使用扩散模型进行图像描述的多模态数据增强

Changrong Xiao[1], **SeanXin Xu**[1], **KunpengZhang**[21]清华大学经济管理学院[2]马里兰大学决策、运营与信息技术系 xcr21@mails.tsinghua.edu.cn xuxin@sem.tsinghua.edu.cn kpzhang@umd.edu

arXiv:2305.01855v1 [cs.CV] 3 May 2023

摘要

图像描述是一项重要的视觉-语言任务，通常需要大量精细标注的图像-描述对，以学习图像与文本之间的内在对齐关系。本文提出了一种多模态数据增强方法，利用一种名为*Stable Diffusion*的最新文本到图像模型，通过高质量生成图像-描述对来扩充训练集。在 MS COCO 数据集上进行的大量实验表明，与多种基准方法相比，我们的方法具有优势；尤其是在训练样本较少时，性能提升显著。此外，使用我们构建的增强数据集训练的模型，也大幅优于以往的无配对图像描述方法。最后，根据质量评估对生成数据进行有针对性的筛选后，训练效率和效果还能进一步提升。

1 引言

图像描述旨在自动生成图像视觉内容的文本描述，是自然语言处理（NLP）与计算机视觉（CV）交叉领域的一项重要任务（Stani¯ut˙eand Šešok，2019年；Atliha andŠešok，2020年；Cornia et al.，2020年；Turkerud andMeng-shoel，2021年）。尽管随着深度学习算法的兴起，尤其是在该领域取得了令人瞩目的成果，但大多数研究仍然聚焦于模型优化、提取信息丰富的特征或开发更好的训练技术。数据是影响模型性能的关键维度，尽管近期受到的关注有所增加（Feng etal.，2021a；Turkerud andMengshoel，2021年）。

在监督式图像描述任务中，通常需要大量带有精细标注的图像文本对（Hendricks 等人，2016年；Zhu 等人，2022年）。现有的图像描述数据集，例如 MS COCO（Lin 等人，2014年）要求人工标注人员用描述性句子为图像添加标签，这一过程既繁琐又耗时。此外，收集的图像和标注的图像描述可能不完整且缺乏多样性，从而限制了在此类数据集上训练的模型的泛化能力（Zhu 等人，2022年）。

为了解决数据问题，一些研究人员尝试利用无配对图像和文本，因为它们很容易分别获取（Hos-sainet al.，2019年；Kim et al.，2019年；Laina et al.，2019年；Zhu et al.，2022年）。这项任务称为无配对图像描述（UIC），近期受到了广泛关注。还有一些研究人员利用数据增强来扩充训练数据并提升模型性能，包括对图像（Wang et al.，2018年；Katiyar and Borgohain，2021年）和文本描述（AtlihaandŠešok，2020年；Turkerud and Mengshoel，2021年；Atlihaand Šešok，2022年）。

在这项工作中，我们尝试一种更具挑战性的情形：仅提供文本数据，不提供图像。我们应用 *Stable Diffusion* (Rombach 等人，2022年），这是一种近期发展成熟的文本到图像模型，可根据给定的文本输入生成高质量图像。我们进一步创建了多个合成数据集，通过基于文本和/或图像的数据增强来扩充图像-描述对。在这些合成数据集上训练后，图像描述模型的性能可显著超越先进的 UIC 方法。我们还通过大量实验表明，在我们的多模态增强数据集上训练的模型优于以往的图像描述数据增强方法，尤其是在真实标注数据有限时。为进一步提升性能，我们依据质量度量对增强数据进行评估和筛选。总之，本文的贡献如下：

- 我们开发并实现了一种多模态数据增强方法，以提升图像描述模型的性能。据我们

所知，这是首次尝试在图像描述任务中同时对图像和文本进行增强。

- 我们进行了大量实验，结果表明，在零样本和少样本场景中，由大型预训练文本到图像模型生成的高质量合成数据能够显著提升图像描述的质量。我们的研究也成功应用了*Stable Diffusion* 模型。

- 最后，我们计划向社区发布代码和增强数据集[1]。凭借有效的图像-描述对生成与质量评估方法，我们证明，无需投入大量成本进行人工标注，也能为有监督的视觉语言任务构建训练数据集。

2 相关工作

2.1 图像描述中的数据问题

近年来，受益于深度学习算法，图像描述模型发展迅速。编码器-解码器架构是最常见且有效的模型框架之一，它使用卷积神经网络（CNN）编码器提取图像特征，并使用循环神经网络（RNN）将其解码为自然语言（Hossain 等人，2019年）。注意力机制（Xu 等人，2015年；Lu 等人，2017年；Anderson 等人，2018年；Huang 等人，2019年）和 Transformer（Li 等人，2019年；Cornia 等人，2020年）已得到广泛应用，并显著提升了性能。

采用全监督方法训练图像描述模型，需要大量配对的图像-描述数据。缺少训练数据时，即使是最先进的模型也很难取得良好表现。为解决标注数据有限这一问题，研究人员提出了无配对图像描述（UIC）和数据增强方法。

无配对图像描述UIC 任务旨在利用分别从各种来源（例如网络）轻松获取的无配对图像和图像描述训练模型，并生成图像描述。由于获取配对图像和文本的成本高昂，这一任务受到了研究人员的广泛关注。例如，Gu 等人（2018年）利用额外的中文实现了语言枢轴描述信息。Feng 等人（2019年）提出了一种无配对图像描述框架，通过学习视觉概念检测器来实现这一目标。Laina 等人（2019年）利用数据集之外的大规模文本语料库，学习了图像与句子的共享多模态嵌入。最近，Zhu 等人（2022年）仅借助图像级类别标签，解决了 UIC 的视觉概念识别阶段问题。另一方面，半监督学习方法也发挥了作用。Chen 等人（2016年）基于文本数据生成缺失的视觉信息。Kim 等人（2019年）利用生成对抗网络为未标注图像分配伪标签。大多数 UIC 研究的共同点是，它们会利用图像描述数据之外的数据，并且或多或少需要一些辅助信息或昂贵的标注。

数据增强 旨在创建更多样化的训练数据。它们已广泛应用于各种机器学习任务，分别对图像和文本进行增强。然而，增强技术很少用于视觉-语言任务，例如图像描述（Atliha 和 Šešok，2020年）。

图像描述研究中的大多数数据增强工作只关注图像或描述其中一方，而非同时增强两者。Wang 等人 (2018 年）；Katiyar 和 Borgohain (2021年) 使用裁剪、翻转和镜像等标准图像变换来处理图像数据。然而，当变换扭曲图像语义时，这些图像增强方法可能会引入噪声。对于描述增强，Atliha 和 Šešok(2020年) 使用 BERT（Devlin 等人，2018年）进行同义词替换和句子释义改写。词语排列/替换（Cui 等人，2018年）、回译（Turkerud 和 Mengshoel，2021年）以及其他自然语言处理增强技术也曾用于以往研究。

然而，同时修改图像和文本的多模态增强仍是一个相对少有探索的方向（Hartmann 等人，2022年）。Feng 等人（2021b）提出了一种结合 CutMix（Yun 等人，2019年）和描述编辑的方法：从另一张图像中剪切出图块并插入其中，同时修改描述，使其能够准确描述新图像。但作者并未实现这一方法，也没有提供实验结果。

2.2 文本到图像合成

文本到图像合成是一项具有挑战性的多模态任务，旨在生成以描述性文本为条件的高质量图像

[1] https://github.com/Xiaochr/Multimodal-Augmentation-Image-Captioning

。Du 等人，2022年；Yang 等人，2022年）。该领域此前一直由生成对抗网络（生成对抗网络）（Goodfellow 等人，2014年）、变分自编码器（变分自编码器）（Kingma 和 Welling，2013年）以及其他生成模型主导。然而，这些模型存在一些缺陷。例如，生成对抗网络难以优化（Mescheder 等人，2018年），且只能处理变化有限的数据（Brock 等人，2018年；Karras 等人，2019年）；变分自编码器生成高分辨率图像的效率更高，但生成样本的质量仍不够理想（Rombach 等人，2022年）。

近年来兴起的一类深度生成模型——扩散模型（Ho 等人，2020年；Song 等人，2021年），展现出强大的能力，并在文本到图像合成方面超越了生成对抗网络（Dhariwal 和 Nichol，2021年）。尽管 *DALLE-2*（Ramesh 等人，2022年）和*Imagen*（Saharia 等人，2022年）等最先进的模型能够生成质量出奇高的图像，但其推理过程成本高昂且耗时（Rombach 等人，2022年）。在本研究中，我们采用改进后的潜在扩散模型 *Stable Diffusion*（Rombach 等人，2022年）作为文本到图像组件。*StableDiffusion* 能够以适中的计算资源和耗时，生成可与最先进模型媲美的高分辨率图像。

在图像描述任务中，生成模型很少得到应用。Kim 等人 (2019 年) 使用 *Cycle-GAN* (Zhu 等人，2017年) 作为其无配对图像描述方法的基线，但结果证明其效果较差。在最新的研究中，Li 等人(2022年) 利用 *Stable Diffusion* 作为文本到图像模型，并使用*Flamingo* (Alayrac 等人，2022年) 作为其逆向模型，证明了能够实现原始图像最佳重建效果的描述就是该图像的最佳描述。然而，尽管扩散模型在文本到图像生成方面取得了令人瞩目的成果，但它们尚未被用于提高生成图像描述的质量。

3 方法

我们提出的图像描述系统由两部分组成。第一部分，我们实现图像-描述对生成，以构建合成数据集。还可以通过文本或图像增强方法进一步扩充该数据集。第二部分介绍如何基于第一部分构建的数据，对选定的两个图像描述模型进行训练和评估。

3.1 多模态增强

我们使用 MS COCO（Lin 等人，2014年）作为基础数据集进行数据增强。MS COCO 是一个规模庞大且广泛使用的图像描述数据集，共包含 123,287 张图像和 616,767 条图像描述。

在我们的多模态增强中，我们首先应用 *Stable Diffusion* 为每条 COCO 描述生成一张图像，同时丢弃含有不适宜工作场所观看的内容的图像。随后，我们将生成的图像与 COCO 真实描述配对，构成基础合成数据集（记为 SD_{base}）。

先前研究发现，使用描述更多样化的图像训练图像描述模型，可以取得更好的性能（De-vlinet al.，2018年；AtlihaandŠešok，2020年）。因此，对于 SD_{base} 中的每张生成图像，仅使用一条描述可能不够，扩充每张生成图像的描述是可取的。本研究采用两种策略扩充图像描述：（i）基于 COCO 真实描述；（ii）通过文本增强（例如释义改写）自动生成描述。由此得到的数据集分别记为 SD_{true} 和SD_{para}。图1 以 COCO 数据中的一个示例说明了构建三个增强数据集的方式（即，一个示例包含一张图像及其对应的 5 条描述）。

- SD_{base} 由 5 个图像-描述对组成，每条真实描述对应一张生成图像：(合成图像 1，真实描述 1)、 ···、(合成图像 5，真实描述 5)。

- SD_{true} 基于所有 5 条描述均相关这一假设进行扩展 SD_{base}。因此，根据一条描述生成的每张图像都会与其余 4 条描述及其自身建立映射关系：（来自真实描述 1 的合成图像 1，真实描述 1），

  ···,（来自真实描述 1 和真实描述 5 的合成图像 1），···,（来自真实描述 2 和真实描述 1 的合成图像 2），···,（来自真实描述 2 和真实描述 5 的合成图像 2），
  ···，（来自真实描述 5 的合成图像 5，真实描述 1），···，（来自真实描述 5 的合成图像 5，真实描述 5）。因此，SD_{true}仅通过一个示例就扩展为 25 个图像-描述对。

- SD_{para} 基于 SD_{base}构建。每条真实描述都通过 k 次释义改写进行增强。因此，SD_{para} 包含：（合成图像1，k 条合成图像描述）、 ···、（合成图像 5，k 条合成图像描述）。请注意，我们丢弃了

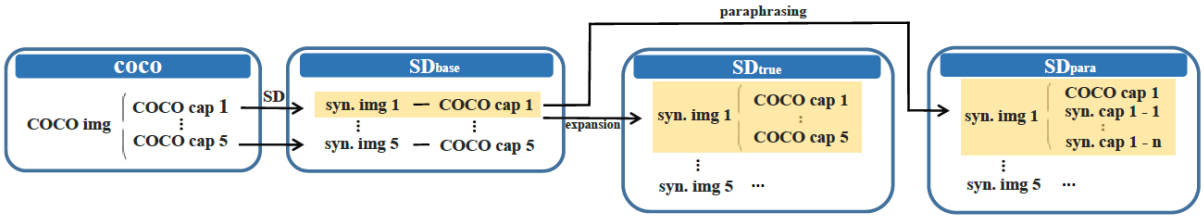


图 1：一个示例，展示如何使用文本到图像生成和描述扩展构建三个合成数据集（SD_{base}、SD_{true} 和 SD_{para}）。*COCO* 图像/描述 表示原始 COCO 数据集中的图像/图像描述，合成图像/描述表示合成图像/图像描述。

那些生成的释义完全相同。因此，SD_{para} 中每张生成图像对应的图像描述数量为 $k+1$（即一个真实图像描述和 $k \leq 5$ 个增强图像描述）。

接下来，我们创建了若干额外的数据集，同时考虑标注数据（即图像-描述对）可用量有所不同的情况。具体而言，一种典型情况是：仅有一小部分 COCO 图像-描述对可用，但所有图像描述均可用。在这种情况下，我们提出的多模态增强方法可用于增加图像-描述对。例如，我们首先保留这些已提供的图像-描述对。对于其余图像描述，我们按照 SD_{base} 中的方法进行增强，以创建更多图像-描述对。此类新数据集记为n% $COCO + SD_{base}$，其中 n 是 COCO 原始图像-描述对所占的百分比。例如，10% $COCO + SD_{base}$ 表示 COCO 中 10% 的真实图像-描述对，加上由其余 90% 图像描述生成的 SD_{base}。同理，可以得到 n% $COCO + SD_{true}$ 和n% $COCO + SD_{para}$。

与我们的多模态增强方法不同，现有许多单模态方法。例如，释义改写是一种广泛采用的文本增强方法，通过类似于Atliha 和 Šešok （2020年）的大型语言模型实现。对于图像增强，已有研究采用随机翻转与随机透视变换相结合的方法（Wang 等人，2018年；Katiyar 和 Borgohain，2021年）。基于这些方法，我们进一步创建了两个数据集。生成的释义改写文本被添加到 COCO 数据集中，作为额外的图像描述，记为 $COCO_{text}$。我们用增强后的图像替换原始 COCO 图像，从而得到 $COCO_{image}$。

3.2 选定的图像描述模型

接下来，我们将介绍两种选定的图像描述模型，并在其上评估增强数据集的有效性。我们的目标是探究实验结果是否因模型而异，以及我们构建的数据集能否普遍用于改进图像描述模型。为此，我们选用相对简单且规模较小的 FC模型作为图像描述模型。我们还选择了另一种更先进的基于 Transformer 的模型，以检验结果的稳健性。

FC模型 是许多图像描述研究中经常使用的模型（Vinyals 等人，2015年；Karpathy 和 Fei-Fei，2015年；Rennie 等人，2017年；Luo 等人，2018年）。在 FC模型中，首先嵌入 CNN 架构，以提取每幅图像的视觉特征。然后，通过长短期记忆（LSTM）模块（Hochreiter 和 Schmidhuber，1997 年）处理提取的特征嵌入，以生成图像描述。

基于 Transformer 的模型 近年来，Transformer（Vaswani 等人，2017年）模型提升了各种深度学习任务的性能，包括图像描述（Cornia 等人，2020年；Luo 等人，2021年；Zhou 等人，2022年）。本研究没有选择特定的基于 Transformer 的图像描述模型，而是采用了一种通用架构：首先通过自底向上的方法提取视觉特征（Anderson 等人，2018年），然后应用带有自注意力机制的基础 Transformer 模块，将视觉特征解码为文本描述。

4 实验与结果

4.1 实验设置

在本研究中，我们使用*Stable Diffusion* 模型的 1-4[2] 版本，该版本在*huggingface* 上实现，用于根据给定的图像描述生成合成图像。我们保留默认设置。在生成过程中，模型参数保持冻结状态。系统会自动检测 NSFW 图像，而我们对图像进行

[2]https://huggingface.co/CompVis/stable-diffusion-v1-4

再次生成这些图像，直至其符合纳入合成数据集的要求。最终，共生成了 566,747 张合成图像。

用于 SD_{para} 的合成描述扩展和用于 $COCO_{text}$ 的文本增强采用相同的释义改写方法，该方法基于真实的 COCO 描述：使用预训练的 T5 模型（Raffel 等人，2020年）进行释义改写[3]。对于图像增强，我们应用 $RandomHorizontalFlip$ 和 $RandomPerspective$ 由*torchvision*实现的变换函数，并将两种变换的概率均设为0.5。

特征提取、模型训练和模型评估的实现主要改编自 Luo 等人（2018年）[4]。我们采用全监督方法和更有效的 CIDEr 分数优化方法（Luo 等人，2018年），基于训练集的 Karpathy 划分（Karpathy 和 Fei-Fei，2015年）训练图像描述模型。此外，我们还采用了提前停止策略，FC模型和基于 Transformer 的模型的最大训练轮数分别设置为 30 和 15。对于模型评估，我们使用图像描述任务的标准评估指标，包括 BLEU（Papineni 等人，2002年）、METEOR（Banerjee 和 Lavie，2005年）、ROUGE（Lin，2004年）、CIDEr（Vedantam 等人，2015年）和 SPICE（Anderson 等人，2016年）。所有评估指标均与 Luo 等人（2018年）保持一致，COCO 数据集和合成数据集的所有评估均基于 Karpathy 测试集划分。我们将随机种子设置为 42，并使用 2 块 RTX A4000 16G GPU 运行所有实验。接下来，我们将展示三组实验及其结果，以证明我们提出的多模态数据增强方法的有效性，并在此基础上开展图像描述任务及其评估。

4.2 实验 1：完全合成数据集

在本实验中，我们使用 3 个数据集训练模型，这些数据集中的所有图像均为合成图像：SD_{base}、SD_{para}和 SD_{true}。我们将图像描述模型（即 FC 和 Transformer）的性能与先进的 UIC 方法进行比较。需要注意的是，UIC 任务假设存在未配对但真实的图像，这比我们的场景更加有利，因为在我们的场景中不存在任何真实图像。实验结果如表1所示。由于两个模型的结果相当一致，这里展示 FC模型的结果，并且请参见附录中基于 Transformer 的模型部分。

训练集	B4	M	R	C	S
完全配对的 COCO	28.7	24.4	52.4	92.4	17.6
Gu 等人(2018年)	5.4	13.2	-	17.7	-
Feng 等人 (2019年)	18.6	17.9	43.1	54.9	11.1
Laina 等人(2019年)	19.3	20.2	45.0	61.8	12.9
Zhu 等人 (2022年)	21.5	20.1	45.8	65.7	13.6
我们的 SD_{base}	23.6	21.7	48.6	75.4	15.3
我们的 SD_{para}	23.4	21.6	48.9	75.5	15.4
我们的 SD_{true}	**25.6**	**22.6**	**50.1**	**81.0**	**15.6**

表 1：我们的多模态增强方法与 UIC 的性能比较。第一行展示了在原始 COCO 数据集上训练的 FC模型的性能，作为比较基准；中间几行列出了先前研究报告的 4 种常用 UIC 基线的性能；底部几行展示了在我们的 3 个合成数据集上训练的 FC模型的性能。请注意，“B4”“M”“R”“C”和“S”分别代表“BLEU-4”“METEOR”“ROUGE”“CIDEr”和“SPICE”。

我们得到以下观察结果：（1）使用我们的基础多模态增强数据集（SD_{base}）进行训练，在所有指标上都大幅优于 UIC 基线。换言之，预训练的文本到图像模型可以根据给定的描述生成图像，用于训练图像描述模型。图像描述模型可以通过完全监督方式，在合成图像-描述对上进行训练，而不是在 UIC 任务中基于真实但未配对的图像-描述数据，以半监督或无监督方式进行训练。即使没有真实图像，合成数据在下游图像描述任务中也表现出令人惊讶的效果，显示出合成数据生成的巨大应用潜力。（2）使用合成释义扩展图像描述（SD_{para}）可以取得更好的结果。在 SD_{true}上进行训练后，性能进一步得到显著提升。这可能是因为通过释义改写获得的文本，其多样性低于人工标注的真实图像描述。这也与先前的研究结果一致（Atliha 和 Šešok，2020年；Turkerud 和 Mengshoel，2021年），研究人员发现，如果训练集中的图像拥有更多且多样性更高的图像描述，模型性能会得到提升。

4.3 实验 2：少样本学习

接下来，我们打算了解在以下情况下多模态数据增强方法的性能：

[3]https://huggingface.co/Vamsi/T5_Paraphrase_Paws[4]https://github.com/ruotianluo/ImageCaptioning.pytorch

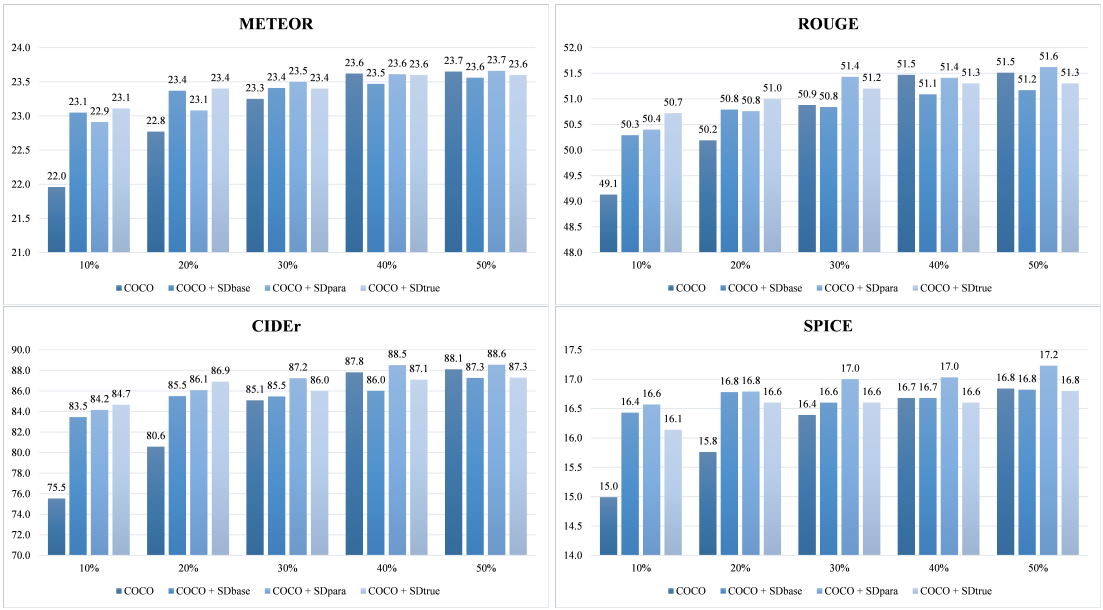


图2：使用不同 COCO 数据比例的三个多模态增强数据集的图像描述性能。五组分别为包含 10% 至 50% COCO 数据的情形。在每组中，从左到右仅展示 $n\%$COCO 、 $n\%$ COCO 与 SD_{base}， $n\%$ COCO 与 SD_{para}，以及 $n\%$ COCO 与 SD_{true}。图中仅展示了部分评估指标。更多详情请参见附录中的表 10 。

真实标注数据有限。为此，我们首先按不同的比例对原始 COCO 数据进行抽样，比例范围为 10% 到 50%，然后对其应用文本、图像和多模态数据增强方法。此外，我们假设 COCO 中的所有图像描述均可用，因此多模态数据增强的目的是补全缺失的图像。这一假设是合理的，因为仅获取文本信息相对容易。

训练集	B4	M	R	C	S
10% COCO	24.3	22.0	49.1	75.5	15.0
10% $COCO_{text}$	23.8	21.6	49.1	74.6	15.0
10% $COCO_{img}$	24.1	22.0	49.0	75.7	15.0
10% COCO + SD_{base}	25.8	23.1	50.3	83.5	16.4
10% COCO + SD_{para}	26.0	22.9	50.4	84.2	**16.6**
10% COCO + SD_{true}	**26.2**	**23.1**	**50.7**	**84.7**	16.1

表2：COCO 中有 10% 的图像对可用时 FC模型的性能。

从结果（见桌子2）中我们发现，当训练实例的数量非常有限（10%）时，针对单一模态（图像或文本）的基础数据增强几乎不会影响图像描述的性能（即比较第 2、3 行与第 1 行）。相比之下，我们的多模态数据增强可以显著提升性能（即比较最后 3 行与第 1 行）。此外，合成图像的描述扩展可以进一步提升性能（即比较第 5 行：10% $+ SD_{para}$ 与第 4 行：10% $+ SD_{base}$），这与实验 1 中的发现一致。

所有实验结果从 10% 到 50% COCO 数据汇总于图2中。随着真实标注数据规模增大，数据增强带来的性能提升逐渐减小。当真实数据达到40%时，使用 SD_{base} 增强的效果变得比仅使用40% 真实数据更差。一种可能的解释是，COCO 中的40% 真实数据已足以训练出性能不错的模型。添加合成图像可能会引入更多噪声，从而损害性能。

在三种多模态增强方法中，当真实数据非常有限（10%）时，合成图像的真实描述扩展效果优于自动改写扩展。然而，当真实数据量增加至 $\geq$ 40% 时，真实描述扩展无法达到 n % COCO 基线的同等效果。尽管如此，在这种情况下，自动改写仍然能够提升模型性能。这可能是因为改写模型引入了 COCO 词汇表之外的新颖词汇。

4.4 实验 3：合成数据筛选i ng

在本节中，我们尝试利用增强数据集进一步提升图像描述任务的性能。

我们注意到，*Stable Dif-fusion* 生成的图像可能与文本描述不匹配，例如遗漏重要物体或出现不自然的失真。我们非常希望筛选出与文本高度契合的高质量图像。以往的研究提出了多种方法来实现这一目标。例如，（Salimansetal.，2016年）提出了 Incep- tion Score（IS），用于评估 GAN 生成的图像。（Heuselet al.，2017年）提出了弗雷歇初始距离（FID），将其作为衡量图像质量的黄金标准。这些指标可能无法直接应用于我们的场景，部分原因在于，它们衡量的是生成图像相对于真实图像的失真程度，或两个分布之间的差异，而我们的目标是找出与文本更匹配的图像。

此外，评估不应依赖真实的参考图像和文本。因此，在我们的语境中评估图像质量时，我们考虑以下三个方面：

- 图像本身的质量，由近期提出的一种无参考图像质量评估（NR-IQA）指标衡量，该指标名为*MUSIQ* (Ke 等人，2021年)。

- 合成图像与相应输入文本之间的相似度。*CLIPScore*(Hessel 等人 , 2021 年）是一种基于 CLIP嵌入、用于衡量图像描述性能的无参考指标，可用于衡量生成的图像-描述对的一致性。

- 合成图像是否反映了描述中提及的重要物体。*VIFI-DEL* (Madhyastha 等人，2019年）是一种新提出的图像描述任务质量评估方法，它使用物体检测模型识别图像中的主要物体，然后计算物体与描述文本之间的相似度。

我们使用 *MUSIQ*、*CLIPScore* 和 *VIFIDEL* 作为数据筛选标准，并发现 *CLIPScore* 是三者中最有效的指标。选取*CLIPScore*最高的前 50% 数据，可以取得与使用完整合成实验 1 中的合成数据集（见表 3）。这表明，没有必要使用完整的合成数据训练模型。基于质量评估进行数据选择，可以在不牺牲性能的情况下提高训练效率。

训练集	B4	M	R	C	S
COCO	28.7	24.4	52.4	92.4	17.6
SD_{base}	**23.6**	21.7	48.6	**75.4**	15.3
已选择 SD_{base}	23.5	**22.2**	**48.9**	75.3	**15.6**
SD_{para}	**23.4**	21.6	**48.9**	**75.5**	15.4
已选择 SD_{para}	23.2	**22.0**	48.7	75.0	**15.6**
SD_{true}	**25.6**	22.6	50.1	81.0	15.6
已选择 SD_{true}	25.5	**22.7**	**50.2**	**81.0**	**15.7**

表3：在筛选出 CLIPScore*CLIPScore*排名前50%的数据后，使用三个增强数据集训练的 FC 模型的性能。

由于在三种增强数据集中，数据选择对 SD_{true}的效果更好，我们进一步利用 SD_{true} 探究数据量有限时数据筛选的影响。我们在桌子 4 中发现，数据筛选确实能提升模型性能。即使真实数据量增加，性能仍有显著提升，超过了 SD_{para}带来的改进。这表明，当真实标注数据有限时，适当的数据筛选既能提高训练效率，也能提升图像描述性能。

训练集	B4	M	R	C	S
10% COCO + SD_{para}	26.0	22.9	50.4	84.2	**16.6**
10% COCO + SD_{true}	26.2	23.1	50.7	84.7	16.1
10% COCO + 已选择 SD_{true}	**26.8**	**23.3**	**51.1**	**86.2**	16.5
20% COCO + SD_{para}	26.3	23.1	50.8	86.1	**16.8**
20% COCO + SD_{true}	26.9	23.4	51.0	86.9	16.6
20% COCO + 已选中 SD_{true}	**27.1**	**23.5**	**51.2**	**87.2**	16.7
30% COCO + SD_{para}	27.0	23.5	51.4	87.2	**17.0**
30% COCO + SD_{true}	26.9	23.4	51.2	86.0	16.6
30% COCO + 已选中 SD_{true}	**27.0**	**23.7**	**51.4**	**87.8**	16.9
40% COCO + SD_{para}	27.3	23.6	51.4	**88.5**	**17.0**
40% COCO + SD_{true}	27.1	23.6	51.3	87.1	16.6
40% COCO + 已选中 SD_{true}	**27.3**	**23.7**	**51.4**	88.3	17.0
50% COCO + SD_{para}	27.5	23.7	**51.6**	88.6	**17.2**
50% COCO + SD_{true}	27.1	23.6	51.3	87.3	16.8
50% COCO + 已选中 SD_{true}	**27.6**	**23.7**	51.5	**89.3**	16.9

桌子4：采用基于前 50%*CLIPScore* 标准的合成数据筛选时，FC 模型在有限数据设置下的性能。

然而，SD_{base} 和SD_{para} 带来的提升并不显著。而且，在某些情况下，基于 Transformer 的模型的图像描述性能甚至会下降（见附录中的桌子11）。因此，*CLIPScore*和另外两项指标并非筛选适用于图像描述任务的高质量数据的绝对标准，尽管已有研究证明它们在图像描述质量评估方面表现良好（Madhyastha 等人，2019年；Hessel 等人，2021年）。

以往的一项图像分类任务研究也报告了生成数据质量与下游任务表现之间存在类似的差异（Ravuri 和 Vinyals，2019年）。作者发现，尽管 GAN 生成的图像获得了接近真实图像的高分，但使用完全由合成图像构成的数据训练的分类模型，其准确率远低于使用真实图像训练的模型。沿着这一思路，我们分别计算了 COCO 数据和合成数据的三项指标。我们发现，合成数据的三项质量度量与真实数据的相应指标水平相近（见表 5）。然而，若在训练时完全以合成数据替代真实数据（例如实验 1），模型表现会远逊于使用真实数据训练的模型（即 CIDEr 分数：81。0 对比 92。4）。由此，我们将 Ravuri 和 Vinyals (2019年) 的研究发现拓展到了图像描述任务。

	MUSIQ	CLIPScore	VIFIDEL
COCO 数据	69.8	78.4	35.8
合成数据	69.6	80.1	34.4

桌子5：使用真实与增强图像-描述对对多模态数据质量进行比较。

5 结论

我们针对图像描述任务开发了一种多模态数据增强方法，利用扩散模型在图像合成方面的强大能力。在 MS COCO 数据集上，对于两种典型的图像描述模型，该方法优于单模态方法，尤其是在真实标注数据量有限的情况下。与使用完全合成图像的 UIC 方法相比，该方法的表现也显著更好。

我们的研究是一次早期尝试，通过两个逆向过程——文本到图像和图像到文本——将图像模态与文本模态结合起来。我们通过实证验证了将合成多模态数据用作训练集的有效性，并且通过数据过滤还能进一步提升性能。尽管合成数据无法取代真实数据，但未来仍值得探索多模态数据合成的潜力，以及其在各种下游应用中的用途。

局限性

下面我们讨论多模态数据增强方法的三个局限性。（1）最值得注意的是合成数据的质量，这是许多数据生成研究中常见的挑战。我们探索了一些用于图像-描述对的质量评估指标，并测试了它们在图像描述任务中的表现。有效且具有普适性的多模态数据质量评估仍然是一个有待未来研究解决的开放问题，我们认为这是一个有价值的研究方向。（2）合成图像的质量和灵活性受限于所应用的文本到图像模型的能力。例如，如果我们打算为人脸识别任务生成合成训练数据集，由于 *StableDiffusion* 不擅长绘制人脸，因此不太可能取得良好表现。（3）即使使用 *StableDiffusion* 这类轻量级模型，生成大型合成图像数据集仍需要大量计算资源。在本研究中，生成完整的合成 COCO 图像耗时约 2 周，并使用了 2 个 RTX A4000 16G GPU，这对从业者和研究人员而言都是一项较大的负担。

伦理声明

本研究使用了在海量公开可用的在线图像上预训练的 *Stable Diffusion*。然而，有人指出，其中包含了若干未经授权收录的在线艺术家创作的画作。这是否损害艺术家的原创性，或涉及隐私问题，仍是激烈争论的话题。这些大型预训练文本到图像模型的开发者已采取措施消除伦理方面的顾虑，例如从训练集中移除未经授权的图像。

这对我们而言可能是一个潜在问题，但我们关注的是这些文本到图像模型的能力。我们生成更高质量图像-描述对的方法可以应用于更广泛的领域（例如，在教育领域为视障人士提供辅助），从而对社会和人类福祉产生更大的影响。

参考文献

Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katie Millican、Malcolm Reynolds 等。2022年。Flamingo：一种用于少样本学习的视觉语言模型。*arXiv* 预印本 *arXiv:2204.14198*。

Peter Anderson、Basura Fernando、Mark Johnson 和 Stephen Gould。2016年。SPICE：用于图像描述评估的语义命题。载于欧洲计算机视觉会议，第 382–398 页。Springer。

Peter Anderson、Xiaodong He、Chris Buehler、Damien Teney、Mark Johnson、Stephen Gould 和 Lei Zhang。2018年。用于图像描述和视觉问答的自底向上和自顶向下注意力。载于*IEEE*计算机视觉与模式识别会议论文集，第 6077–6086 页。

Viktar Atliha 和 Dmitrij Šešok。2020年。使用 BERT 进行图像描述的文本增强。应用科学，10(17):5978。

Viktar Atliha 和 Dmitrij Šešok。2022年。用于压缩图像描述模型的文本增强。在*2022*年 *IEEE* 电气、电子与信息科学开放会议（*eStream*），第 1–4 页。IEEE。Satanjeev Banerjee 和 Alon Lavie。2005年。Meteor：一种自动机器翻译评估指标，与人工标注判断的相关性更高。在机器翻译和/或摘要的内在与外在评估指标 *ACL* 研讨会论文集，第 65–72 页。Andrew Brock、Jeff Donahue 和 Karen Simonyan。2018年。用于高保真自然图像合成的大规模生成对抗网络训练。*arXiv* 预印本 *arXiv:1809.11096*。Wenhu Chen、Aurelien Lucchi 和 Thomas Hofmann。2016年。用于图像描述的半监督框架。*arXiv* 预印本 *arXiv:1611.05321*。Marcella Cornia、Matteo Stefanini、Lorenzo Baraldi 和 Rita Cucchiara。2020年。用于图像描述的网格记忆 Transformer。在*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 10578–10587 页。Yin Cui、Guandao Yang、Andreas Veit、Xun Huang 和 Serge Belongie。2018年。学习评估图像描述。在*IEEE* 计算机视觉与模式识别会议论文集，第 5804–5812 页。Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：用于语言理解的深度双向 Transformer预训练。*arXiv* 预印本 *arXiv:1810.04805*。

Prafulla Dhariwal 和 Alexander Nichol。2021年。扩散模型在图像合成方面胜过生成对抗网络。神经信息处理系统进展，34:8780–8794。

Yifan Du、Zikang Liu、Junyi Li 和 Wayne Xin Zhao。2022年。视觉语言预训练模型综述。*arXiv* 预印本 *arXiv:2202.10936*。

Steven Y. Feng、Varun Gangal、Jason Wei、Sarath Chan- dar、Soroush Vosoughi、Teruko Mitamura 和 Ed- uard Hovy。2021a。自然语言处理数据增强方法综述。收录于计算语言学协会论文集：*ACL-IJCNLP 2021* 研究成果，第 968–988 页，线上。计算语言学协会。

Steven Y Feng、Varun Gangal、Jason Wei、Sarath Chandar、Soroush Vosoughi、TerukoMitamura 和 Eduard Hovy。2021年b。自然语言处理数据增强方法综述。*arXiv* 预印本 *arXiv:2105.03075*。

Yang Feng、Lin Ma、Wei Liu 和 Jiebo Luo。2019年。无监督图像描述。载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，第 4125–4134 页。

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. Jiuxiang Gu、Shafiq Joty、Jianfei Cai 和 Gang Wang。2018年。通过语言转换实现无配对图像描述。载于欧洲计算机视觉会议（*ECCV*）论文集，第 503–519 页。Mareike Hartmann、Aliki Anagnostopoulou 和 Daniel Sonntag。2022年。用于图像描述的交互式机器学习。*arXiv* 预印本 *arXiv:2202.13623*。Lisa Anne Hendricks、Subhashini Venugopalan、Marcus Rohrbach、Raymond Mooney、Kate Saenko 和 Trevor Darrell。2016年。深度组合式描述：无需配对训练数据即可描述新颖物体类别。载于*IEEE*计算机视觉与模式识别会议论文集，第 1–10 页。Jack Hessel、Ari Holtzman、Maxwell Forbes、Ronan Le Bras 和 Yejin Choi。2021年。CLIPScore：一种无需参考文本的图像描述评估指标。*arXiv* 预印本 *arXiv:2104.08718*。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。神经信息处理系统进展，30。Jonathan Ho、Ajay Jain 和 Pieter Abbeel。2020年。去噪扩散概率模型。*CoRR*，abs/2006.11239。

Sepp Hochreiter 和 Jürgen Schmidhuber。1997年。长短期记忆。*Neural Computation*，9(8):1735–1780。

MD Zakir Hossain、Ferdous Sohel、Mohd Fairuz Shiratuddin 和 Hamid Laga。2019年。深度学习在图像描述领域的综合综述。*ACMComputingSurveys(CsUR)*，51(6):1–36。

Lun Huang、Wenmin Wang、Jie Chen 和 Xiao-Yong Wei。2019年。用于图像描述的注意力上的注意力。载于*IEEE/CVF*国际计算机视觉会议论文集，第 4634–4643 页。

Andrej Karpathy 和李飞飞。2015年。用于生成图像描述的深度视觉-语义对齐。载于*IEEE*计算机视觉与模式识别会议论文集，第 3128–3137 页。

Tero Karras、Samuli Laine 和 Timo Aila。2019年。用于生成对抗网络的基于风格的生成器架构。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 4401–4410 页。

Sulabh Katiyar 和 Samir Kumar Borgohain。2021年。使用深度堆叠式长短期记忆网络、上下文词嵌入和数据增强进行图像描述。*arXiv* 预印本 *arXiv:2102.11237*。

Junjie Ke、Qifei Wang、Yilin Wang、Peyman Milan- far 和 Feng Yang。2021年。Musiq：多尺度图像质量变换器。载于*IEEE/CVF*国际计算机视觉会议论文集，第5148–5157页。Dong-Jin Kim、Jinsoo Choi、Tae-Hyun Oh 和 In So Kweon。2019年。使用极少量监督数据进行图像描述：对抗式半监督学习方法。*arXiv* 预印本 *arXiv:1909.02201*。Diederik P Kingma 和 Max Welling。2013年。自动编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*。Iro Laina、Christian Rupprecht 和 Nassir Navab。2019年。探索使用共享多模态嵌入进行无监督图像描述。载于*IEEE/CVF*国际计算机视觉会议论文集，第7414–7424页。Guang Li、Linchao Zhu、Ping Liu 和 Yi Yang。2019年。用于图像描述的纠缠变换器。载于*IEEE/CVF*国际计算机视觉会议论文集，第8928–8937页。Hang Li、Jindong Gu、Rajat Koner、Sahand Shar- ifzadeh 和 Volker Tresp。2022年。DALL-E 和 Flamingo 能否相互理解？*arXiv* 预印本 *arXiv:2212.12249*。Chin-Yew Lin。2004年。ROUGE：用于摘要自动评估的工具包。载于*Text Summarization Branches Out*，第74–81页。

Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。2014年。Microsoft coco：情境中的常见物体。载于欧洲计算机视觉会议，第740–755页。Springer。Jiasen Lu、Caiming Xiong、Devi Parikh 和 Richard Socher。2017年。知道何时观察：通过视觉哨兵实现图像描述的自适应注意力。载于*IEEE*计算机视觉与模式识别会议论文集，第375–383页。Ruotian Luo、Brian Price、Scott Cohen 和 Gregory Shakhnarovich。2018年。用于训练描述性图注的可区分性目标。载于*IEEE*计算机视觉与模式识别会议论文集，第6964–6974页。Yunpeng Luo、Jiayi Ji、Xiaoshuai Sun、Liujuan Cao、Yongjian Wu、Feiyue Huang、Chia-Wen Lin 和 Rongrong Ji。2021年。用于图像描述的双层协同 Transformer。载于*AAAI*人工智能会议论文集，第35卷，第2286–2293页。Pranava Madhyastha、Josiah Wang 和 Lucia Specia。2019年。VIFIDEL：评估图像描述的视觉保真度。*arXiv*预印本 *arXiv:1907.09340*。Lars Mescheder、Andreas Geiger 和 Sebastian Nowozin。2018年。生成对抗网络的哪些训练方法确实能够收敛？载于国际机器学习会议，第3481–3490页。PMLR。Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002年。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集，第311–318页。Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。探索统一文本到文本 Transformer 的迁移学习极限。机器学习研究期刊，21(140):1–67。Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022年。利用 CLIP潜变量进行分层文本条件图像生成。*arXiv*预印本 *arXiv:2204.06125*。Suman Ravuri 和 Oriol Vinyals。2019年。条件生成模型的分类准确率得分。神经信息处理系统进展，32。Steven J Rennie、Etienne Marcheret、Youssef Mroueh、Jerret Ross 和 Vaibhava Goel。2017年。用于图像描述的自批判序列训练。载于*IEEE*计算机视觉与模式识别会议论文集，第7008–7024页。Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bjorn Ommer。2022年。利用潜扩散模型进行高分辨率图像合成。载于*IEEE*计算机视觉与模式识别会议论文集（计算机视觉与模式识别会议）。

Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、BurcuKaragol Ayan、SSara Mahdavi、Rapha Gontijo Lopes 等。2022年。结合深度语言理解的照片级文本到图像扩散模型。*arXiv*预印本 *arXiv:2205.11487*。

Tim Salimans、Ian Goodfellow、Wojciech Zaremba、Vicki Cheung、Alec Radford 和 Xi Chen。2016年。改进生成对抗网络训练的技术。神经信息处理系统进展，29。

Yang Song、JaschaSohl-Dickstein、Diederik P Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。2021年。通过随机微分方程实现基于得分的生成建模。载于国际学习表征会议。

Raimonda Stani¯ut˙e 和 Dmitrij Šešok。2019年。图像描述的系统性文献综述。*AppliedSciences*，9(10):2024。

Ingrid Ravn Turkerud 和 Ole Jakob Mengshoel。2021年。使用深度学习进行图像描述：通过回译进行释义改写来增强文本。载于 *2021*年 *IEEE* 计算智能研讨会系列（*SSCI*），第 01–10 页。IEEE。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。《Attention is all you need》。神经信息处理系统进展，30。

RamakrishnaVedantam、C Lawrence Zitnick 和 Devi Parikh。2015年。《Cider：基于共识的图像描述评估》。载于*IEEE*计算机视觉与模式识别会议论文集，第4566–4575页。

Oriol Vinyals、Alexander Toshev、Samy Bengio 和 Dumitru Erhan。2015年。《Show and tell：一种神经图像描述生成器》。载于*IEEE*计算机视觉与模式识别会议论文集，第3156–3164页。

Cheng Wang、Haojin Yang 和 Christoph Meinel。2018年。《结合深度双向长短期记忆网络和多任务学习的图像描述》。*ACM*多媒体计算、通信与应用汇刊（*TOMM*），14(2s)：1–20。

Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhadinov、Rich Zemel 和 Yoshua Bengio。2015年。《Show, attend and tell：利用视觉注意力生成神经图像描述》。载于国际机器学习会议，第2048–2057页。PMLR。

Ling Yang、Zhilong Zhang、Yang Song、Shenda Hong、Runsheng Xu、Yue Zhao、Yingxia Shao、Wentao Zhang、Bin Cui 和 Ming-Hsuan Yang。2022年。扩散模型：方法与应用的全面综述。*arXiv* 预印本 *arXiv:2209.00796*。

Sangdoo Yun、Dongyoon Han、Seong Joon Oh、Sanghyuk Chun、Junsuk Choe 和 Youngjoon Yoo。2019年。CutMix：利用可定位特征训练强大分类器的正则化策略。载于*IEEE/CVF*国际计算机视觉会议论文集，第6023–6032页。

Yuanen Zhou、Zhenzhen Hu、DaqingLiu、Huixia Ben 和 Meng Wang。2022年。《用于图像描述的紧凑型双向Transformer》。*arXiv*预印本 *arXiv:2201.01984*。

Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A Efros。2017年。《利用循环一致对抗网络进行非配对图像到图像翻译》。载于*IEEE*国际计算机视觉会议论文集，第2223–2232页。

Peipei Zhu、Xiao Wang、Yong Luo、Zhenglong Sun、Wei-Shi Zheng、Yaowei Wang 和 Changwen Chen。2022年。《通过图像级弱监督视觉概念识别实现无配对图像描述》。*arXiv*预印本 *arXiv:2203.03195*。

合成数据集

我们的增强数据集的描述性统计数据如桌子6所示。

	# 图像	# 图像描述	词汇表大小
COCO训练集	113,287	566,435	9,486
$COCO_{text}$	113,287	2,423,730	19,952
$COCO_{image}$	113,287	566,435	9,486
SD_{base}	566,747	566,747	9,486
SD_{true}	566,747	2,835,615	9,486
SD_{para}	566,747	2,423,730	19,952

表6：COCO训练集及其增强数据集的描述性统计。

由于 $COCO_{text}$ 不会改变原始 COCO 数据集中的图像，而 $COCO_{image}$ 会将所有 COCO 图像替换为转换后的图像，因此两者的图像数量保持不变。三个增强 SD 数据集中的图像均基于 COCO 图像描述生成，因此图像数量与 COCO 图像描述的数量相同。文本增强应用于 $COCO_{text}$和 SD_{para}。因此，图像描述的数量增加了。同时，释义模型引入了新词，因此这两个数据集的词汇量也得到了扩展。表 7 列出了若干释义后图像描述的示例。

在生成合成图像时，我们并未刻意调整 *StableDiffusion* 的提示词以获得更高质量的图像。因为在生成如此庞大的合成数据集时，提示词调优会耗费大量时间。而且，图像美学并不是我们的】【。

COCO 图像描述 → 改述
一个站在电脑键盘前的小男孩。 →一个小男孩站在电脑键盘前。
一名男子正在厨房里制作披萨。 →在厨房里，一名男子正在制作披萨。
一位女士在炉灶前吃蔬菜。 →一位女士在炊具前吃蔬菜。
小浴室里有一个马桶和一个水槽。 →一间有马桶和水槽的浴室。
一条挤满车辆、亮着停车灯的城市街道。 →一条车辆拥堵、亮着停车灯的城市街道。

表 7：COCO 图像描述及其对应释义改写示例

重点。因此，COCO 图像描述直接用作输入，不进行任何修改。

大多数生成图像中的物体清晰可辨（参见图8的前两行），而其他图像则因文本到图像模型的局限性而存在问题。例如，*Stable Diffusion* 不擅长绘制人手和面部，物体也可能出现奇怪的扭曲（参见图8的后两行）。尽管如此，我们已经证明，利用这些不完美的合成图像进行训练，图像描述模型仍能取得相当不错的表现。

由于用于生成图像的*Stable Diffusion* 模型和用于释义改写的 T5 模型都是开源模型，因此生成图像和文本数据不存在版权问题。我们还开放了合成 COCO 数据集，供后续研究免费使用。

B 实验结果

此处列出全部三项实验的详细结果。

表 9 展示了实验 1 中 FC模型和基于 Transformer 的模型的结果；在该实验中，我们使用三个合成 SD 数据集训练了图像描述模型。

在实验 2 中，我们对有限的 COCO 数据进行数据增强，将采样的 COCO 数据集与我们的合成 SD 数据集混合。基线模型在这些增强数据集上进行训练，并比较我们提出的方法与基线数据增强方法在图像描述方面的性能。FC模型和基于 Transformer 的模型的结果分别如表10 和表 11所示。COCO 数据集与..... 相结合后的性能

所选合成数据集也列于表 10 和表 11 中，便于比较。

我们在实验 3 中检验了合成数据质量与其下游性能之间的关系。使用筛选后的 SD_{true}数据集训练的两个基线模型的图像描述性能见桌子 12，并与SD_{true} 和 SD_{para}进行了比较。数据稀缺情形下数据质量筛选的结果也见桌子10 和桌子11。

COCO 图像	生成图像
	
	
	
	

表 8：COCO 图像和合成图像示例。COCO 图像位于左列，右列为对应的生成图像。

C 图像描述示例

为了让读者更直观地了解应用所提出的多模态数据增强后的改进，本节展示了一些图像描述示例。示例图像取自 COCO 测试集，训练后的模型生成的相应图像描述列于右侧。

模型	训练集	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
FC模型	COCO	72.2	54.6	39.7	28.7	24.4	52.4	92.4	17.6
	SD_{base}	68.4	49.3	34.2	23.6	21.7	48.6	75.4	15.3
	SD_{para}	69.2	50.1	34.5	23.4	21.6	48.9	75.5	15.4
	SD_{true}	**69.5**	**50.9**	**36.1**	**25.6**	**22.6**	**50.1**	**81.0**	**15.6**
T模型	COCO	75.5	59.2	44.7	33.5	27.4	55.8	111.3	20.6
	SD_{base}	70.0	52.0	37.2	26.2	23.8	50.6	87.5	17.5
	SD_{para}	**71.5**	**53.5**	38.1	26.8	23.9	51.2	89.0	18.0
	SD_{true}	70.4	53.1	**38.9**	**28.4**	**25.1**	**52.2**	**94.1**	**17.8**

表 9：在三个合成数据集上训练的 FC模型和基于 Transformer 的模型的图像描述性能。

训练集	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
10% COCO	68.0	49.4	34.6	24.3	22.0	49.1	75.5	15.0
10% $COCO_{text}$	68.0	49.6	34.5	23.8	21.6	49.1	74.6	15.0
10% $COCO_{img}$	67.5	49.1	34.4	24.1	22.0	49.0	75.7	15.0
10% COCO + SD_{base}	69.9	51.8	36.8	25.8	23.1	50.3	83.5	16.4
10% COCO + SD_{para}	70.5	52.3	37.0	26.0	22.9	50.4	84.2	**16.6**
10% COCO + SD_{true}	70.0	51.9	36.9	26.2	23.1	50.7	84.7	16.1
10% COCO + 已选择 SD_{true}	**70.6**	**52.5**	**37.6**	**26.8**	**23.3**	**51.1**	**86.2**	16.5
20% COCO	69.2	50.9	36.3	25.7	22.8	50.2	80.6	15.8
20% $COCO_{text}$	69.2	51.0	35.9	24.9	22.2	49.9	79.1	15.7
20% $COCO_{img}$	69.4	51.5	36.9	26.3	22.9	50.4	82.5	16.0
20% COCO + SD_{base}	70.7	52.5	37.4	26.5	23.4	50.8	85.5	16.8
20% COCO + SD_{para}	**71.5**	**53.0**	37.5	26.3	23.1	50.8	86.1	16.8
20% COCO + SD_{true}	70.6	52.4	37.6	26.9	23.4	51.0	86.9	16.6
20% COCO + 已选择 SD_{true}	70.7	52.6	**37.8**	**27.1**	**23.5**	**51.2**	**87.2**	**17.0**
30% COCO	70.2	52.2	37.4	26.6	23.3	50.9	85.1	16.4
30% $COCO_{text}$	70.3	52.4	37.2	26.1	22.8	50.9	83.1	16.1
30% $COCO_{img}$	69.9	52.1	37.4	26.7	23.4	50.9	85.4	16.3
30% COCO + SD_{base}	70.5	52.3	37.3	26.3	23.4	50.8	85.5	16.6
30% COCO + SD_{para}	**71.6**	**53.5**	**38.3**	27.0	23.5	51.4	87.2	**17.0**
30% COCO + SD_{true}	70.4	52.4	37.6	26.9	23.4	51.2	86.0	16.6
30% COCO + 已选择 SD_{true}	70.9	52.9	38.0	**27.0**	**23.7**	**51.4**	**87.8**	16.9
40% COCO	71.0	53.1	38.3	27.4	23.6	51.5	87.8	16.7
40% $COCO_{text}$	70.3	52.4	37.2	26.2	22.7	51.0	83.6	16.1
40% $COCO_{img}$	70.7	52.9	38.0	27.2	23.6	51.3	87.5	16.6
40% COCO + SD_{base}	70.8	52.7	37.6	26.6	23.5	51.1	86.0	16.7
40% COCO + SD_{para}	**71.6**	**53.7**	**38.5**	27.3	23.6	**51.4**	**88.5**	**17.0**
40% COCO + SD_{true}	70.4	52.7	37.8	27.1	23.6	51.3	87.1	16.6
40% COCO + 已选择 SD_{true}	70.9	52.9	38.1	**27.3**	**23.7**	51.4	88.3	17.0
50% COCO	71.3	53.5	38.5	27.5	23.7	51.5	88.1	16.8
50% $COCO_{text}$	70.4	52.6	37.4	26.4	22.9	51.1	84.3	16.3
50% $COCO_{img}$	71.1	53.3	38.5	27.7	23.8	51.8	88.3	16.8
50% COCO + SD_{base}	70.8	52.7	37.7	26.8	23.6	51.2	87.3	16.8
50% COCO + SD_{para}	**71.8**	**53.9**	**38.7**	27.5	23.7	**51.6**	88.6	**17.2**
50% COCO + SD_{true}	70.8	52.8	37.9	27.1	23.6	51.3	87.3	16.8
50% COCO + 已选择 SD_{true}	71.2	53.4	38.5	**27.6**	**23.7**	51.5	**89.3**	16.9

表10：使用有限真实COCO数据的FC模型的数据增强性能，包括每组最后一行的数据质量筛选性能。

训练集	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
10% COCO	69.3	51.5	37.1	26.3	24.1	50.8	88.2	17.6
10% $COCO_{text}$	69.5	51.9	36.9	25.9	23.9	50.5	87.1	17.3
10% $COCO_{img}$	68.7	50.8	36.3	25.7	23.6	50.5	85.3	17.0
10% COCO + SD_{base}	72.3	54.8	39.8	28.7	24.9	52.1	95.6	18.7
10% COCO + SD_{para}	**72.3**	54.5	39.2	27.7	24.7	51.8	94.1	**18.8**
10% COCO + SD_{true}	71.5	54.5	40.1	29.4	**25.5**	53.0	97.5	18.6
10% COCO + 已选中 SD_{true}	72.1	**54.9**	**40.5**	**29.5**	25.4	**53.1**	**98.3**	18.7
20% COCO	71.0	53.5	39.1	28.2	25.3	52.3	94.8	18.9
20% $COCO_{text}$	71.7	53.9	38.5	27.0	24.6	52.2	92.7	18.5
20% $COCO_{img}$	70.6	53.3	38.9	28.2	25.2	52.1	94.2	18.5
20% COCO + SD_{base}	72.5	55.1	40.1	28.8	25.7	52.8	99.2	19.4
20% COCO + SD_{para}	**74.0**	**56.6**	41.5	30.1	25.7	53.3	100.5	19.2
20% COCO + SD_{true}	73.2	56.5	**42.1**	**31.1**	**26.2**	**53.9**	**103.4**	**19.4**
20% COCO + 已选中 SD_{true}	72.0	55.0	40.6	29.7	25.8	53.1	99.9	19.2
30% COCO	72.5	55.3	40.8	29.8	25.7	53.2	99.7	19.4
30% $COCO_{text}$	73.5	56.4	41.5	29.8	25.1	53.3	97.3	19.0
30% $COCO_{img}$	71.3	54.0	39.7	29.0	25.8	52.7	98.2	19.0
30% COCO + SD_{base}	73.5	56.4	41.7	30.6	25.9	53.6	102.4	19.5
30% COCO + SD_{para}	**74.4**	**57.6**	**42.8**	31.2	26.1	54.1	104.2	19.5
30% COCO + SD_{true}	73.8	57.0	42.6	**31.4**	**26.5**	**54.6**	**104.6**	**19.8**
30% COCO + 已选择 SD_{true}	72.4	55.5	41.2	30.4	26.1	53.6	102.2	19.2
40% COCO	73.4	56.4	42.2	31.2	26.5	54.2	103.9	20.0
40% $COCO_{text}$	73.3	55.8	40.6	28.8	25.2	53.0	97.0	19.2
40% $COCO_{img}$	73.0	55.8	41.3	30.4	26.2	53.8	101.5	19.6
40% COCO + SD_{base}	74.1	57.0	42.4	31.1	26.3	54.2	104.5	19.9
40% COCO + SD_{para}	**74.9**	**57.9**	43.0	31.7	26.5	54.5	106.5	**20.0**
40% COCO + SD_{true}	74.1	57.6	**43.3**	**32.2**	**26.9**	**55.0**	**106.9**	19.9
40% COCO + 已选择 SD_{true}	72.8	56.1	41.8	31.0	26.2	54.1	103.1	19.4
50% COCO	74.3	57.3	42.5	31.0	26.7	54.6	105.1	20.0
50% $COCO_{text}$	74.4	57.6	42.6	31.1	25.7	54.1	102.3	19.5
50% $COCO_{img}$	73.6	56.5	42.1	31.1	26.5	54.1	103.3	19.7
50% COCO + SD_{base}	73.8	57.0	42.5	31.5	26.4	54.2	105.5	19.8
50% COCO + SD_{para}	**74.9**	**58.3**	**43.7**	**32.6**	**26.8**	**55.0**	**107.8**	**20.1**
50% COCO + SD_{true}	74.6	58.0	43.4	32.2	26.7	54.8	107.1	19.9
50% COCO + 已选择 SD_{true}	73.2	56.5	42.3	31.2	26.7	54.4	104.7	19.8

表 11：使用有限真实COCO数据的基于 Transformer 的模型的数据增强性能，包括每组最后一行的数据质量筛选性能。

模型	训练集	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
FC模型	COCO	72.2	54.6	39.7	28.7	24.4	52.4	92.4	17.6
	SD_{base}	**68.4**	**49.3**	**34.2**	**23.6**	21.7	48.6	**75.4**	15.3
	已选择 SD_{base}	67.4	48.8	34.0	23.5	**22.2**	**48.9**	75.3	**15.6**
	SD_{para}	**69.2**	**50.1**	34.5	**23.4**	21.6	**48.9**	**75.5**	15.4
	已选择 SD_{para}	67.5	48.8	**33.8**	23.2	**22.0**	48.7	75.0	**15.6**
	SD_{true}	69.5	50.9	36.1	**25.6**	22.6	50.1	81.0	15.6
	已选择 SD_{true}	**69.6**	**51.3**	**36.3**	25.5	**22.7**	**50.2**	**81.0**	**15.7**
T模型	COCO	75.5	59.2	44.7	33.5	27.4	55.8	111.3	20.6
	SD_{base}	**70.0**	**52.0**	37.2	**26.2**	23.8	50.6	87.5	17.5
	已选择 SD_{base}	69.9	52.0	**37.2**	26.1	**24.1**	**50.7**	**87.9**	**17.9**
	SD_{para}	**71.5**	**53.5**	**38.1**	**26.8**	23.9	**51.2**	89.0	**18.0**
	已选择 SD_{para}	70.9	53.0	37.9	26.7	**24.0**	51.1	**89.4**	17.9
	SD_{true}	**70.4**	**53.1**	**38.9**	**28.4**	**25.1**	**52.2**	**94.1**	17.8
	已选择 SD_{true}	70.3	52.9	38.5	27.7	24.7	51.8	93.0	**17.8**

表 12：FC模型和基于 Transformer 的模型在完全合成数据集上训练，并根据排名前 50% 的 *CLIPScore* 标准选择高质量数据后的图像描述性能。

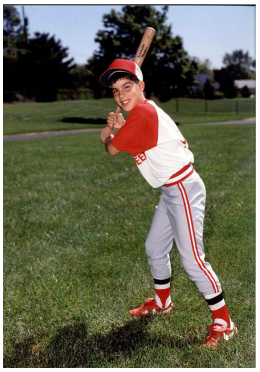
图像	图像描述
	**人工标注**：烘焙披萨配红色番茄和绿色橄榄。 **FC 模型**： **10% COCO**：一张披萨上[illegible]放在白色盘子上 **10% COCO$_{text}$**：一张披[illegible]盘子上 **10% COCO$_{img}$**：一张披萨[illegible]桌子上 **10% COCO + SD$_{base}$**：[illegible]披萨上有很多配料 **10% COCO + SD$_{para}$**：一份披[illegible]在白色盘子上 **10% COCO + SD$_{true}$**：一张披[illegible]
	**人工标注**：这名足球运动员正在将球重新带回比赛中。 **FC 模型**： **10% COCO**：一名男子正在玩棒球 **10% COCO$_{text}$**：一名穿着[illegible]衬衫和白[illegible]车、正在进行棒球 **10% COCO$_{img}$**：一名男子正在玩足球，地点是[illegible]球场上 **10% COCO + SD$_{base}$**：一名男子正在踢足球 **10% COCO + SD$_{para}$**：一个穿绿色衬衫的男人和[illegible]球 制服 **10% COCO + SD$_{true}$**：一名男子身[illegible]红色衬衫正在踢
	**人工标注**：一只白色的大狗坐在一位老人旁边的长椅上。 **FC 模型**： **10% COCO**：一只狗和[illegible]狗在长椅上 **10% COCO$_{text}$**：[illegible]奇上 **10% COCO$_{img}$**：一只狗和[illegible]狗坐在长椅上 **10% COCO + SD$_{base}$**：一只狗坐在长椅上，旁边还 **10% COCO + SD$_{para}$**：一[illegible]坐在长椅上，旁边有一只狗 **10% COCO + SD$_{true}$**：一[illegible]子 坐在长椅上，旁边有一只狗
	**人工标注**：一个穿着棒球制服的小男孩手握球棒，做好了挥棒的准备。 **基于 Transformer 的模型**： **COCO**：[illegible]子拿着一根[illegible]公园里 **SD$_{base}$**：一名男子[illegible] **SD$_{para}$**：[illegible]子握着一[illegible]在球场上 **SD$_{true}$**：[illegible]子握着一[illegible]在球场上

表 13：在 COCO 数据集和我们的合成数据集上训练的基线模型生成的图像描述。红色表示检测到错[illegible]的物体或词语使用不当（例如，重复已有词语）；蓝色突出显示正确的物体或词语[illegible]当使用。我们可以观察到，我们的方法优于基线模型。此外，我们的方法生成的图像描述质量接近人工撰写的句子。