ANNA：使用经过筛选的新闻图注进行抽象式文本到图像合成

Aashish Anantha Ramakrishnan Sharon X. Huang Dongwon Lee宾夕法尼亚州立大学，斯泰特学院，宾夕法尼亚州，美国{aza6352，suh972，dongwon}@psu.edu

arXiv:2301.02160v2 [cs.CV] 1 Jul 2024

摘要

近年来，文本到图像合成领域的发展更多地聚焦于使用带有描述性提示词的数据集来提升生成样本的质量。然而，新闻数据等领域中存在的现实世界图像-图注对，并不会使用简单且直接的描述性提示词。由于图注同时包含图像内容和潜在上下文线索的信息，因此它们在本质上属于抽象式图像描述。在本文中，我们推出ANNA，这是一个从在线新闻文章中提取的抽象式新闻图注数据集。我们通过在ANNA上进行基准测试，探索当前文本到图像合成模型在零样本和微调设置下，使用抽象式图像描述生成新闻领域特定图像的能力。生成的图像将根据视觉质量以及与真实样本之间的图像-描述对齐程度进行评判。通过实验，我们表明，不同的微调技术在理解抽象式图像描述方面取得的成功有限，但仍无法持续学习句法特征与上下文特征之间的关系。该数据集可在 https://github.com/aashish2000/ANNA获取。

1 引言

近年来，得益于生成建模方法的进步以及更强计算能力的普及，图像生成取得了突飞猛进的发展（Tsirikoglou 等人，2020年）。随着基于从互联网海量图文对中挖掘的数据进行模型预训练的范式不断发展（Schuhmann 等人，2021年），文本到图像（T2I）合成等领域日益受到关注。这推动了生成器在各种应用中的使用，例如在线内容创作、艺术合成（Ramesh 等人，2022年），以及深度伪造生成等更具恶意的用途（Zhang，2022年）。随着互联网新闻媒体和社交网站逐渐成为人们首选的信息传播形式



图 1：COCO 图注数据集（Chen 等人，2015年）中的描述性图像描述示例（左），以及 ANNA 的抽象式图像描述示例（右）。以 蓝色突出显示的词语直接对应视觉实体，而以 红色突出显示的词语则展现了其中的上下文线索和句法变化。

生成建模，尤其是语义相关的图像生成，可能对新闻媒体行业产生相当重大的影响。

通常，文本到图像合成会采用描述性提示词，将显著的视觉对象列为关键词，并说明它们的空间排列方式和物理属性。然而，现实世界中的图像描述（例如新闻媒体中的图像描述）在语言表达上丰富多样，旨在向读者传达尽可能多的信息，同时不重复显而易见的事实（Federico，2016年）。这在新闻媒体中十分重要，因为图像用于辅助文章叙事，并向读者提供额外信息。非描述性图像描述会运用句法意识（Deacon 和 Kieffer，2018年）和上下文线索，以高效传达情境信息。语用推理等语言学理论（Grice，1975年）支持这样一种观点：文本的信息量受其与上下文相关性的影响，尤其是在新闻图注中（Vedantam 等人，2017年）、（van Miltenburg 等人，2016年）（Nie 等人，2020年）。新闻图注的抽象程度也各不相同，

有些示例只关注图像可见特征中的一小部分，以传达特定信息。大型视觉与语言模型已展现出涌现的常识推理能力，能够处理各种理解任务。我们选择从**抽象式**（超越描述性）图像描述生成新闻图像这一任务，来评估它们生成高度匹配结果的能力。

当前用于文本到图像合成的数据集，要么聚焦于狭窄领域并配有简单、描述性的提示词，要么包含来自众多线上来源、仅经过最低限度筛选的图文对。包含情境信息以及图像描述的领域专用数据集并不多见。此外，尽管大多数模型都将输出图像更高的视觉质量作为性能优越的指标，但人们很少关注评估输出图像与输入文本图像描述之间的相关性。当处理其特征由于非描述性而仅与真实图像部分对齐的图像描述时，这一点就变得更加重要。为评估新闻图像合成任务，我们设计了 ANNA，这是一个包含抽象式新闻图像描述对的数据集。抽象式图像描述可以促使文本到图像合成模型有效识别这些不同的特征类型及其相对重要性，并在生成的图像中恰当地呈现这些特征。鉴于当前的文本到图像架构会隐式地区分视觉内容与情境特征，我们在 ANNA 上详细展示了其成功与失败案例的可视化结果，以及为了生成图像特征而更好地理解句子结构的必要性。

本文的贡献可概括如下：

- 我们介绍了 ANNA，这是一个包含约 3 万组来自热门媒体组织的抽象式图像描述图像-图注对的数据集
- 我们展示了当前开源的 T2I 架构如何通过零样本和微调设置理解抽象式图像描述
- 我们采用一套详尽的评估指标，从生成图像质量、与真实图像的相似度，以及人类偏好与参考描述的一致性等方面，对流行的文本到图像架构进行基准评测

2 相关工作

文本到图像合成 生成对抗网络（GAN）（Goodfellow 等人，2014年）的出现推动了文本到图像生成领域的重大创新。早期基于 GAN 的方法利用注意力网络，在选定的数据集中学习图像与文本之间的关联（Zhang 等人，2017年）、（Xu 等人，2018年）、（Zhu 等人，2019年）。然而，这些模型需要大量特定任务的数据才能取得良好性能。随着基于 Transformer 的架构（Vaswani 等人，2017年）通过从海量线上数据集中学习、在视觉和语言特征编码方面取得更大成功，诸如 CLIP（Radford 等人，2021年）这样的预训练多模态编码器显著提升了多模态嵌入的质量，并提供了更好的输入条件（Crowson 等人，2022年）、（Zhou 等人，2022年）。扩散模型（Sohl-Dickstein 等人，2015年）的发展进一步提升了生成图像的视觉保真度。扩散模型支持生成更高分辨率的图像，并能对生成过程进行细粒度控制（Nichol 等人，2022年）、（Ramesh 等人，2021年）、（Ding 等人，2021年）。随着基于语言模型的纯文本编码器取得成功，大型视觉模型采用基于大型语言模型（LLM）的编码器进行文本到图像生成，利用其语言理解能力（Saharia 等人，2022年）。文本到图像个性化是另一个重要研究方向，模型借助少量样本示例进行训练，以生成特定概念和物体的图像（Kumari 等人，2023年）、（Ruiz 等人，2023年）。

潜在扩散模型 潜在扩散模型（LDM）是一类将训练过程分为两个不同阶段的扩散模型。这些模型在扩散过程中使用一种经压缩的潜在空间，该空间在感知上等效。为此，模型采用自动编码器模型，从而减少总体可训练参数的数量。通过使用基于图像块的对抗目标和感知损失进行训练，可确保潜在空间中的信息损失最小。第二阶段的训练通过为 UNet 增加交叉注意力层，改进扩散模型的条件机制。

数据集 近年来，文献中逐渐转向采用仅评估基准，通过提示词来评判生成器的特定属性

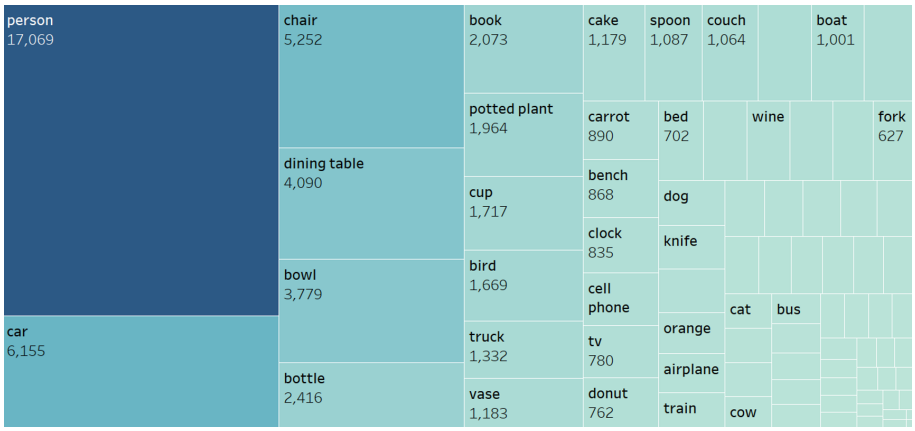


图2：使用矩形树图分析物体频率

表现。PartiPrompts（Saharia 等人，2022年）和 UniBench（Li 等人，2022年）提供了多样的文本提示词，并按风格和难度进行分类。DiffusionDB（Wang 等人，2022年）是一个大规模的提示调优的图像描述对集合，常用于为文生图评估提供图像描述。所有这些基准测试中的图像描述，都只对图像中的物理实体进行简略或详细描述。我们旨在将包含情境上下文信息和复杂句子结构的图像描述纳入ANNA。

3 数据集构建

该ANNA （抽象式新闻图注数据集）旨在使用抽象式图像描述执行新闻图像生成任务。我们的图像取自 NYTimes800K 数据集（Tran 等人，2020年），该数据集包含从新闻机构《纽约时报》（NYT）抓取的新闻文章及其关联的图像-描述对。该数据集最初是为新闻图像描述任务开发的。使用来自 NYT 等信誉良好的媒体机构的新闻图像-描述对，有助于确保数据集的质量。我们着重在数据集中选择具有泛化能力的实体，以观察不同句子结构的影响，以及它们如何影响生成的图像。新闻数据包含大量命名实体，这些实体通常重复率很低，且具有鲜明的外观特征，例如人脸和地理地标。将新闻图像中的命名实体纳入其中，会大幅增加生成任务的复杂度。由于命名实体属性彼此依赖，无法准确生成这些属性还会进一步妨碍上下文特征的表示。为了解决上述问题，我们精心整理数据集，纳入包含充足上下文信息和内容相关信息的图像-描述对。我们选择对命名实体依赖较少、包含更多通用视觉元素的样本，使任务切实可行。

3.1 预处理与筛选方法

原始 NYTimes800K 数据集包含 44.5 万篇新闻文章，以及 79.3 万个图像-图注对。这些文章涵盖了 14 年间发表于 NYT 网站的内容。该数据集以 MongoDB 数据转储的形式提供，供公众访问。预处理的第一步是移除图像和文本中均包含明确实体提及的图像-图注对。我们使用每条图像描述附带的实体标签进行筛选。我们排除所有带有 'PERSON'、'GPE'、'LOC'、'WORK_OF_ART' 和 'ORG' 标签的图像描述。如示例2所示，一些常见的 'GPE'、'LOC' 和 'ORG' 实体在真实标注中并未标记，因此我们将其保留在数据集中。这确保了从数据集中移除不常见的视觉实体。随后，我们还将图像描述的长度限定在 4 至 70 个词之间。少于 4 个词的图像描述不足以提取可用特征，而超过 77 个词的图像描述则无法由我们在实验中采用的基于 CLIP 的文本编码器（Radford 等人，2021年）处理。

经过基于图注的筛选后，我们移除所有前景中清晰可见人脸的图像。使用基于 RetinaFace 的人脸检测器（Deng 等人，2020年），我们又移除了约 1000 张图像。通过这些筛选技术，我们从 NY-Times800K 数据集中提取相关的图像-图注对及对应的文章标题。数据集中的所有新闻图像都统一调整为目标输入分辨率

512×512。为定位源图像中包含感兴趣物体的区域，我们采用基于熵的区域选择（Ma 和 Guo，2004年）。这确保信息损失尽可能少，并有助于优先选择每张图像中的前景物体。免责声明：该数据集取材于真实世界新闻文章，因此可能包含一些读者认为亵渎、粗俗或冒犯性的语言。

3.2 数据集洞察

经过筛选和预处理的 ANNA 版本包含 29625 个图文对。我们按 80%:10%:10% 的比例将数据集划分为训练集、验证集和测试集。报告的所有指标分数均在测试集上计算。为了更好地了解数据集的构成，我们分析了图文对及其来源文章的各种属性。

数据集	唯一词元	标题长度	
		Mean	标准差
COCO Captions 训练集	22767	10.42	0.88
COCO Captions 验证集	16647	10.42	0.87
CC3M 训练集	45896	10.31	3.30
CC3M 验证集	9289	10.40	3.35
ANNA 训练集	17897	**14.1**	**7.75**
ANNA 验证集	1622	13.8	7.60
ANNA 测试集	1649	14.1	7.71

表 1：ANNA 数据集统计

3.2.1 图注统计

在本节中，我们评估了不同的统计度量，以量化数据集中图像描述的分布。表 1 展示了数据集以及训练集、验证集和测试集子集中图像描述的平均长度。不同数据划分的平均图像描述长度相近，且略长于 COCO 图像描述数据集中的平均长度。图像描述长度的标准差也更高，表明其中既有短篇、长篇，也有多行描述。为直观呈现 ANNA 中图像描述的词元多样性，我们通过识别唯一词元的出现情况来考察其中的词语。唯一词元数量使用 spaCy 库计算；该库会对图像描述进行分词和词形还原，并移除所有停用词。随后，对不同词性（POS）进行标注，并选取属于 [普通名词、专有名词、形容词和动词] 类别的唯一词元。这可以提供一个最低限度的保证，确保其中的抽象式图像描述足够长，能够包含充足的内容和上下文特征。我们的分析还确保了训练集、验证集和测试集中图像描述的构成保持一致。

3.2.2 新闻图像分析

除了图像描述，我们还估算图像属性，例如每张图像中可识别物体的数量，以及每张图像检测到的物体平均数量。我们使用基于 YOLO-R 的目标检测器（Wang 等人，2021年）来识别数据集中每张图像中的物体。YOLO-R 检测器使用包含 80 种常见物体类别的 MS-COCO 数据集进行训练（Chen 等人，2015年）。YOLO-R 推理使用 0.4 的置信度阈值。我们发现，ANNA中每张图像平均有 2.57 个物体。图 2 使用矩形树图对数据集中最常出现的物体类别进行可视化。

3.2.3 所选新闻文章类别

本节将介绍用于数据集构建的图像-图注对所来源的新闻文章类型。我们的数据集中共有 123 种不同的文章主题。只有 13 个图像-图注对没有相应的文章类型信息，因此我们在文章主题分析中将其忽略。从图 3中可以看出，Dining、商业、房地产等主题的分布较为均衡。这表明新闻图像-图注对具有多样性，并非仅限于某一特定类型的新闻文章。

4 评估方法

为了了解不同架构如何在 ANNA 上学习抽象式图像描述，我们考察了文献中此前提出的各种文本到图像合成模型。我们为此任务选取了两种主要架构：潜在扩散模型和基于 StyleGAN 的生成器。由于这些模型在描述性图像描述数据集上均取得了最先进的成绩，我们在实验中评估它们在新闻领域特定的抽象式图像描述上的表现，并对结果进行可视化。所有图像均以 512×512 分辨率生成。

文本到图像合成模型我们选择最先进的文本到图像合成模型作为基线，包括稳定扩散（SD）V1.4、V1.5 和 V2.1（Rombach 等人，2022年），用于

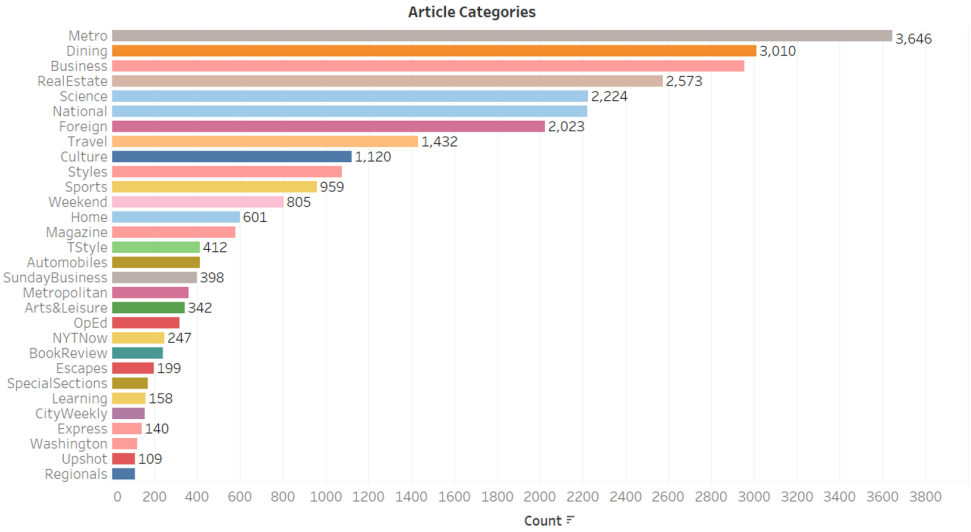


图3：ANNA 中图像-图注对文章类别可视化

模型	FID_{CLIP} (↓)	ImageReward (↑)	HPS V2 (↑)
Stable Diffusion 2.1（ReFL）	9.728	**0.2182**	**0.2470**
Stable Diffusion 2.1（LoRA）	**7.5906**	-0.0081	0.2335
Stable Diffusion 2.1（基础版）	7.9707	0.1041	0.2399
Stable Diffusion 1.5（基础版）	7.7008	-0.0094	0.2312
Stable Diffusion 1.4（基础版）	7.7178	-0.0104	0.2314
Lafite（微调版）	12.5602	-0.933	0.1809
Lafite（基础版）	20.8314	-1.3477	0.1685

表 2：ANNA 上抽象式文本到图像合成的结果

新闻图像生成任务。它们属于潜在扩散模型（LDM），在 COCO Captions（Chen 等人，2015年）等传统基准测试中表现出色。我们还评估了 Lafite（Zhou 等人，2022年），该模型利用预训练的 CLIP 编码器将文本嵌入映射到图像特征空间。它通过仿射变换注入文本条件信息，从而适配无条件 StyleGAN2 生成器（Karras 等人，2020年）。模型使用两个全连接层，将输入文本特征转换为在语义上更接近 StyleGAN图像样式空间的表示。

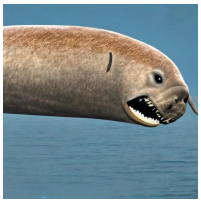
微调方法 我们在Hu 等人，2021年提出的 SD V2.1 低秩适配（LoRA）和奖励反馈学习（ReFL）方法（Xu 等人，2023年）上测试了两种参数高效微调方法，使用的数据集是ANNA 。微调过程中，LoRA 不会直接修改模型权重，而是为模型中的每个 Transformer 层注入可训练的秩分解矩阵。这使其能够在较小的数据集上更快收敛，并获得更好的泛化能力。另一方面，ReFL 在则使用奖励模型向生成器提供反馈，而不是直接在数据集上进行微调。这些奖励模型经过训练，可预测给定图像-描述对的人类偏好分数，并将其作为图像-描述对齐的间接指标。ReFL 仅使用图像描述来微调生成器，无需额外的图像数据。

指标 为了全面评估文本到图像模型在ANNA上生成的样本，我们报告三种不同类型的指标：弗雷歇 Inception 距离（FID）（Heusel 等人，2017年）、ImageReward（Xu 等人，2023年）和人类偏好分数（HPS）V2（Wu 等人，2023年）。FID 用于衡量生成样本相较于真实图像的整体真实感和多样性。由于 ANNA 等数据集的分布与传统 FID 计算所使用的 Inception-V3 存在显著差异（ANNA 因此，我们在测试集中采用更具代表性的Kynkäänniemi 等人，2022年）指标。为了衡量生成图像与真实描述之间的相关性，我们使用 ImageReward。与图像—描述相似度相比

原图	SD 1.5（基础模型）	SD 2.1（基础模型）	SD 2.1（LoRA）	SD 2.1（ReFL）







说明文字：从它的牙齿来看，这种杂交动物的饮食可能更接近海象或髯海豹，而非独角鲸或白鲸。







说明文字： 一名抗议者爬上自由女神像的基座，迫使纪念碑关闭。岛上的约 4,500 名游客被疏散，几小时后，攀登者被安全拘留。







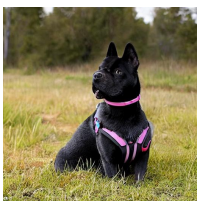
说明文字： 厨房的蓝白相间瓷砖防溅板很漂亮。

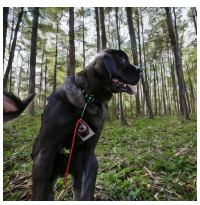






说明文字： 一片死树果园，位于 Pouze Khoon 村附近。十年前，开心果林还是一片葱绿。如今这里滴雨未下，地下水几乎已经枯竭。




说明文字： 有了 RoamEO 基站（左侧，包含一个项圈），狗主人就能接收无线电信号，追踪最远 1.5 英里外动物的位置。



说明文字： 与野生动物保护中心里的这只西表山猫标本不同，活体标本行踪难觅，鲜少露面。

图4：不同文本到图像模型在 ANNA 数据集上的定性比较

诸如 CLIPScore（Hessel 等人，2021年）等指标。ImageReward 使用真实世界中经人工偏好标注和排序的图像-图注对进行训练。同样，HPS V2 也可作为偏好对齐程度的指标。

5 实验

实现细节 在实验中，我们以全监督方式在 ANNA 数据集上训练 Lafite。我们训练了两个版本的 Lafite，分别进行和不进行微调。在基础版本中，我们从头开始在 ANNA 数据集上训练模型，直至收敛，共训练 4400 个 epoch。为进行微调，我们使用 Conceptual Captions（CC3M）数据集的预训练权重（Sharma 等人，2018年）初始化模型，并在 ANNA 数据集上继续训练 1800 个 epoch，直至收敛。两个 Lafite 模型使用的超参数为：itd = 5、itc =10、gamma = 10 和 temp = 0.5。对于稳定扩散模型，我们将 guidance_scale=*7.5*、mixed_precision=fp16 和 inference_steps=*100* 设为相应值。我们将 max_grad_norm=*1* 设为相应值，并进行 1000 个 epoch 的 LoRA 微调。类似地，对于 ReFL，我们将 max_grad_norm=*2*、grad_scale=*0*.001 设为相应值，并进行 200 个 epoch 的微调。两种方法均使用 batch_size=*2*、grad_accumulation_steps=*4*，将 use_ema=设为 True 、 learning_rate=*1*e-05，并将 lr_scheduler=设为 constant。所有报告的指标均为随机种子 42 和 973 下结果的平均值。表 2 展示了定量结果，而图 4 则对我们在 ANNA 数据集上的实验进行了定性比较。

5.1 定量结果

查看表2中的结果可以发现，不同类型的微调稳定扩散模型在特定指标上的表现优于其零样本版本。这支持了我们的假设：预训练数据集主要包含描述性提示词，而非图像描述。根据报告的 FID_{CLIP} 分数，我们发现，稳定扩散（LoRA）在生成高保真样本且与真实图像分布高度一致方面表现最佳。另一方面，稳定扩散（ReFL）在 ImageReward 和人类偏好评分 V2 方面表现最佳。这表明，图像质量与图像-描述相关性及人类偏好并无直接关联。传统的文本到图像模型仅使用图像质量指标来量化模型性能，而这些指标可能并非评估抽象式描述数据集的最佳依据。

5.2 定性结果

从图4的可视化结果中，我们可以观察到，经过微调的稳定扩散模型能够准确捕捉图像描述中提及的关键主体。在示例 1中，图像描述主要描述了动物的牙齿。与所有其他模型相比，Stable Diffusion 2.1（ReFL）展现出更强的句法意识，生成的动物牙齿图像也更准确。同样，在示例 2中，未经微调的模型未能捕捉图像中的攀登者。此外，这些模型也无法理解图像描述中的上下文线索：图像描述明确指出自由女神像周围的观众已被疏散，它们却仍然生成了多名观众。只有 Stable Diffusion 2.1（ReFL）能够充分展现雕像和攀登者的细节。这表明，使用抽象式图像描述进行微调，可以显著提升模型对上下文和内容特征的理解。不过，在某些情况下，即使是经过微调的模型，也无法捕捉图像描述中的关键主体。在示例4中，图像描述将枯树森林描述为图像的主体。然而，所有模型生成的森林里，树木仍然长着绿叶。在

示例 6中，所有模型都未能生成毛绒玩具，反而生成了真正的野猫。当需要在不常见的非典型上下文中生成重要对象时，就会出现这一问题。通过使用更多抽象式图像描述测试这些模型，我们可以更好地理解当前文本到图像模型的局限性，并制定策略以持续提升其性能。

6 讨论与结论

我们的实验展示了现有开源文本到图像架构如何理解新闻媒体等领域特定数据中的抽象式图像描述。我们表明，在评估复杂度更高的图像描述上的文本到图像合成时，仅关注图像质量或图像-描述相关性是不够的。我们提出的数据集ANNA有助于推动开发语言理解能力更强的文本到图像模型，使其能够在生成过程中整合推理和上下文线索。尽管现有的微调方法在提升模型性能方面展现出潜力，但要同时优化生成图像的图像保真度和人类偏好对齐，它们仍面临困难。这凸显了开发更复杂架构的必要性，使其能够理解抽象式图像描述中不同特征类型的细微差别。随着数据集规模持续扩大，扩大图像人工标注的规模以满足需求会带来巨大的额外负担。描述性提示词需要与参考图像的内容紧密对应，这需要经过多轮评估和筛选，因而成为一项繁琐的人工任务。使用图像的抽象式图像描述可以大幅简化数据集的人工标注流程。此外，ANNA也推动了新闻辅助解决方案的发展。使用关键词和描述性提示词配合当前的图像生成器，需要大量提示工程，才能生成与特定主题相关的图像（Liu 和 Chilton，2022年）。只有采用限制性极强的句式结构和词汇，才能生成高质量图像。经过训练、能够理解抽象式图像描述的模型，可大幅降低对繁复提示工程的需求。同样，更好地区分非描述性提示词中的不同特征类型，也有益于图像检索等相关任务。加入上下文对影响检索结果的质量起着重要作用。

局限性 本文旨在介绍抽象式图像描述的潜力，以推动开发上下文 grounding 能力更强的文本到图像合成模型，尤其是在合成新闻领域特定图像时。尽管新闻文章包含许多命名实体，我们选择将其过滤掉，转而关注可从文本图像描述中推断、并能通过通用视觉概念呈现的上下文特征。开发能够利用外部知识库作为参考、借助命名实体的文本到图像合成架构，将有助于克服这一局限。我们评估方法面临的另一项挑战是 CLIP 等多模态编码器的提示词最大长度限制。由于输入提示词长度限制为 77 个词元，我们能够有效评估的新闻图像生成任务受到标题长度的制约。为应对这一挑战，我们计划尝试其他能够接受更长抽象式图像描述、基于大语言模型的多模态编码器。从指标选择的角度来看，对文本到图像架构根据抽象式图像描述生成的图像开展大规模人工评估，是衡量其相对性能的重要一步。我们计划在未来研究中开展人工评估研究。

潜在的负面社会影响 图像生成架构可能会被滥用于恶意用途，例如传播虚假信息（Zhang，2022年）和生成神经网络生成的假新闻（Zellers 等人，2019年）。我们当前的预处理流程会移除大多数包含命名实体的图像，即公众人物和国家重要地点，从而降低风险。不过，我们也意识到，在处理新闻媒体图像时，上下文相关的深度伪造图像会带来威胁。未来研究方向包括了解文本到图像模型在多大程度上可用于神经网络假新闻生成，以及确定适当的检测策略。

致谢

本研究部分由美国国家科学基金会（NSF）项目 #1820609 资助。部分研究成果使用了 CloudBank 提供的计算资源（https://www.cloudbank.org/），该资源由 NSF 项目 #1925001 资助。

参考文献

Xinlei Chen、Hao Fang、Tsung-Yi Lin、Ramakrishna Vedantam、Saurabh Gupta、Piotr Dollar 和 C La wrence Zitnick。2015年。Microsoft COCO captions：数据收集与评估服务器。*arXiv* 预印本 *arXiv:1504.00325*。

Katherine Crowson、Stella Biderman、Daniel Kornis、Dashiell Stander、Eric Hallahan、Louis Castricato 和 Edward Raff。2022年。VQGAN-CLIP：通过自然语言引导进行开放领域图像生成与编辑。*arXiv:2204.08583 [*计算机科学*]*。

S Hélène Deacon 和 Michael Kieffer。2018年。理解句法意识如何促进阅读理解：来自中介效应和纵向模型的证据。《教育心理学杂志》，110(1)：72–86。

Jiankang Deng、Jia Guo、Evangelos Ververas、Irene Kotsia 和 Stefanos Zafeiriou。2020年。Retinaface：真实环境中的单次拍摄多层级人脸定位。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第5203–5212页。IEEE。

Ming Ding、Zhuoyi Yang、Wenyi Hong、Wendi Zheng、Chang Zhou、Da Yin、Junyang Lin、Xu Zou、Zhou Shao、Hongxia Yang 和 Jie Tang。2021年。CogView：通过变换器掌握文本生成图像。

Stephanie Federico。2016年。这是 NPR 的照片说明指南。*NPR*。

Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。2014年。生成对抗网络。载于神经信息处理系统进展，第27卷。Curran Associates, Inc.

Herbert P Grice。1975年。逻辑与会话。载于言语行为，第41–58页。Brill。

Jack Hessel、Ari Holtzman、Maxwell Forbes、Ronan Le Bras 和 Yejin Choi。2021年。CLIPScore：一种用于图像描述生成的无参考评估指标。载于*2021*年自然语言处理实证方法会议论文集，第7514–7528页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。2017年。采用双时间尺度更新规则训练的生成对抗网络会收敛到局部纳什均衡。载于神经信息处理系统进展，第30卷。Curran Associates, Inc.

Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2021年。LoRA：大语言模型的低秩适配。*arXiv [*计算机科学*.CL]*。

Tero Karras，Miika Aittala，Janne Hellsten，Samuli Laine、Jaakko Lehtinen 和 Timo Aila。2020年。使用有限数据训练生成对抗网络。载于第*34*届神经信息处理系统国际会议论文集，NIPS’20 中第1015篇文章，第12104–12114页，美国纽约州雷德胡克。Curran Associates Inc. Nupur Kumari、Bingliang Zhang、Richard Zhang、Eli Shechtman 和 Jun-Yan Zhu。2023年。文生图扩散的多概念定制。载于*2023*年 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）。IEEE。

Tuomas Kynkäänniemi、Tero Karras、Miika Aittala、Timo Aila 和 JaakkoLehtinen。2022年。ImageNet类别在 Fréchet初始距离中的作用。*arXiv [*计算机科学*.CV]*。

Daiqing Li、Huan Ling、Seung Wook Kim、Karsten Kreis、Sanja Fidler 和 Antonio Torralba。2022年。BigDatasetGAN：通过像素级标注合成 ImageNet。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第21330–21340页。

Vivian Liu 和 Lydia B Chilton。2022年。文本到图像生成模型提示工程的设计指南。载于*2022*年计算机系统中的人为因素会议论文集，CHI ’22 第384号文章，第1–23页，美国纽约州纽约市。计算机协会。

Matthew Ma 和 Jinhong K Guo。2004年。面向内置摄像头移动设备的自动图像裁剪。在 *2003*年国际 *VLSI* 技术、系统与应用研讨会。技术论文集。（*IEEE* 编号：*03TH8672*），第 710–711 页。IEEE。

Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。2022年。GLIDE：利用文本引导的扩散模型实现照片级真实感图像生成与编辑。*arXiv:2112. 10741 [*计算机科学*]*。

Allen Nie、Reuben Cohn-Gordon 和 Christopher Potts。2020年。语用学问题敏感图像描述。在计算语言学协会发现：*EMNLP 2020*，第1924–1938页，线上。计算语言学协会。

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。2021年。从自然语言监督中学习可迁移视觉模型。*arXiv:2103.00020 [cs]*。

AdityaRamesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022年。使用 CLIP 潜变量进行分层文本条件图像生成。*arXiv:2204.06125 [cs]*。

Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、ChelseaVoss、Alec Radford、Mark Chen 和

Ilya Sutskever。2021年。零样本文本到图像生成。*arXiv [*计算机科学*.CV]*。

Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bjorn Ommer。2022年。使用潜在扩散模型进行高分辨率图像合成。载于*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。IEEE。

Nataniel Ruiz、Yuanzhen Li、Varun Jampani、Yael Pritch、Michael Rubinstein 和 Kfir Aberman。2023年。DreamBooth：针对主体驱动生成微调文本到图像扩散模型。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。IEEE。

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。2022年。具备深度语言理解能力的逼真文生图扩散模型。*arXiv [*计算机科学*.CV]*。

Christoph Schuhmann、Richard Vencu、Romain Beaumont、RobertKaczmarczyk、Clayton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev 和 Aran Komatsuzaki。2021年。LAION-400M：包含 4 亿组经 CLIP 筛选的图文对的开放数据集。*arXiv[*计算机科学*.*计算机视觉*]*。

Piyush Sharma、Nan Ding、Sebastian Goodman 和 Radu Soricut。2018年。Conceptual captions：一个经过清理、上位词化的图像替代文本数据集，用于自动图像描述。收录于计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），第2556–2565页，澳大利亚墨尔本。计算语言学协会。

Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan 和 Surya Ganguli。2015年。使用非平衡热力学进行深度无监督学习。收录于第*32*届国际机器学习会议论文集，第2256–2265页。PMLR。

Alasdair Tran、Alexander Mathews 和 Lexing Xie。2020年。Transform and Tell：实体感知新闻图像描述。在*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 13035–13045 页。openaccess.thecvf.com。

Apostolia Tsirikoglou、Gabriel Eilertsen 和 Jonas Unger。2020年。视觉机器学习图像合成方法综述。计算机图形论坛，39(6):426–451。

Emiel van Miltenburg、Roser Morante 和 Desmond Elliott。2016年。图像描述中的语用因素：否定的情形。载于第五届视觉与语言研讨会论文集，第54–59页，德国柏林。计算语言学协会。

Ashish Vaswani、Noam M Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是你所需要的一切。神经信息处理系统，30:5998–6008。Ramakrishna Vedantam、Samy Bengio、Kevin Murphy、Devi Parikh 和 Gal Chechik。2017年。基于上下文无关监督的上下文感知图像描述。载于*2017*年 *IEEE*计算机视觉与模式识别会议 *(CVPR)*，第1070–1079页。IEEE。Chien-Yao Wang、I-Hau Yeh 和 Hong-Yuan Mark Liao。2021年。你只需学习一种表征：用于多项任务的统一网络。*arXiv [*计算机视觉*]*。

Zijie J Wang、Evan Montoya、David Munechika、HaoyangYang、BenjaminHoover 和 Duen Horng Chau。2022年。DiffusionDB：面向文本到图像生成模型的大规模提示词图库数据集。*arXiv [*计算机视觉*]*。

Xiaoshi Wu、Yiming Hao、Keqiang Sun、Yixiong Chen、Feng Zhu、Rui Zhao 和 Hongsheng Li。2023年。人类偏好评分 v2：用于评估文本到图像合成的人类偏好的可靠基准。*arXiv [*计算机视觉*]*。

Jiazheng Xu、Xiao Liu、Yuchen Wu、Yuxuan Tong、Qinkai Li、Ming Ding、Jie Tang 和 Yuxiao Dong。2023年。ImageReward：学习和评估人类对文本生成图像的偏好。*arXiv[*计算机科学*.CV]*。

Tao Xu、Pengchuan Zhang、Qiuyuan Huang、Han Zhang、Zhe Gan、Xiaolei Huang 和 Xiaodong He。2018年。AttnGAN：利用注意力生成对抗网络进行细粒度文本生成图像。载于*2018*年*IEEE/CVF* 计算机视觉与模式识别会议，第 1316–1324 页，美国犹他州盐湖城。IEEE。

Rowan Zellers、Ari Holtzman、Hannah Rashkin、Yonatan Bisk、Ali Farhadi、Franziska Roesner 和 Yejin Choi。2019年。防范神经网络生成的假新闻。*Adv. Neural Inf.Process. Syst.*，32。

Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和 Oriol Vinyals。2017年。理解深度学习需要重新思考泛化。*arXiv:1611.03530 [*计算机科学*]*。

Tao Zhang。2022年。深度伪造生成与检测：综述。多媒体工具与应用，81(5)：6259–6276。

Yufan Zhou、Ruiyi Zhang、Changyou Chen、Chunyuan Li、Chris Tensmeyer、Tong Yu、Jiuxiang Gu、Jinhui Xu 和 Tong Sun。2022年。迈向无需语言的文本生成图像训练。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第17907–17917页。openaccess.thecvf.com。

Minfeng Zhu、Pingbo Pan、Wei Chen 和 Yi Yang。2019年。DM-GAN：用于文本到图像合成的动态记忆生成对抗网络。载于*2019*年 *IEEE/CVF*计算机视觉会议

与模式识别（*CVPR*），第5795–5803页，美国加利福尼亚州长滩。IEEE。