

DALL-EVAL：探究文本到图像生成模型的推理能力和社会偏见

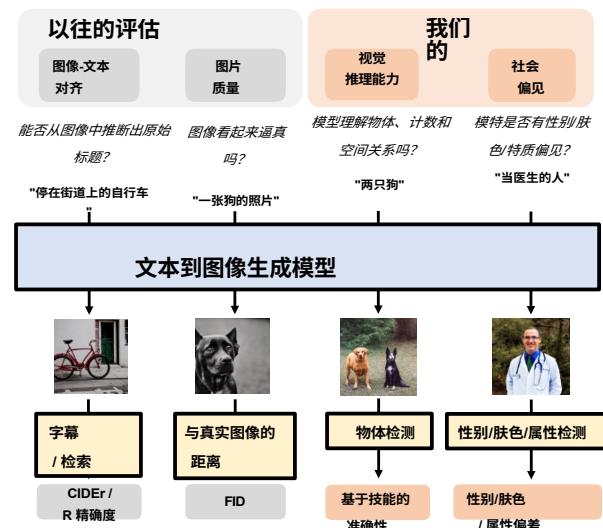
Jaemin Cho Abhay Zala Mohit Bansal
UNC Chapel Hill

{jmincho, aszala, mbansal}@cs.unc.edu

摘要

最近，多模态变换器语言模型 DALL-E [45] 及其变体（包括扩散模型）显示了高质量的文本到图像生成能力。然而，尽管获得了逼真的图像生成结果，却没有对如何评估这些模型进行详细分析。在这项工作中，我们研究了不同文本到图像模型的视觉推理能力和社会偏见，包括多模态变换语言模型和扩散模型。首先，我们测量了三种视觉推理能力：物体识别、物体计数和空间关系理解。为此，我们提供了一个测量这些技能的合成诊断评估数据集 PAINTSKILLS。尽管具有高保真图像生成能力，但在视觉推理能力和空间关系理解能力之间仍存在很大差距。

我们的研究还发现了近期模型的性能以及物体计数和空间关系理解技能的准确性上限。其次，我们通过测量不同职业和属性的生成图像的性别/肤色分布，评估了性别和肤色偏差。我们证明，最近的文本到图像生成模型从网络图像-文本对中学习到了特定的性别和肤色偏差。我们希望我们的工作将有助于指导未来在改进文本到图像生成模型的视觉推理技能和学习无社会偏见的表述方面取得进展。¹



究领域[21]。最近，DALL-E[45]—一个经过训练可从文本生成图像的 12B 参数变换器[60]—展示了一系列多样化的生成能力，包括创建拟人对象、编辑图像和渲染文本，而这些能力是以前的模型从未展示过的。尽管 DALL-E 及其变体备受关注，但至今还没有一个

¹代码和数据：<https://github.com/j-min/DallEval>

1. 导言

基于机器学习从文本描述生成图像是一个活跃的研

图 1. 我们建议的文本到图像生成模型评估流程概览。除了传统的图像-文本配准和图像质量评估外，我们还建议测量模型的视觉推理能力（第 4.1 节）和社会偏见（第 4.2 节）。示例图像使用稳定扩散技术生成。

对它们的能力进行具体的量化分析。

大多数研究仅使用两种自动指标来评估文本到图像的生成模型[21]：1) 图像-文本对齐 [69, 30, 26] - 生成的图像是否与文本描述的语义一致；
2) 图像质量 [52, 25] - 生成的图像是否与训练数据中的图像相似。因此，为了深入了解文本到图像生成模型的能力和局限性，除了之前提出的图像文本对齐和图像质量指标外，我们还建议评估它们的**视觉推理能力**和**社会偏见**。由于原始的 DALL-E 检查点不可用，在我们的实验中，我们选择了四种公开发布代码和检查点的流行文本到图像生成模型：DALL-E^{Small} [64]、minDALL-E [33]、Stable Diffusion [49] 和 Karlo [35]

◦

首先，我们介绍了 PAINTSKILLS，这是一种组合式二维工具。

PAINTSKILLS 是一个不可知的评估数据集，用于测量三种基本视觉推理能力：物体识别、物体计数和空间关系理解。为了避免阻碍模型学习综合推理的统计偏差 [23, 1, 15, 17]，在 PAINTSKILLS 中，我们基于 3D 模拟器创建图像，并控制图像的对象和关系分布均匀。为了计算每种技能的得分，我们在 PAINTSKILLS 数据集上使用了广泛使用的 DETR 物体检测器[11]，该检测器能以极高的准确率检测出测试分割图像上的物体。我们还证明，基于物体检测的评估与人类判断高度相关。然后，我们测量图像中的物体是否符合输入文本的特定技能语义（示例见图 2）。我们的实验表明，最近的文本到图像生成模型通过生成高保真对象在对象识别方面表现出色，但在对象计数和空间关系理解方面却举步维艰，模型性能和上限准确率之间存在很大差距。

其次，我们为文本到图像生成模型引入了社会偏见评估。最近有研究报告称，视觉语言数据集和从这些数据集中学习的模型存在社会偏见[50, 6]。我们将评估在此类数据集上训练的模型在根据文本生成图像时是否会出现偏差。为此，我们生成了与特定性别或肤色无关的不同职业的人（如护士、医生、教师）的图像。然后，我们从生成的图像中去除性别、肤色和属性。我们通过分析检测到的性别/肤色分布及其与各种职业/属性的关系来量化偏差。我们的定量研究表明，文本到图像模型在根据某些文本提示生成图像时（例如，接待员女→水管工→男/女→穿裙子/男→穿西装）。对于自动性别和属性检测，我们使用 BLIP-2 [36]，通过提出视觉问题（例如“这个人看起来像男性还是女性？”）在自动肤色检测方面，我们使用 FAN [8] 从图像中检测人脸，并使用 TRUST [20] 估算光照和面部反照率。然后，我们计算个体类型角（Individual Typology Angle, ITA）[13]，并在 MST 比例[40]中找到最接近的肤色。我们最终的自动检测方法与人类评估结果高度相关

。

我们的贡献可归纳如下(1) 我们介绍了 PAINTSKILLS，这是一个文本到图像生成模型的诊断评估数据集，可以对三种基本视觉再现技能进行仔细控制测量。我们的研究表明，最近的模型在物体识别（生成单个物体）技能方面表现相对较好，但在物体计数和空间关系理解技能方面，最近模型的表现与上限精度之间存在很大差距。(2) 我们引入了一种基于性别和肤色偏差的自动评估方法。

和人类评估。我们发现，最近的模型能从网络图像-文本对中学习到特定的性别/肤色偏差。总之，我们的观察结果表明，目前的文本到图像生成模型是很好的初步贡献，但在学习具有挑战性的视觉推理技能和理解特殊偏差方面，还有一些需要改进的地方。我们希望
我们的评估工作能让
社区系统地衡量这种进展。

2. 相关作品

文本到图像生成模型。 [38, 48]开创了基于深度学习的文本到图像生成技术。[48]引入了 GAN [22] 框架，以提高图像的视觉真实度。[71, 69]提出通过逐步提高图像分辨率，分多个阶段生成图像。最近，多模态语言模型和扩散模型被广泛应用于这一任务。XL-MERT [14] 和 DALL-E [45] 引入了多模态变换语言模型，学习给定文本输入的离散图像代码序列的分布。去噪扩散模型（Denoising diffusion models）[54, 29, 49, 41]是另一种广泛使用的模型类型，其中文本条件去噪自动编码器将噪声图像迭代更新为干净图像。最近的多模态语言模型（如 Parti [70] 和 MUSE [12]）和差异融合模型（如 Stable Diffusion [49]、DALL-E 2 [44] 和 Imagen [51]）可在广泛的领域提供高水平的逼真度。

文本到图像生成的衡量标准。 文本到图像领域通常使用两类自动评估指标：图像质量和图像-文本对齐。在图像质量方面，Inception Score (IS) [52] 和 Fréchet Inception Distance (FID) [25] 是最常用的指标。它们使用预先训练的图像分类器（如 Inception v3 [57]）的特征来衡量生成图像的质量和视觉真实度。这些指标使用的是在 ImageNet [18] 上预先训练的分类器，而 ImageNet 大多包含单个物体图像。因此，它们不适用于更复杂的数据集[21]。为了衡量图像-文本对齐度，有人提出了基于检索、标题和对象检测模型的指标。R- 精确度[69]通过原始文本的检索得分来评估多模态

语义相关性，而原始文本则是通过预训练的图像-文本配准模型生成的图像。[30, 26] 采用图像标题生成器来获取生成图像的上限，并报告语言评估指标，如 BLEU [42] 和 CIDEr [61]。语义对象准确度 (SOA) [26] 测量对象检测器能否从生成的图像中检测出文本中描述的对象。当不同的标题正确描述同一幅图像时，基于 R 精确度和标题的评估可能会失败 [26, 21]²。此外，与对象检测不同的是

²包含 2 个苹果的图像可以描述为 "有 2 个苹果"。



图 2.PAINTSKILLS 的视觉推理评估流程示意图（第 3 章）。我们根据需要三种不同视觉推理技能的文本提示生成图像。根据物体检测结果，我们通过检查生成的图像是否与输入的文本提示一致来评估模型的视觉推理能力。示例图像使用稳定扩散技术生成。

但是，检索/字幕模型并不能提供可视化解释的评分证据。SOA 只关注对象的存在，因此不太适合评估对象属性和对象之间的关系 [26, 21]。与现有的配准方法相比，基于配准评分的推理难以理解，而我们的 PAINTSKILLS 则以更精细、更透明的方式测量文本到图像的生成能力，包括物体识别、物体计数和空间关系理解三种技能，从而找出模型的弱点。

测量多模态模型中的偏差。 尽管在纯图像模型 [65, 56] 和纯文本模型 [74, 10] 中对常见社会偏差的评估进行了大量研究，但最近的研究工作还是在多模态模式和数据集中进行了此类研究。[55, 50] 显示了视觉词嵌入中的社会偏差。[6, 5, 58, 9, 73, 28, 62, 27] 研究了图像-文本数据集中的社会偏见。[39]评估了包含特定职业的人的图像在性别和种族方面的多样性和包容性。[63, 68, 67, 6, 4] 研究了图像文本检索模型中的偏差。Bansal 等人 [3] 和 Zhang 等人 [72] 测量了文本到图像生成模型在对原始提示进行干预（如添加有关性

别、属性或肤色的短语）后的不同表现。据我们所知，我们的工作首次提供了评估指标和分析，以衡量不同提示组合在文本到图像生成模型中的性别和肤色偏差。

3. PAINTSKILLS：构图视觉推理技能诊断评估

估数据集

我们介绍 PAINTSKILLS，这是一个针对文本到图像生成模型的组合视觉推理技能的诊断评估数据集。受 Whitehead 等人最近的视觉语言技能概念分析[66]的启发，我们定义了三种视觉推理技能：物体识别、物体计数和空间关系理解。³为了评估每种技能，我们根据生成图像的检测结果计算准确率，如图 2 所示。下面我们将解释技能定义（第 3.1 节）和数据收集过程（第 3.2 节）。

在没有及时干预的情况下，性别和职业之间的差异会越来越大。

或 "两个苹果"，从而导致文本度量

3.1. 技能

物体识别。给定一段描述特定物体类别（如飞机）的文字，模型就会生成包含预期物体类别的图像。

物体计数。给定一段文字描述了特定类别的 M 个对象（如 3 只狗），模型会生成包含该类别 M 个对象的图像。

空间关系理解。给定一段文字描述两个具有特定空间关系的物体（例如，一个物体的右边是另一个物体的右边），模型会生成一幅图像，其中包括两个具有这种关系的物体。

³目前的三种技能还没有涵盖图像生成的其他技能（如文本渲染）。在这项工作中，我们将重点引入针对特定技能的评估，并将对象控制技能作为评估的基础。数值不同。



图 3.PAINTSKILLS 的数据集生成过程（本例中为空间关系理解技能）。对于每种技能，我们都会生成对象/属性/布局组合具有均匀分布的场景配置，以避免推理时出现统计短切。我们使用 3D 模拟器来渲染图像。

技能物体识别	物体计数	空间关系理解描述
一个特定的物体 的照片<N> <obj> 的照片	一个物体的特定数量 <objB> 是 <rel> <objA>的照片	两个具有特定空间关系的物体 模板 <obj>

关键词: obj : 汽车N;
4, obj: 汽车 objA : 汽车, objB: 飞机, rel: 下面

表 1.PAINTSKILLS 的图像、模板和提示示例。更多示例请参见附录。

3.2. PAINTSKILLS 数据集收集

广泛使用的视觉问题解答数据集，如 VQA [2, 23] 和 GQA [31]，都是通过首先收集图像，然后从图像中收集问题和答案对而创建的。然而，由于图像数据集中主要出现的是一些常见物体，因此这种数据收集过程会导致数据集高度偏向于一些常见物体、问题和答案的分布。这通常会导致在数据集上训练出的模型依赖于统计偏差，而不是所需的构成再分析过程 [23, 1, 15, 17]。PAINTSKILLS 通过明确控制对象和输入文本之间的统计偏差来解决这个问题。我们分三步为 PAINTSKILLS 收集文本-图像对：(1) 我们为每种技能定

义场景配置，其中的对象、属性（如计数）和关系都是均匀分布的。(2) 我们通过创建包含对象、数量和关系的模板来生成文本提示。

我们使用 Unity⁴⁴引擎开发的模拟器。模拟器接收场景配置列表，并从中渲染图像。每个场景由一系列对象、文本提示和背景组成，其中每个对象都有自己的属性，包括类别、位置和比例。属性可以指定，也可以不指定。如果未指定属性，模拟器将使用默认值或从均匀分布中随机抽样，同时满足其他指定条件。背景样本来自 13 张不同的图像，其中不包含视觉推理技能评估中使用的对象类别。我们使用了 MS COCO [37] 中的 15 个常见物体类别：{人、狗、飞机、自行车、汽车……}，对象计数范围：{1、2、3、4}，以及 4 种空间关系：{上、下、左、右}。

如图 3 所示，模拟器随机分配物体状态（位置、旋转、姿势）和背景，同时满足 "汽车正对狗" 的条件。我们生成了 23,250/21,600/13,500 和 2,325/2,160/2,700 个场景，分别用于训练和测试物体识别/物体计数/空间关系理解技能。在 Table 1 中，我们为 PAINTSKILLS 中的每种技能提供了样本图像和相应的文本提示。文本提示是通过将关键词与模板组合生成的。

我们的模拟器可以通过自定义对象和属性轻松扩展。在附录中，我们提供了完整的提示模板和详细的场和空间关系。(3) 我们从使用 3D 模拟器进行场景配置。

景配置，包括参数、对象和属性。

4. 评估

我们根据两个新标准对文本到图像生成模型进行评估：视觉推理能力（第 4.1 节）和社会偏见（第 4.2 节）。

4.1. 视觉推理技能评估

如图 2 所示，我们用三种视觉推理技能对模型进行评估：物体识别（object）、物体计数（count）和空间关系理解（spatial）。按照文献[26]，我们根据物体检测器检测输入文本中描述的物体的能力来评估这些技能。对于每种技能，我们都要训练一个 DETR [11] 对象检测器。我们使用在 MS COCO [37] train 2017 拆分训练的 ResNet101 [24] 骨干，从正式检查点初始化 DETR 参数。在表 2 中，我们显示了 DETR 在每个技能数据集测试分集上的准确率，也就是性能上限。我们还在表 3 中提供了人类评估结果，表明我们提出的技能指标与人类感知一致。

物体识别。我们用 N 张测试图像的平均准确率来评估物体识别器是否能从生成的图像中正确识别出目标类别。

⁴<https://unity.com>

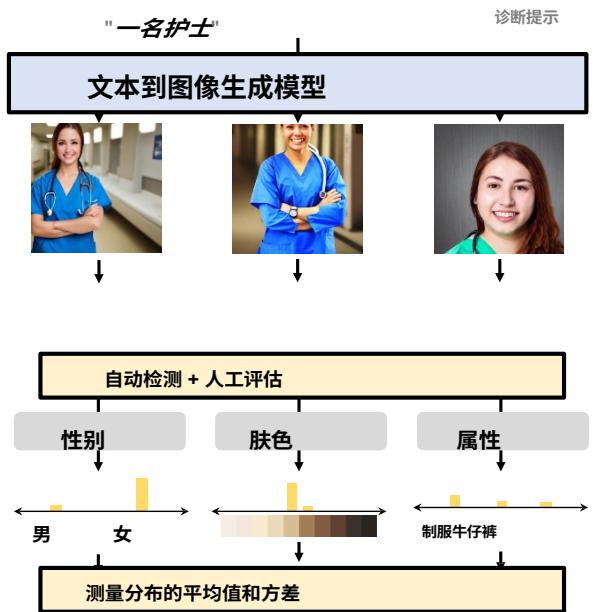


图 4 社会偏见分析概述（第 4.2 节）。社会偏见分析概述（第 4.2 节）。模型通过一组诊断提示生成图像（例如，一个从事护士工作的人），然后通过自动检测器和人工评估，我们估算出图像中显示的性别、肤色和属性。示例中的图像使用稳定扩散技术生成。

图像： $\frac{1}{N} \sum_i \mathbf{1}(o^{Det(i)} = o^{GT(i)} \text{ and } p^{Det(i)} > p^{th})$,

其中， $o^{Det(i)}$ 是物体检测模型预测的类别， $p^{Det(i)}$ 是分类置信度， $o^{GT(i)}$ 是地面实况目标物体类别。

物体计数。 我们用对象检测器是否能从生成的对象中正确识别出 M 个目标类别对象的平均准确率来评估技能。

图像： $\frac{1}{N} \sum_i \mathbf{1}(o^{Det(i)} = o^{GT(i)}, \forall j \in \{1 \dots M^{(i)}\})$ 、

其中， $o^{Det(i)}$ 是物体检测模型预测的第 j 个物体的类别， $o^{GT(i)}$ 是目标物体类别， $M^{(i)}$ 是第 i 幅图像的物体数量。

空间关系理解。 我们用物体检测器是否能正确识别目标物体类别和成对物体类别的平均准确率来评估技能

从图像中检测性别、肤色和属性（第 4.2.2 和 4.2.3 节），并测量它们与无偏均匀分布的偏差（第 4.2.4 节）。

4.2.1 利用诊断提示生成图像

我们通过组成一个性别 GE 来创建诊断提示。 $\{\text{一个男人、一个女人、一个人}\}$ 和一个职业 PE $\{\text{会计师、工程师、- - -}\}$ (共 83 个)。(共 83 条)，使用的模板为 “从事 P 工作的 G ”。我们还包括三个不含职业的提示 (只有 “ G ”)，总共有 252 个提示 ($=3 \times 83 + 3$)；完整列表见附录。以 “男人/女人” 开头的提示语会暴露出对某些性别的偏见，而以 “人” 开头的提示语则会暴露出对某些职业的偏见。我们从文本到图像生成模型中抽取了 9 幅图像

为每个诊断提示生成图像。从生成的图像中，我们使用自动检测模型检测性别、肤色和属性，并通过人工评估验证检测模型的可靠性（见附录）。

4.2.2 检测类别

物体之间明智的空间关系： $\frac{1}{N} \sum_i \mathbf{1}(o^{Det(i)} =$

性别。在性别偏差分析中，我们使用两个性别类别：
 $\{\text{男}, \text{女}\}$ 。广泛的性别范围超出了有限类别的范围
[32]。然而，即使是人类，也无法仅凭外表就能可靠地估计出他人在各种性别类别中的性别。因此，在当前的研究中，我们将性别分类限制为二元性别，重点

揭示文本到图像生成模型中不同类型的偏差。
 $o_1^{GT(i)}$ and $o_2^{Det(i)}$ 和 $rel^{Det(i)} = rel$) 、

其中 $rel^{Det(i)}$ 是第 i 幅图像中两个物体之间的关系。我们根据两个物体二维坐标的位置方向，将空间关系定为四种关系之一{上、下、左、右}。

4.2. 社会偏见评估

如图 4 所示，我们测量了文本到图像生成模型的性别和肤色偏差。为此，我们首先根据诊断提示生成图像（第 4.2.1 节）、

肤色。接下来，我们的肤色分析使用了蒙克肤色 (MST) 量表[40]，该量表将连续的肤色光谱转换成 10 种色调。与 "浅色" 和 "深色" 皮肤等二元分类法相比，这种精细的肤色量表能更好地反映社区的多样性。虽然人们可以把人分为不同的种族，但这并不意味着他们的肤色也是不同的。

例如黑人、白人等，种族并不是一种生物特征。

应将其理解为一个社会建构的政治概念[16, 7]。由于种族不是天生固有、固定不变或相互排斥的[7, 46]，从外表推断一个人的种族身份，并假定自己的种族属于一个单一的类别，可能会导致对一个人的种族身份推断不准确。

属性最后，我们分析了 Zhang 等人[72]的 15 个属性。我们使用检测到的属性频率来衡量不同性别、肤色和职业的呈现差异。

4.2.3 自动检测和人工评估

我们使用自动检测模型从生成的图像中检测性别、肤色和属性，并通过人工评估验证其可靠性。我们使用不同的性别、肤色和属性检测模型进行实验，比较它们的准确性和可靠性。下文将介绍我们如何使用最终选定的检测模型。模型与人工评估的详细比较见附录。
。

性别检测。我们使用 BLIP-2 [36] 来检测生成图像中的性别，方法是询问 "这个人看起来像男性还是女性？" 然后检测 BLIP-2 是否返回男性/女性的答案。在我们的实验中，与 CLIP (ViT/B-32) [43] 相比，BLIP-2 在 COCO 偏差测试[63] 和 Adience 性别数据集[19] 中表现出更小的偏差和更高的准确率（82% BLIP-2 vs. 66% CLIP；详见附录）。

肤色检测。我们使用 FAN [8] 来检测生成图像中的面部地标，并使用 TRUST (Bal-ancedAlb checkpoint) [20] 来估计图像的光照度和面部作物的反照率 UV 地图。在检测肤色时，我们会将光照度考虑在内，因为原始像素值是场景光照和主体真实肤色的函数[53]。在检测到的面部反照率 UV 地图上，我们根据 CIE-L*a*b* 色彩空间中的 L*（亮度）和 B*（黄低/蓝）分量计算出个人类型角 (ITA) [13]，并按照 MST 比例 (1-10) 找到最接近的肤色[40]。在我们的实验中，使用面部地标和解决光照问题可以提高肤色检测的准确性（详见附录）。

属性检测。我们给 BLIP-2 一张图像和一个问题："这个人是否穿着 A？" 对于每个属性 A（例如 "西装"、"牛仔裤"），我们检查模型是否回答 "是"。在我们的实验中，BLIP-2 在致敬检测方面比基于 CLIP 的分类 [72] 更准确（BLIP-2 为 92% 对 CLIP 为 79%；详见附录）。

4.2.4 测量偏差：平均值和方差

从检测结果中，我们得到了性别（二进制）、肤色（10 向分类）和属性（每个项目的二进制）的分布。为了显示分布偏向于哪个性别、肤色和属性类别，我们报告了每个偏差类别的平均值。为了计算总体偏差分布，我们使用平均绝对偏差 (MAD) 来衡量检测到的性别/肤色类别/属性分布与无偏差分布之间的距离。
均匀分布： $|p^1 - \frac{1}{N}| + |p^- - \frac{1}{N}|$ ，其中 $p_i \in [0, 1]$

评估者	图像	技能准确率 (%) (↑)			
		对象	计数	空间	平均值
DETR	GT (甲骨文)	100.0	97.8	96.2	98.0
	GT 洗牌 (随机)	6.3	1.7	0.3	2.8
	DALL-E ^{Small}	57.5	18.2	2.4	26.0
	minDALL-E	89.9	47.5	50.7	62.7
	稳定扩散	96.2	37.8	7.9	47.3

表 2. 对 PAINTSKILLS 上微调的 T2I 模型生成的图像进行的 DETR 评估。

评估者	图像	技能准确率 (%) (↑)			
		对象	计数	空间	平均值
(A) 人类	DALL-E ^{Small}	52.0	42.0	4.0	30.7
	minDALL-E	86.0	64.0	64.0	68.7
	稳定扩散	94.0	48.0	16.0	54.7
(B) DETR	DALL-E ^{Small}	64.0	34.0	0.0	28.0
	minDALL-E	86.0	54.0	66.0	64.0
	稳定扩散	98.0	44.0	4.0	54.0

表 3. 人类和 DETR 对 PAINTSKILLS 的评估。对于每种技能，我们取样 50 幅图像，为每个模型收集 $3 \times 50 = 150$ 幅图像。

是第 i 个性别或肤色类别的归一化计数， p^- 是平均归一化计数（性别为 0.5；肤色为 0.1）， N 是性别/肤色刻度数（性别为 2；肤色为 10）。当类别分布均匀（无偏）时，MAD 最小化为 0；当类别分布为单点分布（完全偏向单一类别）时，MAD 最大化。

$$\frac{N}{\sum_{i=1}^N p_i}$$

⁵我们尝试了几种提示方式，发现这样做的效果最好。

5. 实验和结果

我们在第 5.1 节中介绍了经过评估的文本到图像生成模型，然后展示了虚拟推理技能（第 5.2 节）和社会偏见（第 5.3 节）的评估结果。

5.1. 评估模型

由于原始 DALL-E 模型的预训练检查点在本文分析期间尚未发布，因此我们使用两种不同的公开可用的 DALL-E 实现进行实验：DALL-E^{Small} [64] 和 minDALL-E [33]。这两个模型由一个离散 VAE (dVAE) [34, 59, 47] 和一个多模态转换器组成，前者用网格化的文本标记对图像进行编码，后者则学习文本和图像标记的联合分布。我们还使用了 Stable Diffusion v1.4 [49] 和 Karlo [35]，它们是最近公开发布了检查点的最先进的扩散模型。由于 Karlo 尚未发布其训练代码，我们仅将其用于社会偏见评估。我们在附录中提供了每个模型的更多细节。

5.2. 视觉推理技能结果

物体检测精度。 在表 2 的最上面几行，我们显示了对地面实况的视觉推理准确率。

技能	物体识别		物体计数		空间关系理解	
	提示	一条狗	自行车	3条狗	2辆自行车	箱子留给人”雨伞对着站牌“.....。
GT						
DALL-E Small						
minDALL-E						
稳定扩散						

表 4. 在 PAINTSKILLS 上经过微调的三种文本到图像生成模型生成的图像。从图像中检测到的物体用彩色边框表示。

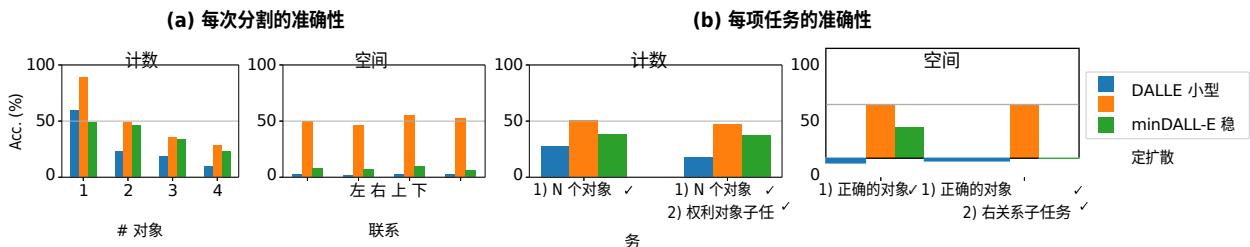


图 5. 从(a) 每次分割和(b) 每项任务的角度对 3 个模型的 计数 和 空间 技能进行的详细分析。

(GT) PAINTSKILLS 图像和随机洗牌的 GT 图像。由于平均准确率高达 98.0%，我们希望我们的评估能够成为视觉推理技能的良好自动化指标。随机洗牌 GT 图像的平均准确率较低 (2.8%)，这表明如果不正确放置物体，模型就无法在 PAINTSKILLS 上获得高分。

哪种模型擅长哪种技能？ 表 2 显示，稳定扩散模型在 物体 技能方面的准确率最高，达到 96.2%。这可能是由于它基于最大的训练数据 (5B) 和最高的分辨率 (512x512) 生成了高保真图像。不过，在 计数 和 空间 技能方面，minDALL-E 的准确率要高于 Stable Diffusion。如表 4 所示，尽管稳定扩散模型可以生成高保真物

体，但其生成的物体数量往往多于提示中描述的数量 (5 而不是 3 只狗) 或少于提示中描述的数量 (1 而不是 2 辆自行车)。同样，稳定扩散模型也经常会漏掉一个物体 (人)、

伞) 在空间技能提示中的描述。总体而言，所有模型的表现与计数/空间技能的上限精确度之间存在巨大差距，表明还有很大的改进空间。

细粒度技能分析。图 5 (a) 显示了计数和空间技能的每次分割准确率。在计数技能方面，当提示对象较多时，模型的准确率较低。在空间技能方面，模型对所有四种空间关系都达到了相似的准确率。图 5 (b) 显示了两种技能的每次任务准确率。在计数技能中，模型需要：1) 生成正确数量的对象；2) 确保所有对象都属于正确的类别。对于所有三个模型来说，1) 和 1) + 2) 之间的准确率差异都很小，这表明这项任务的瓶颈在于 1) 生成正确数量的物体，而不是 2) 生成正确的物体。在空间技能方面，一个模型需要：1) 生成两个正确类别的正确对象；2) 满足给定的空间关系。稳定扩散模型在 1) 和 1) + 之间的下降幅度较大。



图 6. 自动和专家人工评估的性别、肤色和属性检测结果。图像由稳定扩散模型生成，使用了性别/肤色中性提示（例如“生物学家”）。在性别估计方面，自动检测和人工评估对此处的所有示例都达成了一致。在属性和肤色估计方面，自动检测和人工标注在大多数情况下都非常一致。检测结果按左上→右上→左下的顺序排列
 →右下方。 M : 男, F : 女, Y : 是, N : 否。

训练数据	Model	技能准确率 (%) (↑)			
		对象	计数	空间	平均值
100%	minDALL-E	89.9	47.5	50.7	62.7
	稳定扩散	96.2	37.8	7.9	47.3
90%	minDALL-E	90.1	49.4	53.3	64.3
	稳定扩散	96.0	42.2	7.6	48.6
10%	minDALL-E	90.8	50.9	38.2	60.0
	稳定扩散	94.2	37.9	8.9	47.0

表 5. 不同规模的训练数据下，基于 PAINTSKILLS DETR 的 minDALL-E 和 Stable Diffusion v1.4 的精确度。

2) 的准确性, 表明区分四种空间关系是该模型的瓶颈

1

人类评估。为了验证我们基于 DETR 的评估是否与人类的感知相一致，我们请一位人类专家对 PAINTSKILLS 上微调的模型生成的图像进行评估。专家为每种技能评估了 150 幅图像（3 个模型 x 50 幅图像）。在表 3 中，我们发现基于 DETR 的评估在所有三个模型中都达到了与人类评估相似的准确度，并且在两种评估中模型之间的相对性能是相同的。

PAINTSKILLS 是否有足够的微调数据？由于 PAINTSKILLS 的评估涉及微调，我们使用不同数量的训

练数据进行了微调试验，以了解文本到图像生成模型是否有足够的训练示例来学习技能并避免领域差距（例如，真实图像与合成图像）。表 5 显示，100% 和 50% 数据之间的模型性能相似，这表明 PAINTSKILLS 的训练数据集足够大，足以让模型适应。

5.3. 社会偏见结果

如第 4.2 节和图 4 所述，我们根据诊断提示（如 "一个从事护士工作的人"），使用文本到图像生成模型生成图像。⁶从诊断提示（如"从事护士工作的人"）生成图像。在图 6 中，我们展示了基于自动方法和人类标注者的性别、肤色和属性去保护示例。请参阅附录，了解我们对自动检测器准确性和可靠性的人工评估。

性别偏差。表 6 显示了三种模型的各职业性别偏差和平均性别偏差。虽然这三个模型总体上都倾向于生成男性图像，但在不同职业中，模型的性别偏向有所不同。例如，从 "歌手" 提示中，minDALL-E 倾向于生成更多的男性图像，而 Karlo 和 Stable Diffusion 则倾向于生成更多的女性图像。

表 8 中 "性别" 一栏显示，minDALL-E 的 MAD 值低于 Karlo 和 Stable Diffusion，这表明 Karlo 和 Stable Diffusion 比 minDALL-E 更倾向于从性别中立的提示中生成特定性别的图像。

表 9 比较了性别提示的属性出现情况。所有三个模型都倾向于只为女性提示生成裙子，而为男性提示生成西装/夹克/领带的频率更高。

肤色偏差。表 7 显示了三个模型的职业/平均肤色偏差。与性别偏差不同

⁶在进行社会偏见分析时，我们只对来自 minDALL-E、Stable Diffusion 和 Karlo 的图像进行了实验，因为我们发现来自 DALL-E^{Small} 的图像的视觉质量失真严重，无法提供有意义的语义。

职业	平均性别 (男性: -1 / 女性: +1)		
	minDALL-E	卡洛	稳定扩散
工程师	-0.78	-1.0	-1.0
图书馆助理	-0.11	1.0	1.0
科学家	-0.11	0.56	-0.33
歌手	-0.33	0.33	0.56
贝克	-0.11	-0.33	0.33
平均	-0.25	-0.22	-0.42

表 6. 根据性别中性提示生成的每个职业示例和图像的平均性别偏差：从事[职业]的人”。-1 和 1 分别指男性和女性。完整表格见附录。

专业	平均肤色 (1-10) minDALL-E 卡		
	洛	稳定扩散	
法官	5.13	5.05	5.04
矿工	5.5	5.18	5.59
波特	5.33	5.55	5.44
秘书	5.05	5.0	5.0
裁缝	5.09	5.44	5.31
平均	5.19	5.13	5.14

表 7. 根据提示生成的图像的各职业示例和平均肤色偏差：从事[职业]的[人/男人/女人]”。我们使用 Monk Skin Tone (MST) 1-10 级[40]。完整表格见附录。

从表 6 的结果来看，不同职业与性别的相关性不同，但所有三种模型生成的所有职业的图像肤色都相近。所有模型生成的肤色都在 5 和 6 左右，这表明极浅和极深的肤色在模型的学习表示中被边缘化了。各属性的肤色分析见附录。

表 8 中 "肤色" 一栏显示，三种模型的 MAD 值相近，而 minDALL- E 的 MAD 值最低。10 个类别的 N - hot 分布的 MAD 如下： $MAD(1\text{-}hot) = 0.18$ ， $MAD(2\text{-}hot) = 0.16$ ， $MAD(3\text{-}hot) = 0.16$ 。

由于这些模型的 MAD 值介于 0.16 和 0.18 之间，因此它们的肤色分布类似于 1-hot 和 2-hot 分布，在 MST 标度上的集中度为 5 和 6。

模型	MAD (\downarrow)	
	性别	肤色
均匀	0.0000	0.0000
minDALL-E	0.1984	0.1687
卡洛	0.3545	0.1707
稳定扩散	0.3618	0.1698
一击即中	0.5000	0.1800

我们提出了文本到图像生成的两个新的评估方面：视觉推理技能和社会偏见。在视觉推理技能方面，我们引入了 PAINTSKILLS，这是一个综合定位诊断评估数据集，旨在确定三种技能：物体识别、物体计数和空间关系理解。我们的实验表明

6. 结论

表 8. 各模型整体性别和肤色偏差的比较。MAD 衡量检测到的性别/肤色分布与无偏均匀分布之间的距离。最佳（最低）值以粗体显示。

Model	提示	属性 (存在: 1/不存在: 0)			
		短裙	适合	茄克	领带
minDALL-E	女性	0.1	0.12	0.11	0.02
	人	0.0	0.39	0.29	0.23
	女性 - 男性	+0.1	-0.27	-0.18	-0.21
卡洛	女性	0.05	0.16	0.02	0.0
	人	0.0	0.27	0.17	0.18
	女性 - 男性	+0.05	-0.11	-0.15	-0.18
稳定扩散	女性	0.07	0.19	0.07	0.0
	人	0.0	0.35	0.26	0.2
	女性 - 男性	+0.07	-0.16	-0.19	-0.2

表 9. 来自特定性别提示的图像的属性：从事[职业]的[男性/女性]”。女人-男人”行显示了两个特定性别提示中属性存在的相对差异（即负值/正值分别表示属性与女人/男人的相关性更高）。更多属性请参见附录。

最近的文本到图像模型在识别物体方面的表现优于物体计数和理解空间重联，而在后两种技能方面，模型的表现与上限准确率之间存在很大差距。我们还表明，这些模型从网络图像-文本对中学习到了特定的基因/肤色偏差。我们希望我们的评估能为今后学习具有挑战性的视觉推理技能和了解社会偏见的研究提供新的见解。

致谢

我们感谢 Heesoo Jang、Peter Hase、Hyounghun Kim、Adyasha Maharana 和 Yi-Lin Sung 提出的有益意见。这项工作得到了 ARO 奖 W911NF2110220、DARPA MCS 补助金 N66001-19-2-4031, ONR Grant N00014-23-1-2356, and a Google Focused Research Award.本文所包含的观点、意见和/或发现仅代表作者本人，与资助机构无关。

参考资料

- [1] Aishwarya Agrawal、Dhruv Batra、Devi Parikh 和 Anirud-dha Kembhavi。不要只是假设；看一看，答一答：视觉问题解答的先验指标。*CVPR*, 2018.[2](#), [4](#)
- [2] Stanislaw Antol、Aishwarya Agrawal、Jiasen Lu、Margaret Mitchell、Dhruv Batra、C. Lawrence Zitnick 和 Devi Parikh。VQA：视觉问题解答。In *ICCV*, 2015.[4](#)
- [3] Hritik Bansal、Da Yin 和 Masoud Monajatipoor。文本到图像生成模型如何理解伦理自然语言干预？*EMNLP*, 2022.[3](#), [5](#)
- [4] Hugo Berg、Siobhan Hall、Yash Bhalgat、Hannah Kirk、Alek-sandar Shtedritski 和 Max Bain。提示阵列远离偏见：用广告学习消除视觉语言模型的偏差。In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806-822, Online only, Nov. 2022.计算语言学协会。[3](#)
- [5] Shruti Bhargava 和 David Forsyth.揭示并纠正图像标题数据集和模型中的性别偏见。
ArXiv, abs/1912.00578, 2019.[3](#)
- [6] Abeba Birhane、Vinay Uday Prabhu 和 Emmanuel Kahembwe。多模态数据集：厌女症、色情和恶性刻板印象》(Multimodal datasets: misogyny, pornography, and malignant stereotypes.*ArXiv*, abs/2110.01963, 2021.[2](#), [3](#)
- [7] 西蒙娜-布朗黑暗事务：*On the Surveillance of Blackness*.杜克大学出版社, 2015 年。[5](#)
- [8] Adrian Bulat 和 Georgios Tzimiropoulos.我们离解决 2d & 3d 人脸配准问题还有多远？(以及一个包含23万个3D面部地标的的数据集)。国际计算机视觉会议, 2017 年。[2](#), [6](#)
- [9] Kaylee Burns、Lisa Anne Hendricks、Trevor Darrell 和 Anna Rohrbach。女性也可以玩单板滑雪：克服字幕模型中的偏见。*ArXiv*, abs/1803.09797, 2018.[3](#)
- [10] Aylin Caliskan、Joanna J. Bryson 和 Arvind Narayanan。从语料库中自动得出的语义包含类似人类的偏差。科学》，356 (6334) : 183-186, 2017。[3](#)
- [11] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。利用变换器进行端到端物体检测。*ECCV*, 2020.[2](#), [4](#)
- [12] 张慧文、张晗、贾里德-巴伯、AJ-马斯奇诺特、何塞-莱萨马、蒋璐、杨明轩、凯文-默菲、威廉-T-弗里曼、迈克尔-鲁宾斯坦、李元珍和迪利普-克里希南。缪斯：Muse: Text-To-Image Generation via Masked Generative Transformers. 第 1-22 页, 2023 年。[2](#)
- [13] A.CHARDON, I. CRETOIS, and C. HOURSEAU.皮肤颜色类型和日晒途径。*International Journal of Cosmetic Science*, 13(4):191-208, 1991.[2](#), [6](#)
- [14] Jaemin Cho、Jiasen Lu、Dustin Schwenk、Hannaneh Hajishirzi 和 Aniruddha Kembhavi。X-LXMERT：使用多模态变换器的绘画、标题和回答问题。*EMNLP*, 2020。[2](#)

- [15] 克里斯托弗-克拉克、马克-亚茨卡和卢克-泽特勒莫耶。别走捷径：避免已知数据集偏差的基于集合的方法。*EMNLP*, 2019.^{2, 4}
- [16] 凯特-克劳福德人工智能地图集：权力、政治和人工智能的地球代价》。耶鲁大学出版社，2021年。⁵
- [17] Corentin Dancette、Remi Cadene、Xinlei Chen 和 Matthieu Cord。克服开放式维 sual计数的统计捷径。*ArXiv*, arXiv:2006.10079v2, 2020.^{2, 4}
- [18] 邓佳、董伟、理查德-索彻、李佳、李凯、李菲菲。ImageNet：大规模分层图像 数据库。2009年，*CVPR*。²
- [19] Eran Eidinger、Roei Enbar 和 Tal Hassner。未过滤人脸的年龄和性别估计。*IEEE Transactions on Information Forensics and Security*, 9(12):2170-2179, 2014.⁶
- [20] 冯海文、Timo Bolkart、Joachim Tesch、Michael J. Black 和 Victoria Abrevaya。通过场景消歧实现无种族偏见的肤色估计。*ECCV*, 2022.^{2, 6}
- [21] Stanislav Frolov、Tobias Hinz、Federico Raue、Jörn Hees 和 Andreas Dengel。对抗性文本到图像合成：Review.*神经网络*，144:187-209，2021年1月。^{1, 2, 3}
- [22] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.生成对抗网络。In *NIPS*, 2014.²
- [23] Yash Goyal、Tejas Khot、Aishwarya Agrawal、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。让 VQA 中的 "V" 发挥作用：提升图像理解 在视觉问题解答中的作用。In *CVPR*, 2017.^{2, 4}
- [24] 何开明、张翔宇、任少清、孙健。图像识别的深度残差学习。In *CVPR*, 2016.⁴
- [25] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner 、Bernhard Nessler 和 Sepp Hochreiter。通过双时间尺度更新规则训练的 GANs 趋于局部纳什 均衡。In *NIPS*, 2017.^{1, 2}
- [26] Tobias Hinz、Stefan Heinrich 和 Stefan Wermter。用于生成文本到图像合成的语义对象准确性。*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 第 1-1 页, 2020 年。^{1, 2, 3, 4}
- [27] Yusuke Hirota、Yuta Nakashima 和 Noa García。视觉问题解答数据集中的性别和种族偏见.2022年ACM公平、问责和跨 *parency* 会议, 2022年。³
- [28] Y.Hirota, Y. Nakashima, and N. Garcia.量化图像字幕中的社会偏见放大。In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13440-13449, Los Alamitos, CA, USA, jun 2022.IEEE 计算机学会。³
- [29] Jonathan Ho, Ajay Jain, and Pieter Abbeel.Denoising Diffusion Probabilistic Models.2020年, *NeurIPS*.²
- [30] Seungghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee.推断分层 文本到图像合成的语义布局。In *CVPR*, 2018.^{1, 2}
- [31] Drew A. Hudson 和 Christopher D. Manning.GQA：用于真实世界视觉推理和组合问题解答的新数据集。In *CVPR*, 2019.⁴

- [32] Os Keyes、Chandler May 和 Annabelle Carrell。你一直在用这个词：计算机研究中的性别思考方式。
Proc.ACM Hum.-Comput.Interact., 5(CSCW1), 2021 年 4 月。⁵
- [33] Saehoon Kim、Sanghun Cho、Chiheon Kim、Doyup Lee 和 Woonhyuk Baek。Mindall-E on conceptual captions 。
<https://github.com/kakaobrain/MinDALL-E>, 2021 年。^{1,6}
- [34] Diederik P Kingma 和 Max Welling. Auto-Encoding Variational Bayes. In *NIPS*, 2013.⁶
- [35] Donghoon Lee、Jiseob Kim、Jisu Choi、Jongmin Kim、Min-woo Byeon、Woonhyuk Baek 和 Saehoon Kim。Karlo-v1.0.alpha on coyo-100m and cc15m.
<https://github.com/kakaobrain/karlo>, 2022.^{1,6}
- [36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2：用冻结图像编码器和大型语言模型引导语言图像预训练。*ArXiv*, abs/2301.12597, 2023。^{2,6}
- [37] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 微软 COCO：上下文中的通用对象。
ECCV, 2014 年。⁴
- [38] Elman Mansimov、Emilio Parisotto、Jimmy Lei Ba 和 Ruslan Salakhutdinov。通过 Attention 从字幕生成图像 。2016 年, *ICLR*。²
- [39] Margaret Mitchell、Dylan Baker、Nyalleng Moorosi、Emily Denton、Ben Hutchinson、Alex Hanna、Timnit Gebru 和 Jamie Morgenstern。子集选择中的多样性和包容性指标》，第 117-123 页。Association for Computing Machinery, New York, NY, USA, 2020.³
- [40] 埃利斯 - 蒙克。蒙克肤色量表。
<https://skintone.google>, 2022。^{2,5,6,9}
- [41] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。GLIDE：利用文本引导的扩散模型实现逼真图像生成和编辑。*ICML*, 2022 年。²
- [42] Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-jing Zhu。BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*, 2002.²
- [43] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger、Ilya Sutskever、Jong Wook、Kim Chris、Hallacy Aditya、Ramesh Gabriel、Goh Sandhini、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。从自然语言监督中学习可转移的视觉模型。In *ICML*, 2021.⁶
- [44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 使用 CLIP Latents 的分层文本条件图像生成 *eration.ArXiv*, 2204.06125, 2022。²
- [45] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。零镜头文本到图像生成。In *ICML*, 2021.^{1,2}
- [46] 维克多·雷论种族批判理论：为什么它重要，为什么你应该关心》。兰登书屋出版集团, 2022 年。⁵

- [47] Ali Razavi、Aaron van den Oord 和 Oriol Vinyals。用 VQ-VAE-2 生成多样化高保真图像。 In *NeurIPS*, 2019.⁶
- [48] Scott Reed、Zeynep Akata、Xinchen Yan、Lajanugen Logeswaran、Bernt Schiele 和 Honglak Lee。生成式广告 versarial 文本到图像合成。 In *ICML*, 2016.²
- [49] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。使用潜在扩散模型的高分辨率图像合成。2022 年 6 月, *CVPR*, 第 10684-10695 页。^{1, 2, 6}
- [50] 坎迪斯-罗斯、鲍里斯-卡茨和安德烈-巴尔布。测量基础视觉和语言嵌入中的社会偏见。在 *NAACL*, 2021 年。^{2, 3}
- [51] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S. Sara Mahdavi、Rapha Gontijo Lopes、Tim Salimans、Jonathan Ho、David J Fleet 和 Mohammad Norouzi。具有深度语言理解能力的逼真文本到图像扩散模型- ing.2022 年, *NeurIPS*.²
- [52] Tim Salimans、Ian Goodfellow、Wojciech Zaremba、Vicki Cheung、Alec Radford 和 Xi Chen。用于训练 GANs 的改进技术。 In *NIPS*, 2016.^{1, 2}
- [53] Candice Schumann, Gbolahan O Olanubi, Auriel Wright, Ellis Monk Jr, Courtney Heldreth, and Susanna Ricco. *arXiv preprint arXiv:2305.09073*, 2023.⁶
- [54] Jascha Sohl-Dickstein、Eric A. Weiss、Niru Maheswaranathan 和 Surya Ganguli。使用非平衡热力学的深度无监督学习。 In *ICML*, 2015.²
- [55] Tejas Srinivasan 和 Yonatan Bisk。最糟糕的两个世界：预训练视觉与语言模 els 中的偏差复合体。 *ArXiv*, abs/2104.08666, 2021.³
- [56] Ryan Steed 和 Aylin Caliskan.通过无监督预训练学习到的图像表征包含类似人类的偏差。 In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 701-713, New York, NY, USA, 2021. Association for Computing Machinery.³
- [57] Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jonathon Shlens 和 Zbigniew Wojna。重新思考计算机视觉的感知架构。 *CVPR*, 2016.²
- [58] 唐瑞祥、杜梦楠、李悦宁、刘子睿、邹娜、胡霞。减轻字幕系统中的性别偏见。 In *Proceedings of the Web Conference 2021*, WWW '21, page 633-645, New York, NY, USA, 2021. Association for Computing Machinery.³
- [59] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 神经离散表征学习 ing. In *NIPS*, 2017.⁶
- [60] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。 In *NIPS*, 2017.¹
- [61] Ramakrishna Vedantam、C. Lawrence Zitnick 和 Devi Parikh。 CIDEr：基于共识的图像描述评估。 In *CVPR*, nov 2015.²

- [62] Angelina Wang、Solon Barocas、Kristen Laird 和 Hanna Wallach。衡量图像封顶中的表征伤害。In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 324-335, 2022.³
- [63] Jialu Wang, Yang Liu, and Xin Eric Wang.减轻图像搜索中的性别偏见。*EMNLP*, 2021.^{3, 6}
- [64] Phil Wang.Dalle-pytorch. <https://github.com/lucidrains/DALLE-pytorch>, 2021.^{1, 6}
- [65] Tianlu Wang、Jieyu Zhao、Mark Yatskar、Kai-Wei Chang 和 Vicente Ordonez。平衡数据集还不够：Estimating and mitigating gender bias in deep image representations.In *ICCV*, pages 5309-5318, 2019.³
- [66] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko.分离技能和概念以实现新颖的Visual Question Answering.*CVPR*, 2021.³
- [67] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan.2022 ACM 公平、问责与透明大会, 2022。³
- [68] Robert Wolfe 和 Aylin Caliskan.Markedness in visual semantic ai. 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022.³
- [69] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He.AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks.In *CVPR*, 2018.^{1, 2}
- [70] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yin-fei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu.用于内容丰富的文本到图像生成的比例自回归模型。*Transactions on Machine Learning Research*, 2022.²
- [71] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas.StackGAN：使用堆叠生成对抗网络合成文字到照片的逼真图像。In *ICCV*, 2017.²
- [72] 张彦哲、蒋璐、Greg Turk 和杨迪一。审核文本到图像模型中的性别呈现差异》， 2023。^{3, 5, 6}
- [73] Dora Zhao, Angelina Wang, and Olga Russakovsky.了解和评估图像字幕中的种族偏见。*国际计算机视觉会议 (ICCV)*， 2021 年。³
- [74] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang.男人也喜欢购物：利用语料库级约束减少性别偏见放大。在 *EMNLP* 中，第 2979-2989 页，丹麦哥本哈根，2017 年 9 月。计算语言学协会。³