）有些含糊不清。提高错位标注质量是未来的发展方向之一。其次，收集更多关于 Pick-a-Pic（稳定扩散）以外的生成模型的数据，并研究它们对 RAHF 模型的影响也是很有帮助的。此外，虽然我们提出了利用我们的模型改进 T2I 生成的三种有前途的方法，但还有无数其他方法可以利用丰富的人类反馈进行探索，*例如*，如何利用预测的热图或分数作为奖励信号，通过强化学习对生成模型进行微调，以及如何利用预测的热图作为加权图，或如何在从人类反馈中学习时利用预测的不对齐序列来帮助改进图像生成等。我们希望 RichHF-18K 和我们的初始模型能激励人们在未来的工作中探索这些研究方向。

参考文献

[1] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin John- son, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri、