

DALLE-2 看到了双重:单词到概念映射的缺陷 文本到图像模型

Royi Rassin*¹ 巴, Shauli Ravfogel*^{1,2} Yoav Goldberg^{1,2}
伊兰大学²艾伦人工智能研究所{rassinroyi,绍利·拉夫福格尔, yoav.goldberg}@gmail.com

抽象的

我们研究 DALLE-2 将提示中的符号 (单词)映射到其引用 (生成图像中的实体或实体的属性)的方式。我们表明,与人类处理语言的方式形成鲜明对比,DALLE-2 不遵循每个单词在解释中具有单一角色的约束,有时会出于不同目的重复使用相同的符号。

我们收集了一组反映这一现象的刺激:我们证明 DALLE-2 同时描述了名词的两种含义和多种含义;并且给定的单词可以修改图像中两个不同实体的属性,或者可以被描述为一个对象并修改另一个对象的属性,从而在实体之间创建属性的语义泄漏。总而言之,我们的研究强调了 DALLE-2 和人类语言处理之间的差异,并为未来研究文本到图像模型的归纳偏差开辟了道路。

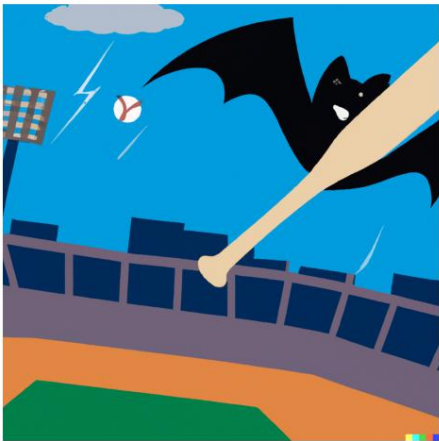


图 1:在提示蝙蝠飞过棒球场的情况下,“bat”一词被实现为两个实体。

一根会飞的木棍或一只会飞的动物,但绝不能同时出现。一旦蝙蝠这个词出现了

在解释中用来表示一个物体 (例如木棍)时,它不能在同一解释中重新用于表示另一个物体 (动物)。同样,在鱼和金锭中,“金”一词用作“锭”的修饰语。

¹ 一旦使用,就

不能重复使用它来修改相同解释中的另一个对象,也不能用作独立对象。²一些语言学家将此属性称为资源敏感性 (Salvucci et al., 2009)。

我们展示的证据表明 形成鲜明对比

¹我们使用“修饰语”一词来指代任何表示影响输出中描述另一个单词的方式。

²请注意,在某些情况下,一个单词可以用来修饰多个对象,例如,a gold Fish and ingot 可以解释为金鱼和金锭。这些情况以非常具体的句法配置体现,并且在语言学中有详细记录。事实上,对此类案例的语言分析要么基于单词的“重复”等机制,要么暗示该单词出现两次并在生成之前被删除的句子的更深层次版本。虽然这些语言理论背后的现实无法得到证实,但呼吁使用“重复”或“删除”(在实现之前)等机制来解释这些现象,凸显了“每个符号单一使用”原则的自然性。

*同等贡献。

对人类来说 文本到图像模型DALLE-2

(Ramesh 等人, 2022)不尊重这一限制。事实上,我们表明DALLE-2展示了

下列行为不符合

单角色原则:

- 单词或短语被解释为两个不同 (同时不兼容)的实体
因此同一场景中被实现为多个对象。例如,

蝙蝠飞过体育场的提示是
映射到显示棒球棒的图像
与飞行的哺乳动物一起 (图1)。

- 一个词被解释为两个词的修饰语
不同的实体,³ 因此是真实的
化为多个对象的属性
场景。例如,提示一条鱼和一个
金锭被映射到显示
金锭和金鱼 (第4.2节)。

- 一个词同时被解释为一个
实体并作为不同实体的修饰符,因此被实现为对象

在场景中以及另一个属性
同一场景中的物体。例如,
提示密封正在打开字母已映射
到显示密封件的图像
信 (图2)。

从视觉上看,我们突出显示的案例映射到以下之一
从图像设计者/用户的角度来看,有两种失败模式:(a)意义模糊的词语
很难分离出来,由此产生的图像往往会表现出意想不到的感觉以及
意图之一 (同音重复)。(b)视觉
图像中一个对象的属性“泄漏”到
图像中的其他物体 (概念泄漏)。⁴

我们还观察到一些二阶效应,
其中修饰词的上位词或它的隐式描述也会触发重复
影响。

总而言之,我们研究的现象为DALLE-2语言能力的局限性提供了
证据,并为未来的研究开辟了道路,以揭示这些局限性是否源于

文本编码、生成模型的问题,

3虽然不在允许这种情况的语法配置下
复制。

⁴重要的是,我们注意到这两种现象很少发生
当提供明确的说明时,例如,提示a
果蝠飞过体育场只产生飞行
哺乳动物。参见哈钦森等人。(2022)讨论
文本到图像生成的规格不足。

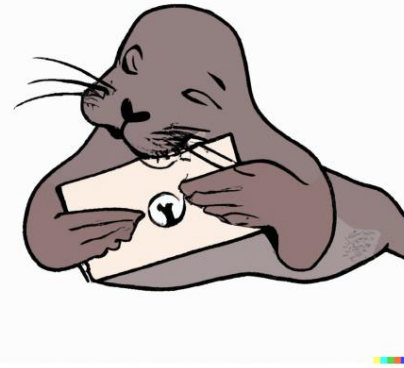


图 2:单词印章既作为对象又作为另一个单词 (字母)的修饰语来实现

提示密封正在打开一封信,从而产生概念
泄漏。提示往往会生成密封的信件,
而不包含的最小差异提示
名词 seal do not。

或两者。更一般地,所提出的方法可以
可以扩展到其他解码场景
过程用于揭示归纳偏差和
文本到图像模型的缺点。

2 以前的工作

DALLE-2引入后立即

(Ramesh 等人, 2022),开始对系统进行分析
浮出水面 (Marcus 等人, 2022 年; Conwell 和 Ull-
man, 2022 年)。Conwell 和 Ullman (2022)表明
DALLE-2不理解空间关系。马库斯等人。(2022)报
告说,如果一个物体
提示中据说有一些属性,那么
图像可能会在某处显示该属性,
然而,不一定是在正确的对象上。我们
进一步分析表明,
相同的属性有时会同时修改
几个不相关的物体。拉梅什等人。(2022)提到我们
分析的现象:他们提到
将单独的属性绑定到单独的对象的问题,并假设这是因为

CLIP (Radford et al., 2021)嵌入不会显式地将属性绑定
到对象。据此,他们
观察到来自解码器的重建
经常混淆属性和对象。并发至
这项工作,哈钦森等人。(2022)讨论了文本到图像生成中的规范不
足,我们发现这种情况与某些相关

我们识别的行为;和匿名(2023)
讨论了几个“泄漏”的案例,并提出了解决方案
干预解码过程的目的
来减轻它们。

3 方法

刺激我们构造语言刺激（提示），它会引发与以下内容不一致的行为：

每个符号单一使用公理。对于第一种情况（被解释为两个实体的单词），我们使用同音异义词，即具有两种不同含义的单词。

DALLE-2通常生成两个对象，一个对应于每种感官。对于修改案例，我们构建以下提示。对于被解释为两个实体的修饰词的单词的情况，我们包括具有两个实体e1和e2的提示以及修饰词m，该修饰词m在语义上与它们两个都兼容，但仅在语法上修改第二个实体（例如，“A Fische1 and a goldm inote2”）。对于同时被解释为实体和修饰语的单词的情况，我们构造包含两个实体e1和e2 的刺激，例如其中一个是在语义上与其他名词相关 例如，“classic buttere1 and花生se2/m”。

附录A.1中的表1、2和3列出了完整的刺激列表以及生成的图像样本。

控制激励为了论证两个实体e1和e2 之间是否存在属性P泄漏，我们需要评估DALLE-2是否不倾向于用该属性来描述e2，而不管 e1 如何。例如，如果DALLE-2在不同上下文中的默认杂货是杂货篮，则很难说正是“杂货附近的篮球”中出现了“篮子”一词，从而引发了DALLE-2生成一个基本的购物篮。考虑到这一点，我们生成了一组控制提示，它们是挑战提示的最小差异变化，其构建目的是通过单独更改n1或n2（具体取决于哪个更改）来防止泄漏的可能性可以防止给定提示中的泄漏。全套控件也详细参见表2和表3。

4 个结果

我们为每个刺激生成 12 个图像，并报告所有刺激和图像的平均值。

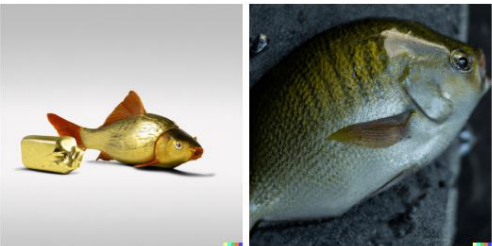
4.1 单个单词实现为多个实体我们有 17 个刺激，引发DALLE-2为单个单词分配两个角色。 216 张图像中 80.3% 出现同音重复。我们发现，脱离上下文，绝大多数词语都是有偏见的



左：一条鲈鱼在太空热带度假胜地闲逛，蒸汽波。右：热门体育赛事中的球迷。

从某种意义上来说，具体来说，在我们的17 个同音异义词中，只有 bass 和 date 揭示了同音异义词的重复，而没有添加上下文。使用贝斯作为提示将返回鲈鱼和吉他贝斯或扬声器贝斯的图像，而对于约会， DALLE-2 将返回一对浪漫的情侣，一起举行约会（水果）。大多数同音异义词重复提示是通过包含与同音异义词的非默认角色相关的上下文来创建的。例如，对于同音风扇， DALLE-2偏向于冷却装置，但如果我们要添加与风扇可能具有的其他可能角色（体育迷）相关的上下文，并且这仍然适用于冷却装置冷却装置，我们将同时获得： 热门体育赛事中的风扇（第 4.1 节顶部）。

4.2 单个词实现为多个实体的修饰语



主要（左）：一条鱼和一个金锭。 对照（右）：一条鱼和一个锭。

给定一对实体(e1, e2),我们观察到未指定的实体e2由DALLE-2 描述,具有e1的属性:在提示A 鱼和金锭中,gold 修饰鱼,并且图像始终包含 a金鱼.6这是

⁵仅使用同音异义词作为提示并不能始终引发包含同音异义词上下文的提示现象。

⁶有趣的是,金鱼也很像金鱼,这表明启动效应也延伸到了化合物,尽管金锭中的“金”一词是具有组合结构的短语的一部分,而“金鱼”不是组合结构。

可能是因为鱼是一个未指定的名词,而黄金是鱼类的友好财产。在控件中,一条鱼和一个元宝、金鱼根本不会出现在输出中。

对于两种属性的情况,我们有 6 个刺激-控制对。刺激提示的输出在 97.2% 的情况下显示共享属性,而控制提示仅在 15.2% 的情况下显示该属性。

4.3 单个词被实现为一个实体和另一个实体的修饰语



主要（左）：斑马和街道。

控制（右）：斑马和碎石街道。

对于实体到属性的情况,我们有 10 个刺激控制对。平均而言,刺激提示在 92.5% 的情况下显示出共享属性,而控制提示仅在 6.6% 的情况下显示出共享属性。

与双属性情况类似, DALLE-2 用 n_1 的属性来描述 n_2 ,但是,这一次, n_1 不是属性,而是实体。为了演示,考虑斑马和街道,这里,斑马是一个实体,但它修改了街道,并且 DALLE-2 不断生成人行横道,可能是因为斑马条纹与人行横道相似。与我们的猜想一致,“斑马和碎石街道”控件指定了一种通常没有人行横道的街道类型,事实上,该提示的所有控件样本都不包含人行横道。当 n_1 是同音异义词时,实体到属性的情况更加微妙,例如,在食物和圆锥形帽子、缩小照片中,单词 cone 通过生成蛋卷冰淇淋来修改食物,而它的控制:食物生日帽、缩小的照片、语义上和物理上对蛋筒的等效替换,不会导致蛋卷冰淇淋的生成。

4.4 二阶刺激



主鸟（左）：一只高大、长腿、长颈的鸟和一个建筑工地。

对照（右）：一只鸟和一个建筑工地。

概念泄漏可以更进一步,通过掩盖影响名词并收到类似的结果:一只高大、长腿、长颈的鸟和一个建筑工地描述了一台起重机和一个建筑工地,但提示中没有任何地方提到了起重机。然而,它会导致 DALLE-2 生成两种类型的起重机:鸟式起重机和建筑起重机。我们还观察到鸟鹤的头部和腿部与鹤具有相同的物理特性。从同样的挑战提示来看,鸟式起重机和建筑起重机之间的界限尤其模糊。当从提示中删除鸟鹤的描述时,不会重复此行为:一只鸟和一个建筑工地。

这表明模型首先将实体映射到概念空间,然后才进行渲染。

4.5 其他文本到图像模型

这项工作的实证重点是 DALLE-2。

在初步实验中,我们发现 DALLE-mini (Dayma 等人, 2021) 一个小得多的模型 没有复制观察到的现象。特别是,仅描述同音异义词的“常见”词义。至于稳定扩散 (Rombach et al., 2021),这种现象似乎出现频率较低,特别是因为在很多时候模型不遵循提示其他模型,例如 (Saharia et al., 2022; Yu et al., 2022),并未公开。我们假设 矛盾的是 这是 DALLE-mini 和 Stable-diffusion 的较低容量事实上,他们没有严格遵循提示,这使得他们在我们检查的缺陷方面显得“更好”。对规模、模型架构和概念泄漏之间关系的彻底评估留待未来的工作。

5.结论

我们证明,在一系列刺激下,
DALLE-2不遵循以下基本原则
语言中符号到实体的映射。这个缺陷是
鉴于DALLE-2是定向的,尤其紧迫
对于普通民众来说,其中大多数人
用户可能没有意识到这种现象。⁷
未来的工作应该将我们诊断的问题追溯到模型中的
特定组件,例如
文本编码或生成解码器,和
研究其对规模和架构的依赖性。此外,事后缓解措施的发展

技术尤其重要,因为大多数
用户缺乏训练模型所需的资源
从头开始。

6 致谢

我们感谢卡洛·梅洛尼的宝贵反馈。
该项目获得了欧洲联盟的资助
欧洲研究委员会 (ERC)
Union 的 Horizon 2020 研究与创新
计划,赠款协议第 802774 号 (iEXTRACT)。

7 局限性

这项工作的一个局限性是关注
DALLE-2是一个商业产品,其来源
代码不公开。正如所讨论的,能力问题阻碍了彻底的检查

我们在其他可用的现象中关注的现象
楷模。我们希望更多类似的型号
今后将向社会公布量表,以便进行更彻底的检查。此外,

我们的分析重点是手动构建的
一组刺激,被设计为表面化
我们以最清晰的方式关注的问题。扩展分析将依赖于我们使用的刺
激的自动或半自动生成,

这是一项具有挑战性的任务。然而,我们相信
如此大规模的分析非常重要,以便
有力地量化我们观察到的问题。

参考

匿名的。2023。[免培训结构化扩散
组合文本到图像合成的指南。](#)

⁷虽然我们关注的问题很容易被忽视
当检查一组随机选择的输出时
大部分看起来都很好 我们精心挑选的检查
刺激清楚地揭示了它们。我们认为这样的分析
很重要,因为人类语言具有许多特征
长尾现象。

提交给第十一届学习表征国际会议。正在审核中。

科林·康威尔和托默·乌尔曼。2022。[测试文本引导图像生成中的关系理
解。](#)

鲍里斯·戴玛 (Boris Dayma),苏拉吉·帕蒂尔 (Suraj Patil),佩德罗·昆卡 (Pedro Cuenca),哈立德·赛夫拉 (Khalid Saiful-lah),
塔尼什克·亚伯拉罕 (Tanishq Abraham),福克·莱·哈克 (Phúc Lê Khac),卢克·梅拉斯 (Luke Melas),
和里托布拉塔·戈什。2021。[达尔·e迷你。](#)

本·哈钦森,杰森·鲍德里奇和维诺德库玛
普拉巴卡兰。2022。场景规格不足
描述到描述的任务。arXiv 预印本
arXiv:2210.05815。

加里·马库斯、欧内斯特·戴维斯和斯科特·阿伦森。2022 年。
[对 dall-e 的非常初步的分析](#)²。

亚历克·雷德福、金钟旭、克里斯·哈拉西、阿迪亚
拉梅什、加布里埃尔·吴、桑迪尼·阿加瓦尔、吉里什
萨斯特里、阿曼达·阿斯科尔、帕梅拉·米什金、杰克·克拉克、
格雷琴·克鲁格和伊利亚·苏茨克弗。2021。[从自然语言中学习可迁移
的视觉模型
监督。](#)

阿迪亚·拉梅什、普拉芙拉·达里瓦尔、亚历克斯·尼科尔、凯西
楚和马克·陈。2022。[具有剪辑潜在特征的分层文本条件图像生成。](#)

罗宾·隆巴赫、安德烈亚斯·布拉特曼、多米尼克·洛伦茨、
帕特里克·埃瑟和比约恩·奥默。2021。[使用潜在扩散模型的高分辨率
图像合成。](#)

Chitwan Saharia,陈伟霆,Saurabh Saxena,Lala
Li,Jay Whang,Emily Denton,Seyed Kamyar Seyed
Ghasemipour,Burcu Karagol Ayan,S. Sara Mah-davi,
Rapha Gontijo Lopes,Tim Salimans,Jonathan
何,大卫·J·弗利特,穆罕默德·诺鲁齐。2022 年。
[逼真的文本到图像扩散模型
深刻的语言理解。](#)

达里奥·D·萨尔武奇 (Dario D Salvucci),尼尔斯·塔特根 (Niels A Taatgen) 和耶尔默·P·博斯特 (Jelmer P Borst)。
2009。走向多任务处理的统一理论
连续体:从并发性能到任务
切换、中断和恢复。SIGCHI 人类因素会议记录

计算系统,第 25 页,第 90 项。

Jiahui Yu,Yuanzhong Xu,Jing Yu Koh,Thang Lu-ong,
Gunjan Baid,Zirui Wang,Vijay Vasudevan,
Alexander Ku,Yinfei Yang,Burcu Karagol Ayan,
本·哈钦森、韩伟、扎拉纳·帕里克、李欣、
张涵、杰森·鲍德里奇和吴永辉。2022 年。
[缩放自回归模型以生成内容丰富的文本到图像。](#)

附录

A.1 数据

完整的图像数据集可以在<https://github.com/RoyiRa>访问
DALLE2-单词到概念映射的缺陷。

下表包含我们用于测试 DALLE-2 参考系统的刺激

迅速的	重复率
一个戴着圆锥帽的人正在吃饭,全身照	7/12
一只蝙蝠飞过棒球场	11/12
懒洋洋地躺在太空热带度假胜地的鲑鱼,蒸汽波	10/12
热门体育赛事的粉丝	12/12
起重机运载重物,背景是鱼	12/12
收集面团的银行家	9/12
两个男人大声吵架,背景是餐厅	10/12
沙漏型身材的模特	12/12
森林里拿着弓的绅士	11/12
一个女人正在往她的眼镜里倒水	9/12
一个人在吃东西的时候被困住了	7/12
一位伟大的统治者	8/12
苹果广告	12/12
一个有金属钉子的人约会	6/12
	7/12
一个有鸡腿的人,全身像董事会会议	12/12
	9/12

表 1:多实体案例的刺激列表（第 4.1 节）。映射到两个实体的名词是
用粗体表示

挑战并控制乱穿马路的人和一	挑战比例	控制比例
只鸟	11/12	0/12
一个人非法过马路,还有鸟食和锥形帽子,缩小了照片	12/12	0/12
食物和生日帽,缩小照片	12/12	0/12
一条鱼和一块金锭	12/12	0/12
一条鱼和一个元宝		
杂货店附近的篮球	12/12	5/12
杂货店附近的足球		
瓢虫和一个人,全身形象	11/12	6/12
一只虫子和一个人,全身形象		
存钱罐旁边的动物	12/12	0/12
桌子旁边有一只动物		

表 2:单词到多个修饰符情况下的质询和控制提示（第 4.2 节）。中的修饰名词
挑战提示用粗体表示,修饰名词用斜体表示

挑战与控制	挑战比例	控制比例
斑马和街道	10/12	0/12
斑马和碎石路		
封印正在打开一封信	10/12	1/12
一只大型水生哺乳动物正在打开一封信		
两个人拿着食物打壁球	12/12	0/12
两个人在台球和桌子上打壁球		
游泳池和酒吧	10/12	0/12
一个人正在用多芬洗发水洗头	12/12	0/12
一个人正在用洗发水、玃狨和大象洗头发		
犀牛和大象 鱼和大象	12/12	0/12
一头犀牛和一头大象	12/12	0/12
十字架和人行道	10/12	2/12
十字架和人行道		
经典黄油和花生	12/12	0/12
经典黄油和黄瓜		
一只豹子和一块布	11/12	5/12
豹子和黑毛巾		

表 3:单词到实体和修饰符情况下的质询和控制提示（第 4.3 节）。修饰符挑战提示中的名词用粗体表示,修饰的名词用斜体表示

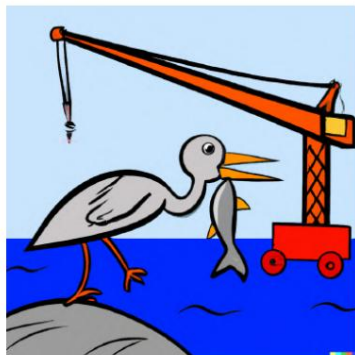
挑战与控制	挑战比例	控制比例
海边的玃狨	12/12	0/12
一只狗在海边		
鳍足类动物正在打开一封信	7/12	1/12
一只大型水哺乳动物正在打开一封信,一只高大、长腿、		
长颈的鸟和一个建筑工地	12/12	0/12
一只鸟和一个建筑工地		
一个人和一只红点的虫子,全身图像	11/12	6/12
一只虫子和一个人,全身形象		

表 4:二阶概念泄漏中的挑战和控制提示（第 4.4 节）。修饰名词挑战中的提示用粗体表示,修饰的名词用斜体表示

单个单词实现为多个实体



一个戴着圆锥帽的人正在吃饭,全身照



起重机运载重物,背景是鱼



热门体育赛事的粉丝



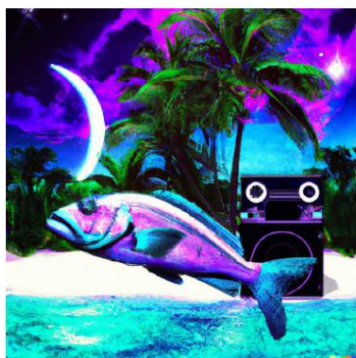
一只蝙蝠飞过棒球场



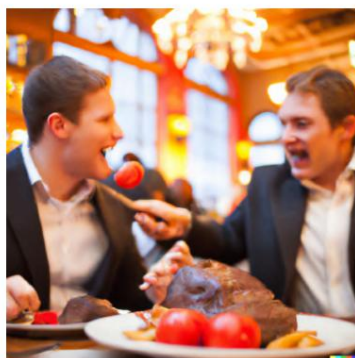
收集面团的银行家



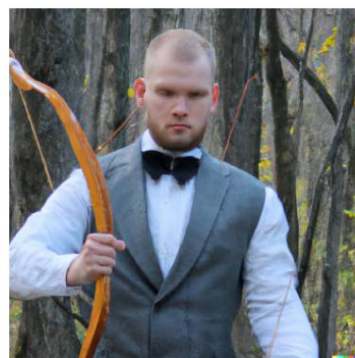
沙漏型身材的模特



懒洋洋地躺在太空热带度假胜地的鲑鱼,蒸汽波



两个男人大声吵架,背景是餐厅



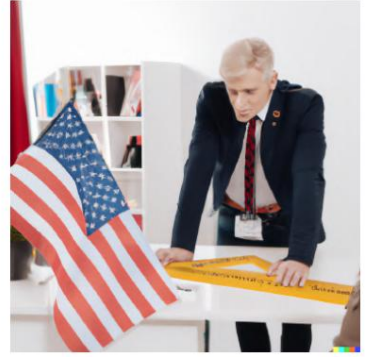
森林里拿着弓的绅士



一个女人正在往她身上倒水
眼镜



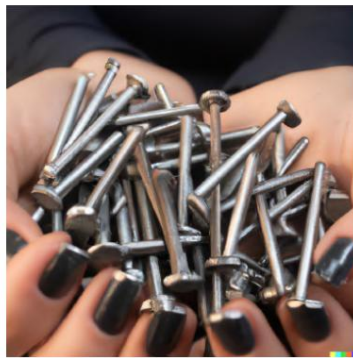
一个人在吃东西的时候被困住了



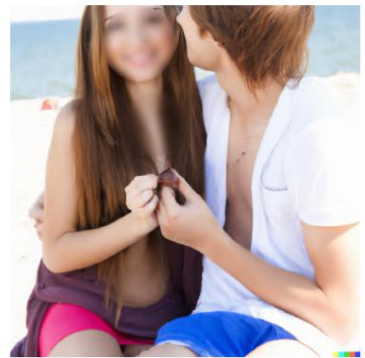
一位伟大的统治者



苹果广告



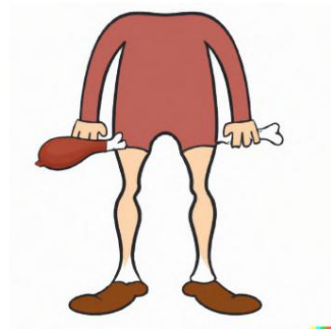
一个有金属钉子的人



日期



董事会会议



一个人有鸡腿,饱了
身体形象

一个词充当两个不同实体的修饰语

粗体项是实体,粗体+下划线>是修饰符。



主要 (左) :乱穿马路的人和一只鸟
对照 (右) :非法过马路的人和一只鸟



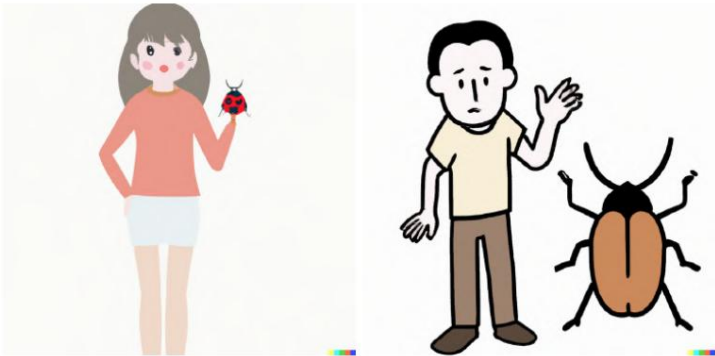
主图 (左) :食物和锥形帽,缩小照片
对照 (右) :食物和生日帽,缩小照片



主图 (左) :鱼和金元宝
对照 (右) :鱼和锭



主要 (左) :杂货店附近的篮球
控制 (右) :杂货店附近的足球



主要 (左) :瓢虫和一个人的全身图像
对照 (右) :一只虫子和一个人,全身图像



主要 (左) :存钱罐旁边的动物
对照 (右) :桌子旁边的动物

同一单词实现为一个实体以及另一个实体的修饰语

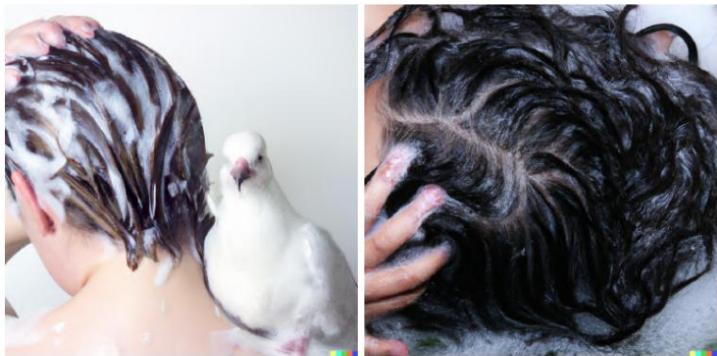
粗体项是实体,粗体+下划线既充当修饰符又充当实体。 _____



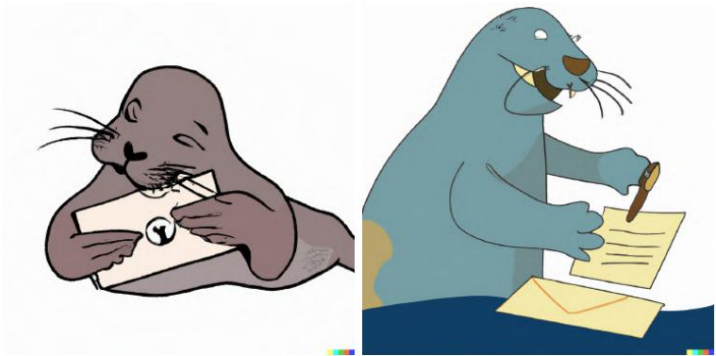
主要 (左) :斑马和街道
控制 (右) :斑马线和碎石街道



主图 (左) :两个人拿着食物打壁球
对照 (右) :两个人打壁球



主图 (左) :一个人正在用多芬洗发水洗头
对照 (右) :一个人正在用洗发水洗头



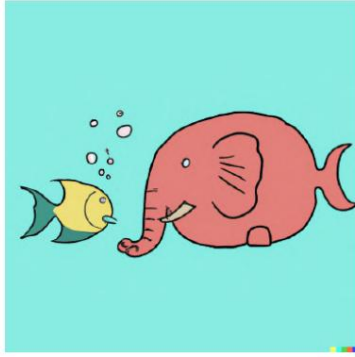
主图 (左) :印章正在打开一封信
控制人员 (右) :一只大型水生哺乳动物正在打开一封信



主要 (左) :一个游泳池和一张桌子
控制区 (右) :游泳池和酒吧



主要 (左) :狢狢和大象
对照 (右) :犀牛和大象



主要（左）：鱼和太象

对照（右）：犀牛和大象



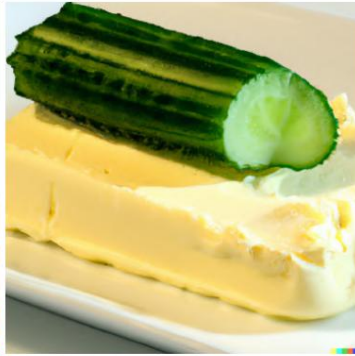
主要（左）：十字架和人行道

控制（右）：十字架和人行道



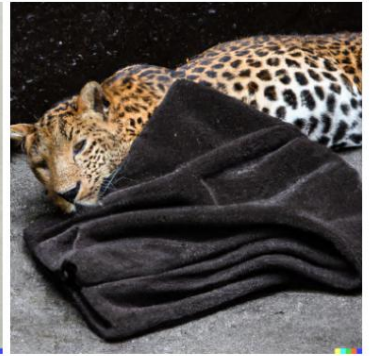
主菜（左）：经典黄油和花生

对照（右）：经典黄油和黄瓜



主要（左）：一只豹子和一块布

对照（右）：豹子和黑毛巾



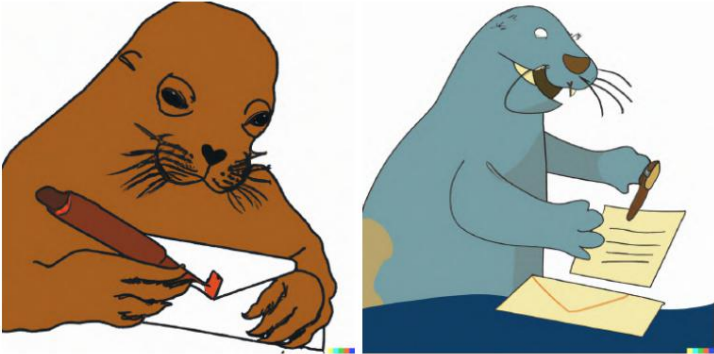
二阶概念泄漏



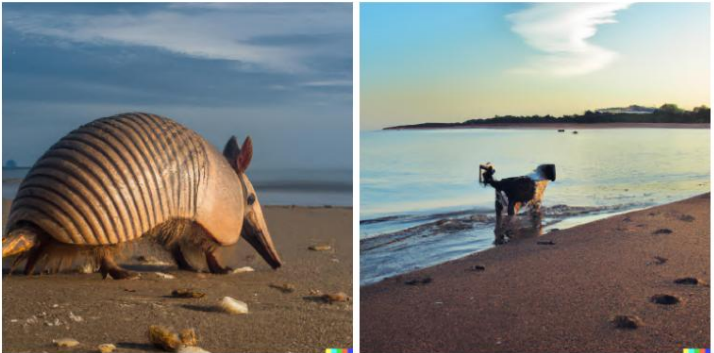
主图（左）：一只高大的长腿长颈鸟和一个建筑工地控制图（右）：一只鸟和一个建筑工地解释:这只高大的长腿长颈鸟被识别为起重机实现为两种起重机。



主图（左）：一个人和一只只有红点的虫子,全身图对照（右）：一个虫子和一个人,全身图说明:有红点的虫子被识别为瓢虫,该人是女性（女士）。



主（左）：鳍足类动物正在打开一封信控制（右）：大型水生哺乳动物正在打开一封信说明:鳍足类动物被解释为海豹,这既作为演员又作为信封上的印章实现。



主图（左）：海边的犴獾对照（右）：海边的狗说明:犴獾有壳,海滩上也有贝壳。