TIAM - 文本到图像生成中的对齐度评估指标

Paul Grimal Le Borgne Olivier Ferret Julien Tourille

Universite´ Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{paul.grimal, herve.le-borgne, olivier.ferret, julien.tourille}@cea.fr

摘要

合成图像生成技术的进步使得评估图像质量变得至关重要。虽然已经提出了几种方法来评估图像的渲染效果，但对于根据提示生成图像的文本到图像（T2I）模型来说，考虑额外的因素是至关重要的，例如生成的图像在多大程度上与提示的重要内容相匹配。此外，虽然生成的图像通常是随机起点的结果，但一般不考虑随机起点的影响。在本文中，我们提出了一种基于提示模板的新指标，用于研究提示中指定的内容与相应生成图像之间的一致性。它能让我们从 指定对象的类型、数量和颜色方面更好地描述对齐情况。我们对最近的几个 T2I 模型进行了多方面的研究。我们的方法还获得了一个有趣的结果，那就是图像质量会因作为图像种子的噪声而发生巨大变化。我们还量化了提示中概念的数量、顺序及其（颜色）属性的影响。最后，我们的方法让我们能够识别出一些种子，它们能生成比其他种子更好的图像，从而为这一研究不足的课题开辟了新的研究方向。

1. 导言

自第一代 GANs [12,26] 问世以来，利用神经模型生成合成图像的能力取得了长足进步。最近，基于扩散的模型[1, 16, 23, 30, 31]通过逐步去噪生成高质量图像，进一步拓展了图像合成的范围。在取得这些成就的同时，如何评估这些合成图像的质量一直是个棘手的问题，其本身也是一个研究课题。为了解决这个问题 [34]，人们提出了几种度量方法 [13,32,37]，但这些方法存在各种局限性 [2,4]。最新的模型以文本图像描述为条件，允许对图像进行精细的

对输出的控制。不过，这也给评估其输出结果增加了一项挑战，即估计所生成的合成图像在多大程度上符合其所依据的文字描述。

虽然 "文本到图像"（T2I）模型显示出强大的语义和合成能力，但要获得与所需条件一致的视觉上愉悦的图像，往往需要生成多幅图像才能获得合适的图像。一个可靠的生成模型应该能够与提示中指定的条件保持一致，而与起始噪声无关。为了解决和研究结果的可变性，我们引入了一种新的指标来评估生成模型根据提示的成功率，即文本-图像配准指标[1] (TIAM)。初始噪声在我们的指标中起着至关重要的作用，使我们能够研究其影响。我们的研究表明，某些初始噪声配置优于其他配置，这表明可以通过选择这些配置来获得更好的合成图像。

最近的研究[5, 9, 28, 31, 33]表明，文本条件扩散模型在文本提示中表达的预期内容与图像中实际生成的内容之间的一致性方面存在三个主要问题：(i) *灾难性忽略，即*提示中描述的一个或多个元素没有生成或有时混合生成；(ii) *属性绑定，即*属性（*如颜色）被绑定到错误的实体上；(*iii) 属性泄漏，即提示中指定的属性被正确绑定，但场景中的一些其他元素也被错误地绑定了该属性（图）。*(ii)*属性绑定，即属性（如颜色）与错误的实体绑定；(iii)属性*泄漏*，即提示中指定的属性被正确绑定，但场景中的某些其他元素也与该属性错误绑定（图 1）。通过 TIAM，我们建议在灾难性问题和属性绑定问题的范围内分析生成模型的成功率。对于属性绑定问题，我们提出了一种可靠的方法来评估颜色与人类感知的一致性。

迄今为止，对某些词语和属性的影响的研究在很大程度上仍然不足。Tang *等人*[33]通过研究词语对生成结果的影响提出了一些见解。通过使用 TIAM，我们进一步了解了文本条件和生成结果之间的关系。依靠提示模板而不是

[1]源代码： https://github.com/grimalPaul/TIAM

(a) 灾难性视

(b) 属性绑定

(c) 属性忽泄漏

图 1.使用 Sta- ble diffusion v1.4 生成的 "*狮子和熊的照片*"、"*蓝猫和黄车的照片* "以及 "*红色公共汽车在街上行驶的照片*" 。

与自然提示相比，TIAM 可以量化提示与图像在句法方面的一致性，尤其是提示中主要实体位置的重要性。

总之，我们的主要贡献是(i) 一种基于提示模板的新方法，可自动量化 T2I 模型*在提示-图像配准方面*的性能；(ii) 对几种扩散模型*进行了*深入研究，并*量化了*它们与灾难性选取和属性绑定问题有关的行为；(iii) 研究了初始扩散噪声（种子）对所有这些模型的影响。这项研究得出的主要结论是：(a) 大多数 T2I 模型的对齐性能随着提示中指定的对象数量的增加而显著下降；
(b) 在实践中，有一些种子能系统地带来更好的结果；TIAM 使我们能够识别这些种子，而且它们在处理我们的研究领域之外的对象时仍能提供更好的结果；(c) 大多数 T2I 模型都能成功地将颜色归因于一个对象，但对象越多，性能就越低。

2. 相关工作

文本到图像模型 在目前最先进的模型中，扩散模型 [16] 的表现可圈可点。这些模型在图像中引入噪声，并学习如何去噪。在推理过程中，模型会从高斯分布中迭代去噪采样，从而得到重建图像。利用自由分类器引导[17]，用户可以通过使用自然语言编写提示来表达自己的想法。然后，利用 CLIP [25] 或 T5 [27]等基础模型，通过编码文本表示来引导扩散过程。值得注意的是，其结果具有高度随机性，通常需要用不同的高斯噪声输入生成多个图像，以符合用户所需的文本和偏好。在这一领域的知名模型中，值得一提的包括 Imagen [31]、Dall-E 2 [28]、Latent Diffusion Model（潜在扩散模型
(LDM) [30] 和 Ediff-I [1]。

T2I 评估 对生成图像的评估主要依赖于入射角分（IS）[32] 和弗雷尔切入射角距离（FID）[13] 等指标，这些指标通常用于评估图像的质量和保真度。然而，这些指标并不能反映所提供的条件与生成图像之间的一致性。为了解决这个问题，人们提出了几种方法来衡量生成图像的内容是否反映了给定的条件。其中一种方法是利用分类模型[29]和对象检测模型（Semantic Object Accuracy，缩写为 SOA）[15]来确定生成的图像是否包含指定对象或符合特定标准。SOA 使用 COCO 中的实际提示，而我们的方法则依赖于提示模板，可以对提示中每个元素的影响进行更精细的分析。另一种特定于文本-图像配准的方法是采用视觉分数相似度测量。它利用了对比模型（如 CLIP），该模型计算图像和文本表示之间的相似度得分。虽然 CLIP 的平均语义表示能力很强，但它对构成的理解能力较差[36]，从而限制了对文本-图像配准的精细评估。我们的方法利用高性能检测器，以可解释的方式控制所要求的属性，并与人类感知保持一致，从而解决了这一问题。

最近的努力引入了创新方法来评估技能和测量 T2I 模型的偏差。值得注意的是，Drawbench[31] 提出了一套有限的文本提示，涵盖 11 种技能，用于评估模型。同样，DALL- EVAL [6] 建议评估三种视觉推理技能：生成一个物体的能力（物体识别）、生成被问物体的精确数量的能力（计数）和放置两个物体的能力（空间推理）。此外，他们还探究了模型表征中的性别和皮肤偏差。Zhang 等人[38]的另一项研究深入探讨了性别刻画的差异，从而对潜在的刻板印象进行了研究。我们的方法侧重于灾难性忽略和属性绑定。我们可以把灾难性忽视方法看作是对象识别方法，只不过是在条件提示中测试一个或多个不同的对象。此外，为了克服对提示中询问对象非常敏感的结果，并准确系统地评估模型的性能，我们探索了所有不同对象和属性的组合，并在每个提示中生成多张图像，以提高结果的随机性。考虑到初始噪声的影响，我们建议进行多种子测试。这种方法提供了一个全面的评估，摆脱了特定标签的限制，从而使评估更精确、偏差更小。与我们同时进行的一项密切相关的工作[11]也得出了一些相同的结论，不过它并没有深入研究起始噪声方面的问题。

3. 方法

我们提出了一种新方法来衡量生成模型在灾难性忽略和/或属性绑定方面的成功率。首先，我们根据一组 单词标签和可能的属性生成多个提示。然后，我们为每个提示生成多个图像，并检测图像上是否存在预期元素，从而得出最终分数。

3.1. TIAM 文本图像对齐度量标准

我们的方法基于用于生成提示集的模板。我们采用了一种受分解表示理论[14, 35]启发的形式主义。提示包含 N 个对象，每个对象都可以用一个属性来限定。因此，提示中 i 位置的对象是属于 A_i 集合的标记，并由 A_i 集合中的一个属性限定。例如，让我们考虑模板 "a *photo of* $\det(o_1, a_1)$ o_1 *and* $\det(o_2, a_2)$ o_2 "，其中 $o_i \in O$ 和 $a_i \in A_i$，$\det(o_i, a_i)$ 是一个行列式，取决于对象或属性（如果存在）。如果 $A_1 = A_2$ = '汽车'、'自行车'、'卡车'，而 $A_1 \neq A_2$ = '蓝色'、'绿色'、'红色'，那么它就能生成 "*一张蓝色汽车和一辆红色卡车的照片* "或 "*一张绿色自行车和一辆蓝色汽车的照片* "这样的提示。

对于这样的模板 t，文本图像对齐度量（TIAM）的通用表达式定义为

$$\mathop{\mathbb{E}}_{\substack{\chi \sim \mathcal{N}(0,I) \\ z \in Z}} [f(G(\chi, t(z)), y(z)) \text{ 与 } Z = \prod_{i=1}^{N} (A_i \times O)_i \tag{1}$$

其中，提示是从模板 t 及其 "潜在概念 "z 中实例化出来的，$y(z)$ 是与模型 $G()$（以提示为条件）从*种子*（用于生成初始噪声）χ *生成的*合成图像预期内容相关的标签。f 是一个评分函数，用于比较地面实况 $y(z)$ 和模型输出（该模型用于检测合成图像中的物体或生成分割图），得出的分数为 0、1 或介于两者之间的连续值，取决于内容是否与地面实况相匹配（见第 3.3 节）。

根据公式 1 的定义，模板可以生成 $\prod_{i=1}^{N} |A_i|$ 的不同提示，其中 N 是心率。的结果。然而，在实际操作中，我们限制推理

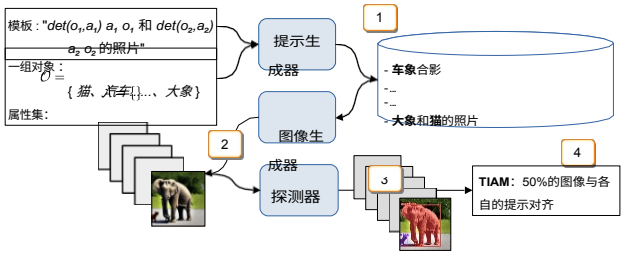


图 2.评估流程概览(1) 生成提示语数据集。(2) 为每个提示生成 $n \geq 16$ 幅图像。
(3) 检测图像中是否存在所要求的标签。(4)
计算 TIAM。在本例中，我们不定义属性。

证明见补充材料。我们还推导出了一种更普遍的情况，即每个位置的属性和对象都可能不同。图 2 概括了全局评估过程。

3.2. 属性定义

本文研究的是颜色属性，但我们的方法也适用于其他类型的属性（*如*尺寸或纹理）。选择符合人类感知的颜色并非易事，因为有无数种可能。我们的选择基于 Berlin 和 Kay 的研究成果[3]。他们定义了 11 种通用的*基本颜色* C：*白色*、*黑色*、*红色*、*绿色*、*蓝色*、*棕色*、*紫色和粉色*、*橙色*、*黄色*和*灰色*。他们要求个人从 329 种颜色（由 Munsell 颜色公司提供）中选择与每种*基本颜色*相对应的芯片，并

的模板，这样一个对象或属性就不会在提示中出现两次。一个简单的例子是，在每个位置上，at- tribute 和对象集都是相同的，那么唯一提示符的数量由以下公式给出：

命题 1： 如果 N，和 i $[1, N]$，$_i =$ ，$_i =$ ，和 (i, j) $[1, N]^2$ s.t $i < j$，我们强制 $a_i = a_j$ 和 $o_i = o_j$ ，因此在等式 1 的上下文中生成的唯一提示数为

来选择最典型的例子。我们使用最典型的美式英语范例的结果，并将颜色从 Munsell 系统转换到 CIELAB 空间。这样做的好处是颜色分布更符合人类的感知，也是比较颜色差异的合适空间。我们注意到*棕色*和*橙色*分别过于接近*黑色*和*白色*，因此将其删除（更多信息见补充资 料）。我们设定
A = 然而，为了确定 CIELAB 空间中像素的实际颜色，我们还要考虑*白色*和*黑色*。

为了确定 $G()$ 是否正确分配了 $_i$ 和 o_i，我们使用了一个分割模型，如果图像上存在 o_i，该模型会对其进行分割。我们将分割图中的所有像素与 CIELAB 空间中的参考颜色进行比较。如果我们检测到至少 40% 的 $_i$，我们就认为绑定成功。

3.3. 主要实施细节

事实上，在我们的案例中，检测器是在真实图像上进行训练的，因此我们认为它在处理草图或 3D 渲染图等方面会有更大的困难。我们的

通过改变句法上下文，模板的一般形式就是上文所举的例子。

在检测和分割任务中，我们使用了 YOLOv8 [18]，这是一个在 80 个 COCO 标签[22]上预先训练过的先进模型。我们将置信度阈值设为 0.25，但进行的一项研究表明，阈值越高（0.4 至 0.8），模型的相对顺序性能越好。研究结果见附录 Mat.，其中详细描述了每次实验中使用的标签及其在生成的提示中的组合。评分函数 f 主要关注的是真阳性，如果我们在提示中至少找到了一个被命名的对象，则认为配准成功。如果对象有属性特征，则每个对象必须在图像中至少出现一次，且属性正确，才算成功。因此，该标准在对齐方面非常严格，要求出现*所有*已命名的对象和属性。这种方法不同于 CLIP 分数，后者提供的是总体趋势，而不是对结果的可理解解释。不过，这种方法不会对其他对象的存在进行惩罚（假阳性），因为在生成的情况下，这并不具有禁止性。实际上，如果两个不同对象的分割掩码重叠，且 IoU 值达到或超过 0.95，我们就会先移除这两个对象，然后再用剩下的对象计算分数。检测到的物体。

4. 研究

我们进行了一项深入研究，以描述和量化几种扩散模型的极限，这些模型因文本调节方法、架构或推理过程的不同而各异。我们考虑了基于 LDM 模型 [30] 的 Stable Diffusion v1-4（SD 1.4）和 Stable Diffusion v2（SD 2），它们使用固定的预训练文本编码器 CLIP [25]，而 CLIP 是基于自动回归架构的。两个版本都使用相同的变异自动编码器 (VAE) 进行训练，但使用了两个不同的 U 网（学习图像去噪的部分）。此外，我们还使用 " 注意并激发"（A&E）[5] 对两个 SD 模型进行了评估，这是一种优化的推理过程，它使用交叉注意图来注意提示中的主题标记。我们沿用了作者的实现方法，只关注对象 o 的标记$_i$，即使该对象有属性特征。我们还考虑了以 CLIP 图像先验和 CLIP 文本先验为条件的 unCLIP 模型 [20, 28]。最后，我们研究了 DeepFloyd IF (IF)，它是受 T5 XXL [27] 文本编码器启发的级联扩散模型[31]。其他细节见附录。

4.1. 初步结果：使用 2 个对象时性能下降

我们报告了第一个实验，以说明我们研究的总体框架。我们考虑了来自

模型	1 个物体	2 个对象
SD 1.4	0.98	0.41
SD 1.4 A&E	0.96	0.64
SD 2	0.99	0.61
SD 2 A&E	0.98	0.65
unCLIP	0.95	0.50
IF	0.99	0.62

表 1.根据提示中对象数量计算的 TIAM，适用于本研究考虑的所有 6 个生成模型。

COCO（详尽清单见附录 Mat.），从而得到一组 = 24 个对象标签。对于 $o_i = o_j$ 的所有可能的标签对（o_i，o_j），我们都会提示 "一张 o_i 和 o_j 的照片"（按预期管理行列式），并通过改变随机种子 χ 生成 64 张合成图片，然后使用 TIAM 估算提示与每张图片的对齐情况。作为参考，我们还对由更简单的提示语 "o_i 的照片 "生成的 64 幅图像进行了同样的试验。

我们在表 1 中报告了所有模型的结果。1 中报告了所有模型的结果。这些模型在生成单个物体的图像方面一直很成功，得分都在 0.95 以上。但是，它们在同时生成两个物体的图像方面却很吃力，最多只有 66% 的图像能正确生成，而提示却非常简单。在最近的文献中，T2I 模型通常表现出很好的定量渲染得分，*例如*在萌芽得分或 FID 方面，我们的经验并没有对此提出质疑。与 Hinz *等人的研究*[15]一样，本实验只表明生成图像的内容与提示之间存在对齐问题。与[15]相比，我们的实验更精确地量化了对齐性能的下降在多大程度上是由于提示中存在多个概念造成的。对于有两个对象的提示，我们在表 2 中报告了对象在提示中出现的次数。2 中报告了根据两个对象在提示中的位置而准确生成它们的次数（表 1 中的分数要求两个对象都存在才有效）。我们发现，在所有模型中，第一个 o_1 比第二个 o_2 更为普遍，第 4.3 节将进一步研究这一现象。

我们注意到，对于 SD v1.4，A&E 推论在总体上大大提高了得分。从表 2 中可以看出，A&E 的超常表现似乎是由于它能够提高数据的出现率。从表 2 中可以看出，A&E 的超常表现是由于它能够通过强制交叉注意图的最小激励来提高 o_2 的出现率。

4.2. 随机种子的重要性

我们考虑了与第 4.1 节相同的 24 个标签和提示 "o_i 和 o_j 的照片"。我们还通过改变随机种子 χ 生成 64 幅同步图像，并计算每条提示的 $f\,(G(\chi,\, t(z)),\, y(z))$ 和对应的

$|0|$ ，

模型	o1	o2
SD 1.4	0.80	0.60
SD 1.4 A&E	0.85	0.75
SD 2	0.83	0.78
SD 2 A&E	0.84	0.80
unCLIP	0.77	0.71
IF	0.86	0.76

表 2.提示语中每个命令的出现比例。
o2 指模板中的位置。

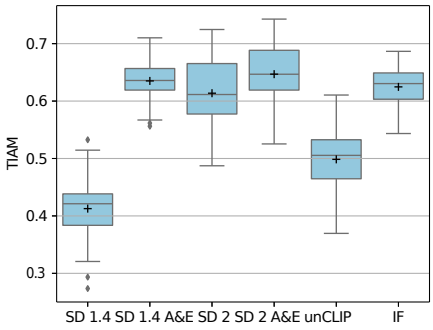


图 3.64 粒种子的每粒种子 TIAM 总计。我们可以看到，一些起始噪声往往无法转换为包含两个实体的图像，而与实体无关。"+"表示平均值。

图像。不过，这次所有 $(o_i, o)_j$[2] 的 64 个种子都是一样的，我们汇总了所有可能的提示和图像的*每个种子的*性能，从而得出每个种子的文本-图像对齐估计值为 24 23 = 552。

图 3 中的结果显示，对于所有 6 个模型来说，其性能随随机种子的不同而变化很大。据我们所知，这些结果在 T2I 生成模型的文献中从未出现过。对于这类模型，人们通常期望所有种子都有相同的机会生成指定的元素。

的问题。实际上，最近的趋势是通过精细的提示[24]来优化 T2I 模型的输出。与这一趋势相反，我们的方法开启了

在选择 "性能种子 "的基础上进行补充优化（见第 4.6 节）。总体而言，最佳种子和最差种子之间存在显著差异。如果通过图 3 的 方框图来考虑四分位数之间的范围和最小-最大范围，就会发现模型之间的平均性能差异远没有它们自身的平均性能差异显著。由于种子的不同，差异较小。

Obviously, enhancing a dependence “to the seed” is convenient from a practical point of view but wrong in all strictness since the initial noises are drawn from a Gaussian at inference. Being more rigorous requires reminding the diffusion models training process [16]. It consists of learning the reverse process of a fixed Markov chain of length T with models that can be interpreted as an equally weighted sequence of denoising autoencoders. The latters are trained to denoise their input, considered as a noisy version of an input training image I. In the case of latent diffusion models [30], the process is embedded into the latent space of a VAE (encoder E + decoder D), such that the autoencoders are U-net networks $\epsilon_\vartheta(x_t, t)$, $t = 1 \dots T$ that denoise the latent code x_t by minimizing the loss $\mathbb{E}_{E(I),\epsilon\sim\mathcal{N}(0,1)} \|\epsilon - \epsilon_\vartheta(x_t, t)\|_2^2$. During inference, the process starts from a latent code $\chi_T \sim \mathcal{N}(0, 1)$, denoises it with the U-net ϵ_ϑ to get the final latent code χ_0, then obtains the synthetic image with VAE decoder as $D(\chi_0)$. This

因此，最后不仅取决于随机种子 χ_T，还取决于 U 网参数 ϑ。*通过*学习这些参数，可以在任何扩散步骤中对潜码进行去噪，因为重参数化技巧[16]可以直接将最终代码表示为：

$$x_{\epsilon_0} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \tag{2}$$

[2]实际上，为了便于比较和连贯，第 4.1 节中的情况也是如此。

其中 ϵ (0，1），而 α_t 取决于扩散过程噪声（详见 [16] 和 [30] 附录 B）。理想情况下，$\bar{\alpha}_T$ 接近于 1，这样同步图像应该来自几乎完美的高斯噪声。然而，在实践中，这一过程并不完美；因此，在推理时，对于给定的 χ_T (0, 1) 所得到的终 χ_0 只是 "理想 χ_0 " 的近似值，而 "理想 χ " 是在完美优化的情况下可以预期的 [7]。更重要的是，由于每个模型都是独立训练的，其潜在空间的结构是相同的，χ_t 在潜在空间中的 "路径 "可能因模型而异。因此，从相同的 χ_T (0, 1) 出发，它们在试图趋向于一个理想的 χ_0 时，会产生截然不同的结果。因此，"好 "种子和 "坏 "种子是每个模型所特有的。

结果取决于种子，表明根据提示的规格，似乎有可能识别出更有可能表现出多个对象的初始噪声。这凸显了在减少对潜在变量的依赖方面不断取得进步的必要性。为了获得稳健的评分并避免种子成功率的高变异性可能造成的影响，我们规定了每个提示生成图像的最低数量。我们确定每个提示生成 32 幅图像即可获得稳健评分。相应的实验报告见附录。

4.3. 灾难性忽视

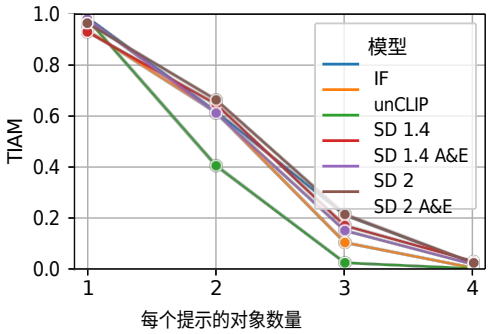
我们将研究这些模型在对象集数量增加时的行为，以及对象顺序在这些模型中的作用。

提示。我们设置 = {*汽车*、*冰箱*、*长颈鹿*、*大象*、*斑马*}，并为包含 1 到 4 个对象的提示计算 TIAM。我们设计了 4 个模板，提示中的物体数量各占一个，并为每个提示生成 32 幅图像，在提示中有 1、2、3 和 4 个物体时（每个模型）分别产生 160、640、1,920 和 3,820 个对齐值。如图 4 所示，当提示对象超过两个时，模型无法持续生成输出。即使使用了 A&E 机制，生成四个对象仍然几乎是不可能的。SD 2 和 IF 显示出相对较好的
但改进幅度不大。

在研究不同物体的出现情况时，可以观察到与初步实验类似的趋势（表 2）。具体来说，模板中初始对象的出现频率往往高于随后插入的对象。图 5 显示了包含四个对象的模板的结果（两个和三个对象的结果见附 图）。这证实了一个观点，即在提示中较早表达的概念在最终图像中出现的机会较多。

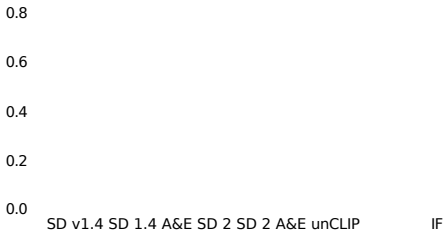
就 SD 而言，以 CLIP 文本编码器为条件，随着对象在提示中的位置越来越远，其出现率呈下降趋势，部分原因可能是编码器的自动回归特性。在自我注意过程中，标记只接收来自其左侧元素（提示语开头）的上下文。因此，标记词没有后续对象的上下文，而最后这些对象则带有前面词语的上下文。因此，U-网络模型会从较远的标记词中接收到更模糊的信号。在 T5 编码器中，所有单词在自我注意过程中都可以互相访问，这一点的解释就不那么清楚了。在 IF 中，我们可以看到第三个和第四个对象出现的几率是一样的，但明显小于第二个位置上的对象，其本身低于提示中的第一个对象。我们假设，在交叉注意过程中，由于训练数据通常会把与图像相关的重要元素放在标题的开头，因此模型学会了更重视位置较早的标记。然而，要证明这一解释，需要对用于预训练六个生成模型的原始训练数据集进行大量的分析工作，而这不在本文的研究范围之内。

最后，我们研究了提示对象之间的语义关系对模型



生成这两个对象的能力的影响。考虑到模板中有两个物体，我们假设 T2I 模型在同一图像中表现两个有语义联系的物体时，失败的次数会更多。我们考虑了来自三个大类（*车辆*、*动物*和 *食物*）的 28 个 COCO 图像，并使用包含两个物体的模板生成图像。根据得到的 TIAM 分数，我们得出了所有对象之间的不相似度量（见附注），并将所有对象的不相似度量投影到图像中。

图 4.每个提示 1 至 4 个对象的 TIAM。



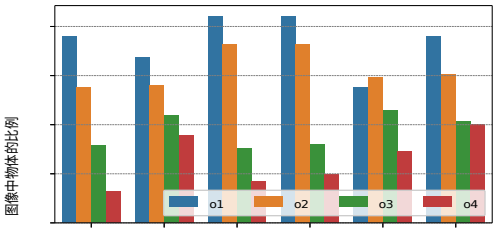
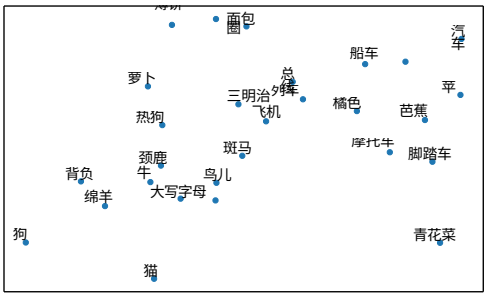


图 5.根据每个对象在提示中的位置计算的出现比例（模板中有 4 个对象）。

图 6.关于 SD 1.4 的对象得分差异的 MDS。



通过多维尺度（MDS）对标签进行分析。由此产生的投影（图 6，SD 1.4）可以解释为，两个标签越接近，模型将它们放在一起的难度就越大。在所有模型中，我们观察到同一类别的标签聚集在一起，尤其是*动物*和*车辆*。这表明存在语义邻近性。在 SD 1.4 和 SD 2 中，我们使用多种方法（如 Wu-Palmer、CLIP 文本嵌入的余弦相似度，甚至是 attention 的键值表示法）测量了两个命名对象之间的语义距离，但与 TIAM 分数的相关性仅为轻微负值。这表明，语义链接可能与配准性能有微小或间接的联系，但无论如何，还需要进一步的研究来澄清这一点。

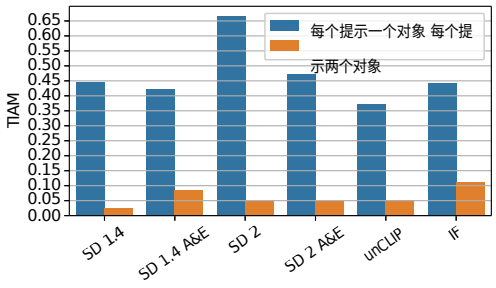


图 7.使用对象和颜色基本事实计算的 TIAM，每个提示一个和两个彩色对象。

	1 个对象		2 个对象	
属性	✗	✓	✗	✓
SD 1.4	0.97	0.95	0.40	0.18
SD 1.4 A&E	0.93	0.94	0.64	0.42
SD 2	0.97	0.97	0.61	0.26
SD 2 A&E	0.96	0.96	0.66	0.32
unCLIP	0.93	0.89	0.61	0.36
IF	0.98	0.90	0.61	0.49

表 3.仅使用对象基本事实计算的 TIAM，提示是否包含属性。

4.4. 属性绑定

在本节中，我们将描述模型将属性应用于对象的能力。让我们考虑两个提示，分别有一个和两个对象，使用相同的集合 $O = \{$*汽车*、*冰箱*、*长颈鹿*、*大象*、*斑马*和第 3.2 节中定义的颜色集合。

我们首先计算 TIAM，而不管属性的正确性如何，并在表 3 中报告结果。3.如果不考虑属性 (✗)，得分与表 1 和表 3 中的得分接近。1 和
添加属性 (✓)对一个对象的影响有限
只要求提供。然而，这对每
当提示符中出现两个对象时的形式。

如果同时考虑物体和颜色的基本事实来估算 TIAM，我们会发现，即使物体出现在图像中，模型也无法为其分配适当的颜色（图 7）。例如，在 SD 1.4 的情况下，95% 的单物体提示都能正确检测到物体，但其中只有约 45% 的物体具有所需的颜色。

在图 8 中，我们报告了每种颜色和对象的归因得分，以深入分析模型的归因能力。我们发现，当观察到物体的特定颜色并不常见时，模型就无法将绑定泛化到物体上。不出所料，与动物相比，为汽车和冰箱指定颜色确实更容易。在训练过程中，模型很可能是根据各种颜色的汽车和冰箱照片进行训练的，因为汽车和冰箱的颜色比较常见。

绿色长颈鹿。

按照第 4.3 节的方法，我们还分析了在涉及 2 个对象的提示中，属性 a_i 和对象 o_i 的位置 i *的*结果。在这种情况下，模型在生成和绑定第一个对象方面取得了很大的成功。然而，我们知道第一个对象更常被生成，因此我们计算了*绑定成功率*，即在正确生成的对象中正确归属对象的得分（图 9）。但$_2$ 对象*的*归属率仍然较低。这进一步证明了提示中的第一个对象对最终生成的影响更大。在补充数据中，我们报告了按颜色区分第一位置属性和第二位置属性的*绑定成功率*结果。我们观察到，当两个对象同时存在时，模型在分配绿色和蓝色时面临更大的困难（与单标签情况平行）。值得注意的是，IF 的表现优于其他模型。

4.5. 与人类、CLIP、BLIP 比较

我们随机选择了 32 对提示-图片，并请 57 位人类对内容配准进行评估。根据弗莱斯卡帕（Fleiss'kappa）[10]，他们的一致可靠性为 0.73，可被视为 "实质性"[19]。一半的提示有两个对象，另外 16 个提示有一个彩色对象。人类与 TIAM 的皮尔逊相关性为 0.82 ($p < 10^{-8}$)。我们将 TIAM 与另外两种基于 CLIP [25] 和 BLIP [21] 的自动方法进行了比较，它们与人类的相关性分别为 0.47 ($p < 10^{-2}$) 和 0.67 ($p < 10^{-4}$)。如附注所述，对于有两个物体和一个彩色物体的图像，TIAM 与人类的配准效果更好。

4.6. 走向噪音开采?

为了突出噪音性能的重要性，我们根据种子的 TIAM 分数来选择种子。在图 10 中，我们展示了使用我们的方法找到的最差和最佳种子的定性结果，以及我们的研究中没有考虑的对象和属性的提示结果。在使用 SD 2 和 IF 时，"好 "种子生成的图像中都能更好地代表两个物体，这表明我们的方法有能力找到能够概括用于确定最佳种子的物体领域之外的物体的种子。我们使用提示中的颜色对最差种子和最佳种子进行了相同的种子选择过程（图 11）。使用 TIAM 确定的最佳种子所生成的图像再次更好地反映了提示。使用 "坏 "种子时，SD 2 会出现属性泄漏或绑定（"蓝月亮 "是黄色的，而 "红狮子 "是蓝色的）。这强调了对提示开始时的噪音的依赖性，而噪音仍然是影响性能的一个重要因素。

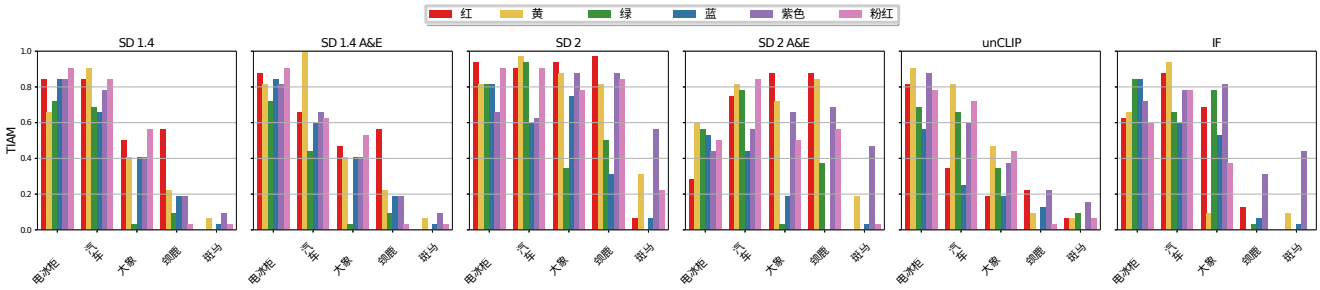


图 8.每种颜色和对象的 TIAM

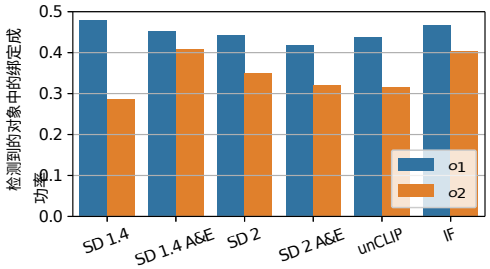


图 9.与检测到的目标相关的颜色归因成功率。

图 10.用相同的提示和 "坏"（左）与 "好"（右）随机种子生成图像：SD 2 "老虎和冲浪板的素描，4K，8K，吉卜力"，种子 23，11：IF "太空老虎和火箭的照片"，种子 41，9。

图 11.用相同的提示和 "坏"（左）与 "好"（右）随机种子生成图像：SD 2 "一幅红狮子和蓝月亮的图画，波普艺术，4K，高度精细"，种子 27，17。*中频* "紫色青蛙和彩虹钢琴的照片"，种子 19，24。

5. 结论、局限和展望

我们提出了一种新指标，用于自动量化 T2I 模型在提示-图像对齐方面的性能。与之前的研究不同，它基于提示模板，可以对提示语法进行更精细的分析。因此，我们发现大多数 T2I 模型的对齐性能会随着提示中指定的对象数量的增加而显著下降，而这种影响对于颜色归属更为关键。扩展

如果 TIAM 包含更多的对象和属性，就需要生成指数数量的提示，因此在实践中会变得非常麻烦。不过，我们在《补充数学》第 12 节中表明，TIAM 可以通过一组 300 条提示可靠地推断出来。我们的指标还允许我们研究输入种子对推理的影响。我们发现，有些种子能系统地生成比其他种子更好的输出图像，而且它还能泛化到用于确定种子的集合之外的对象。这为今后挖掘此类 "好种子 "的研究提供了可能，类似于对文本模型的一些研究[8]，作为提示工程的补充活动，以优化 T2I 模型的输出。为了全面分析 T2I 性能，TIAM 应与反映提示图像对齐以外其他方面的其他指标相结合，如 [6, 13, 31, 32, 38]。

致谢 根据 GENCI 分配的 2022- AD011014009 号拨款，本研究获准使用 IDRIS 的高性能计算资源。在法兰西岛大区议会的资助下，本出版物使用了 FactoryIA 超级计算机。

≈

参考资料

[1] Yogesh Balaji, Seungjun Nah, *Xun* Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu.1, 2

[2] Shane Barratt 和 Rishi Sharma.关于起始得分的说明。In *Proc. ICML 2018 Workshop on Theoretical Foun- dations and Applications of Deep Generative Models*, 2018.1

[3] 布伦特-柏林和保罗-凯*基本色彩术语：Their Universality and Evolution*.加州大学出版社，洛杉矶，1969年。3

[4] Ali Borji.gan评估措施的利弊。*计算机视觉与图像理解*》，179:41-65，2019.1

[5] 希拉-切弗、尤瓦尔-阿拉卢夫、雅伊尔-温克、利奥尔-沃尔夫和丹尼尔-科恩-奥尔。注意与兴奋：基于注意力的文本到图像扩散模型语义引导。*ACM Trans.Graph.*, 42(4), jul 2023.1, 4

[6] Jaemin Cho、Abhay Zala 和 Mohit Bansal。Dall-eval：探索文本到图像生成转换器的推理能力和社会偏见。*ArXiv 2202.04053*，2022。2, 8

[7] Thomas G. Dietterich. 机器学习中的集合方法。In *Multiple Classifier Systems*, pages 1-15, Berlin, Heidelberg, 2000.Springer Berlin Heidelberg.5

[8] Jesse Dodge、Gabriel Ilharco、Roy Schwartz、Ali Farhadi、Hannaneh Hajishirzi 和 Noah Smith。微调预训练语言模型：权重初始化、数据顺序和早期停止。*ArXiv 预印本 arXiv:2002.06305*, 2020.8

[9] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang W a n g . 用于合成文本到图像的无训练结构化扩散引导。*第十一届学习表征国际会议*，2023年。1

[10] J.L. Fleiss.在众多评分者之间测量名义量表的一致性。*Psychological Bulletin*, 76(5):378--382, 1971. 7

[11] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang.文本到图像生成中*的*空间关系基准。2

[12] 伊恩-古德费洛、让-普热-阿巴迪、迈赫迪-米尔扎、徐兵、戴维-沃德-法利、谢尔吉尔-奥扎尔、亚伦-库维尔、尤斯华-本吉奥。生成对抗网*神经信息处理系统进展*》，第 27 卷，2014 年。1

[13] 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。*神经信息处理系统进展*》，第 6629-6640 页，2017 年 12 月。1, 2, 8

[14] 伊琳娜-希金斯、戴维-阿莫斯、戴维-普福、塞巴斯蒂安-拉卡涅雷、洛伊克-马特、达尼洛-雷森德和亚历山大

Lerchner.Towards a definition of disentangled representations. *ArXiv 1812.02230*, 2018.3

[15] Tobias Hinz、Stefan Heinrich 和 Stefan Wermter。生成式文本到图像合成的语义对象准确性。*模式分析与机器智能》（IEEE Transactions on Pattern Analysis and Machine Intelligence）*，44（3）：1552-1565，2022 年 3 月。2, 4

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel.去噪差异融合概率模型。*神经信息处理系统进展》*，第 33 卷，第 6840-6851 页。
库兰联合公司，2020 年。1, 2, 5

[17] Jonathan Ho 和 Tim Salimans.*ArXiv 2207.12598*, 2022.2

[18] Glenn Jocher、Ayush Chaurasia 和 Jing Qiu。2023 年，通过超声心动图测量约洛。4

[19] J.R. Landis 和 G. G. Koch.分类数据的观察者一致性测量。*生物统计学》*，33：159-174，1977 年。7

[20] 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, 2022.4

[21] 李俊楠、李东旭、熊才明和 Steven Hoi。Blip：为统一视觉语言理解和生成进行语言图像预训练的引导。*ICML*, 2022.7

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dolla´r, and C. Lawrence Zitnick.微软 coco：上下文中的通用对象。载于 David Fleet、Tomas Pajdla、Bernt Schiele 和 Tinne Tuyte- laars 编辑的《*计算机视觉 - ECCV 2014*》，第 740-755 页，Cham，2014 年。施普林格国际出版公司。4

[23] Alexander Quinn 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光感图像的生成和编辑。*第 39 届机器学习国际会议论文集》*，*《机器学习研究论文集》第* 162 卷，第 16784-16804 页。PMLR，2022 年 7 月 17-23 日。1

[24] Jonas Oppenlaender.用于文本到图像生成的提示修饰符分类学》，*arXiv 2204.13988*, 2023.5

[25] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、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。*第 38 届国际机器学习大会论文集》*，*《机器学习研究论文集》第* 139 卷，第 8748-8763 页。PMLR，2021 年 7 月 18-24 日。2, 4, 7

[26] Alec Radford, Luke Metz, and Soumith Chintala.使用深度卷积生成对抗网络的非超视化表征学习。2016年*国际表征学习大会。*1

[27] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。用统一的文本到文本转换器探索迁移学习的极限》，*ArXiv 1910.10683*, 2020.2, 4

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen.使用剪辑潜变量的分层文本条件图像生成。*ARXiv 2204.06125*, 2022.1, 2, 4

[29] Suman Ravuri 和 Oriol Vinyals.条件生成模型的分类准确率得分。*神经信息处理系统进展》*，2019 年。2

[30] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bjo¨rn Ommer。利用潜在扩散模型合成高分辨率图像。*IEEE/CVF 计算机视觉与模式识别会议（CVPR）论文集》*，第 10684-10695 页，2022 年 6 月。1, 2,
4, 5

[31] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Raphael Gontijo-Lopes、Burcu Karagol Ayan、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。见 Alice H. Oh、Alekh Agar- wal、Danielle Belgrave 和 Kyunghyun Cho 编著的《*神经信息处理系统新进展》（Ad- vances in Neural Information Processing Systems）*，2022 年。1, 2, 4, 8

[32] Tim Salimans、Ian Goodfellow、Wojciech Zaremba、Vicki Cheung、Alec Radford、Xi Chen 和 Xi Chen。改进的甘斯训练技术见 D. Lee、M. Sugiyama、U. Luxburg、I. Guyon 和 R. Garnett 编辑的《*神经信息处理系统进展*》第 29 卷。Curran As- sociates, Inc., 2016.1, 2, 8

[33] Raphael Tang、Linqing Liu、Akshat Pandey、Zhiying Jiang、Gefei Yang、Karun Kumar、Pontus Stenetorp、Jimmy Lin 和 Ferhan Ture。什么是 DAAM：使用交叉注意力解读稳定的不同融合。In *Proceedings of the 61st An- nual Meeting of the Association for Computational Linguis- tics (Volume 1: Long Papers)*, pages 5644-5659, Toronto, Canada, July 2023.计算语言学协会。1

[34] L.Theis, A. van den Oord, and M. Bethge.关于生成模型评估的说明。*国际学习表征会议*，2016 年 4 月。1

[35] Matthew Trager、Pramuditha Perera、Luca Zancato、Alessan- dro Achille、Parminder Bhatia 和 Stefano Soatto。意义的线性空间：视觉语言模型中的合成结构。In *ICCV 2023*, 2023.3

[36] Mert Yuksekgonul、Federico Bianchi、Pratyusha Kalluri

、Dan Jurafsky 和 James Zou。视觉语言模型何时、为何表现得像词袋？*第十一届学习表征国际会议*，2023年。2

[37] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang.深度特征作为感知度量的不合理效果。In *IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, 2018.1

[38] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang.审核文本到图像模型中的性别呈现差异》，*ArXiv 2302.03675*, 2023.2, 8