arXiv:2304.10824v1 [cs.CV] 21 Apr 2023

重新思考跨模态图文检索基准

Weijing Chen
c
henweijing17@ruc.edu.cn
中国人民大学 中国北京

Linli Yao
linliyao@ruc.edu.cn 中国人民大学 中国北京

Qin Jin*
qjin@ruc.edu.cn 中国人民大学 中国北京

摘要

图文检索是信息检索中一个基础且重要的分支，已受到广泛关注。该任务的主要挑战在于跨模态语义理解与匹配。近期一些研究更加关注细粒度的跨模态语义匹配。随着大规模多模态预训练模型的普及，若干最先进模型（例如 X-VLM）在广泛使用的图文检索基准上取得了近乎完美的性能，即 MSCOCO-Test-5K 和 Flickr30K-Test-1K。本文回顾了这两个常用基准，并发现它们不足以评估模型在细粒度 跨模态语义匹配方面的真实能力。这是因为基准中的大量图像和文本都是粗粒度的。基于这一观察，我们改进了旧基准中的粗粒度图像和文本，并建立了改进后的基准，称为*MSCOCO-FG*和*Flickr30K-FG*。具体而言，在图像方面，我们通过纳入更多相似图像来扩充原有图像池。在文本方面，我们提出了一种新颖的半自动改进方法，只需少量人工投入，便可将粗粒度句子细化为更细粒度的句子。此外，我们在新基准上评估了具有代表性的图文检索模型，以验证该方法的有效性。我们还通过大量实验分析了模型的细粒度语义理解能力。结果表明，即使是最先进模型，在细粒度语义理解方面仍有很大提升空间，尤其是在区分图像中相近对象的属性方面。我们的代码和改进后的基准数据集已公开发布1，希望能推动跨模态检索领域进一步深入研究。

CCS 概念

• **信息系统 → 检索结果评估；多媒体与多模态检索。**

关键词

图文检索，跨模态检索，评估基准

*通讯作者。[1]

https://github.com/cwj1412/MSCOCO-Flikcr30K_FG

允许为个人使用制作本作品全部或部分内容的数字或纸质副本在课堂上使用无需付费，前提是不得为营利或商业利益复制或分发副本，且副本必须附有本声明及首页上的完整引文。对于本作品中由作者以外的其他人拥有版权的部分，必须尊重其版权。注明出处后，可以摘录。以其他方式复制、再版、发布到服务器或分发到列表，均需事先获得明确许可和/或支付费用。请联系 permissions@acm.org 申请许可。*SIGIR '23，2023*年7月*23*日至*27*日，台北，台湾© 2023 版权所有归所有者/作者所有。出版权已授权给 ACM。

ACM ISBN 978-1-4503-9408-6/23/07...$15.00
https://doi.org/10.1145/3539618.3591758

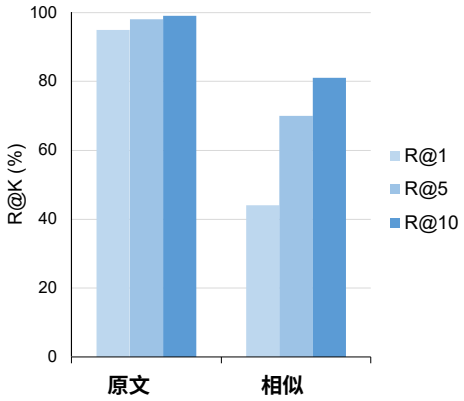

图 1：最先进模型 X-VLM 在不同候选图像上的文本到图像（T2I）检索性能。我们观察到，在相似图像上进行检索时，模型性能显著下降。

ACM 引用格式：
Weijing Chen、Linli Yao 和 Qin Jin。2023年。重新思考跨模态图文检索基准。载于第*46*届国际*ACM SIGIR*信息检索研究与开发大会论文集（*SIGIR '23*），*2023*年*7*月*23*日至*27*日，台湾台北。ACM，美国纽约州纽约市，11页。https://doi.org/10.1145/3539618.3591758

1 引言

图文检索作为跨模态研究的基础课题之一，在搜索引擎、推荐系统和问答系统等现实场景中有着广泛的应用。它要求机器根据文本查询检索图像，或根据图像查询查找匹配的描述。图文检索的关键挑战在于学习不同模态之间的语义对应关系。为实现更好的跨模态检索，当前主流的图文检索研究往往更关注细粒度的跨模态语义理解，并采用各种精心设计的对齐与推理模块，包括注意力机制 [18, 24, 39–41], 基于图的网络 [4, 20, 25], 以及场景图 [36, 38] 等。

近年来，大规模多模态预训练模型 [2, 32, 44] 的引入显著提升了图文检索的性能。人们投入了大量努力，为一些现有预训练模型设计预训练任务，以促进细粒度语义理解，使其在图文检索中取得了近乎完美的表现。多个最先进模型（如 X-VLM [44], ）已在常用基准数据集 MSCOCO-Test-5K 上的图像到文本（图像到文本）检索和文本到图像检索中，分别取得了 R@10 为 98.2 和 90.5 的成绩。



图 2：当前基准数据集中"粗粒度图像和文本"问题示意图：(a) 候选图像差异很大，使模型无需细粒度语义理解即可轻易匹配目标图像；(b) 然而，当候选图像相似时，如果缺乏良好的细粒度语义理解能力，模型便无法完成匹配；(c) 粗粒度文本会导致基准数据集中的图像查询意外匹配到非目标图像；(d) 细粒度文本能够很好地区分相似的图像查询。

当前模型近乎完美的表现令人振奋，也引发了我们的好奇：这些模型是否真的具备可与人类相媲美的跨模态语义知识，尤其是细粒度语义理解能力？还是说，现有基准数据集不足以验证模型在跨模态细粒度语义理解方面的真实能力？因此，我们开展了一项小型文本到图像检索实验来检验这一假设。具体而言，我们设计了两种测试设置。首先，从 MSCOCO-Test-5K 中随机选取 100 个图文对，作为 mini-test(original) 设置。然后，对于 mini-test(original) 中给定的文本查询，我们从 MSCOCO-Test-5K 和辅助来源（见第 3.1 节）中，特意挑选 99 张与其对应目标图像相似的图像；这些图像与目标图像共同构成更具挑战性的文本到图像检索候选集，称为 mini-test(similar) 设置。图 1 展示了 X-VLM 在这两种小型实验设置[2] 下的表现，可以观察到，从 mini-test(original) 到 mini-test(similar)，性能显著下降。这些观察促使我们重新审视当前基准数据集，并发现了两个问题。

首先，**用于基准测试的图像是"粗粒度"的**，这意味着图像检索池的规模较小，且池中的图像差异很大（或者说池在语义上较为稀疏），因此无需细粒度语义理解，就能轻易辨别检索目标。两种常见的图文检索基准数据集 MSCOCO-Test-5K [23] 和 Flickr30K-Test-1K [42], 分别仅包含 5000 张和 1000 张图像。此外，这些图像是从 Flickr30K 和 MSCOCO 中随机选取的。规模有限且语义稀疏的检索候选池会导致图文检索评估存在系统性弱点。如图 2(a) 所示，对于给定的查询"一位戴着眼镜和围巾的女性在人行道上滑滑板"，检索池包含各种各样的图像，其中大多数在语义上与查询无关。在这种情况下，模型可以轻易地从候选图像池中选出目标图像而无需进行细粒度语义理解。然而，如果要求模型从相似的候选图像中检索，模型可能会因缺乏足够的细粒度语义理解能力而失败，如图 2(b) 所示。

其次，**用于基准测试的文本是"粗粒度"的**。基准数据集中的大量文本描述不够详细。在 MSCOCO 和 Flickr30K 中，每张图像都由人工标注了五个句子，其中一些描述更细致（例如"长颈鹿站在一片绿地上，背景中有一栋房子"），而另一些则较为粗略（例如"长颈鹿站在一片田野里"）。然而，粗粒度文本可能会给图文检索评估带来问题：不同的图像查询可能都匹配到同一条粗粒度文本，如图 2(c) 所示。这种评估情形并非能够验证模型能力的有效测试样本。相反，如果文本描述更加细致，就可以作为验证模型细粒度语义理解能力的有效样本，如图 2(d) 所示。

为解决当前基准数据集中上述描述的缺陷，我们提出改造现有基准数据集并构建相应的新基准数据集，即 **MSCOCO-FG** 和**Flickr30K- FG**。我们首先采用相似图像来扩充原始候选池，以获得更大且语义更密集的候选池。具体而言，我们利用多模态模型从原始候选池和辅助来源中搜索与目标图像相似的图像，然后将包括目标图像在内的所有相似图像组合起来，形成新的图像池。此外，我们提出了一种新颖的半自动方法来细化粗粒度文本。具体来说，我们首先利用视觉-语言预训练模型检测潜在的粗粒度文本，然后通过提示补充新细节来丰富这些文本。最后，对细化文本进行一些人工校正。

我们进一步开展了大量实验，在新的基准数据集上测试各种现有模型。实验结果表明，在旧基准数据集上表现近乎完美的模型，在我们的新基准数据集上已无法保持完美表现，这说明当前的图文检索模型在跨模态细粒度语义理解方面仍

[2]mini-test(similar) 设置下的性能是对全部 100 个查询取平均所得。

有很大的提升空间。本研究的主要贡献包括：

- 我们回顾了当前的图文检索基准数据集，并发现它们不足以评估模型的细粒度语义理解能力。
- 我们通过利用相似图像扩充图像池，并以半自动化的方式改进粗粒度文本，对现有基准数据集进行改造，同时构建用于图文检索的新型细粒度基准数据集 MSCOCO-FG 和 Flickr30K-FG。这些新基准数据集将予以公开，以支持更深入的研究。
- 我们在新的基准数据集上评估了若干具有代表性的图文检索模型，以验证我们方法的有效性，并进一步分析它们理解跨模态细粒度语义的能力。

2 相关工作

2.1 图文检索

多年来，人们提出了各种图文检索方法，大致可分为两类。

非预训练模型。早期工作 [7, 8, 28] ，如 VSE++[7] ，主要使用卷积神经网络（CNN）提取固定网格特征来表示图像，但这类特征只能提供局部像素级信息，难以理解高层语义概念。为解决这一问题，[18]后续工作[24, 39, 41]采用对象检测器，以更好地编码图像。例如，SCAN [18] 采用 Faster R-CNN [34] 检测图像中的对象，并将其与句子中的词语对齐。近年来，研究人员探索了多种方法 [4, 20, 30, 31, 36, 38, 40, 45] ，以实现图像与文本之间更细粒度的语义对齐。具体而言，VSRN [20]利用图卷积网络（GCN） [16]和门控循环单元（GRU） [1] 进行局部—全局语义推理。SGM [38] 构建外部场景图 [14] ，以增强视觉关系的学习。DIME [31] 则旨在动态学习模态交互模式。

视觉语言预训练模型。 视觉语言预训练（VLP）旨在通过自监督任务，从大量图文对中学习视觉与语言之间的对齐关系。在下游图文检索任务上进行微调后，视觉语言预训练模型的性能显著优于非预训练模型。其架构大致可分为两种类型：单流和双流。采用单流结构的模型 [2, 9, 15, 19, 22, 46]使用同一个编码器对图像和文本进行编码，并通过不同的预训练任务学习视觉与语言之间的对齐关系。例如，UNITER [2] 采用统一的 Transformer 编码器，在共同的空间中学习图像区域和词语的上下文化嵌入。采用双流结构的模型 [6, 13, 21, 26, 32, 43, 44]首先使用相互独立的编码器分别对图像和文本进行编码，然后通过协同注意力层实现跨模态语义交互，从而提供更加灵活的图像和文本编码方式。具体而言，CLIP[32] 采用基于 ResNet-50 [10]/ViT [5] 的图像编码器，并使用 Transformer 对文本进行编码。在 4 亿个图文对上进行对比式预训练后，两个独立的编码器获得了为两种模态编码通用表征的能力。

为了研究先前模型在细粒度跨模态语义理解与对齐方面的能力，我们选择了若干具有代表性的非预训练和预训练模型，在我们新的基准数据集上进行评估，并在接下来的实验中提供详细分析。

2.2 图像-文本数据集

随着跨模态研究受到越来越多的关注，各种跨模态数据集，尤其是图文数据集，已被提出，以支持深入探索。根据数据收集方法，现有的图文数据集可分为两类。

带有人类标注的数据集。早期，研究人员先从互联网上爬取图像，然后手动为图像标注标签、区域和文本描述，以此创建图像文本数据集。继 Hodosh 等人 [12], 之后，Young 等人 [42]推出 Flickr30K，其中包含从 Flickr 网站收集的 31,783 张日常活动和场景图像，并通过众包（Amazon MechanicalTurk）为每张图像标注了 5 条描述。Lin 等人 [23] 提出了一个规模更大、内容更详细的数据集 MSCOCO，其中包含 123,287 张属于 91 个常见物体类别的图像（同样收集自 Flickr），并为每张图像手动标注了边界框、分割区域和 5 条描述。随后，Krishna 等人[17] 基于 MSCOCO 提出了 Visual Genome，为视觉区域补充了描述，以提供更丰富的图像细节。由于质量较高，这些数据集被广泛用于训练和评估各种图像文本任务。

经过筛选的网络资源数据集。随着视觉-语言预训练模型在自然语言处理领域表现出色，研究人员倾向于扩大图像-文本数据集的规模，以训练更好的多模态模型。为克服人工标注数据集的局限性，他们大多利用网络资源构建未经人工标注的数据集。Thomee 等人 [37] 从 Flickr 收集了约 9,900 万张图像及其元数据（包括标签和位置），构建了 YFCC-100M。由于网络数据往往存在噪声，SBU Captioned Photo Dataset [29]和 Conceptual Captions [35] 提取原始 HTML 文本中的替代文本，并按照若干预设规则进行清理，以获得更准确、详细的图像描述。除了依赖替代文本，一些数据集还采用其他文本来源。例如，RedCaps [3]收集社交媒体用户撰写的说明文字，构建图像-文本对。尽管这些数据集在数量上优于标注数据集，但即使采用精心设计的筛选流程，其质量仍有缺陷。因此，它们大多用于预训练阶段，不适合用于评估。

总而言之，带有人类标注的图文数据集更适合用于基准评测，因此，常见的图文检索基准主要基于人工标注的数据集，例如 MSCOCO 和 Flickr30K。然而，这些基准仍不足以评估细粒度图文检索。为此，我们提出通过半自动化方法改造现有基准，构建新的基准，从而支持细粒度评估并节省人工重新标注的成本。

3 方法

传统的图文检索基准 B 通常包含 N 张图像，每张图像都有 M 条对应的文本描述，通常用于评估两个子任务：文本到图像

（文本到图像）检索和图像到文本（图像到文本）检索。文本到图像检索旨在根据文本查询，从整个候选图像池 I 中搜索目标图像；图像到文本检索则要求模型根据图像查询，从候选文本池 D 中至少检索出一条目标文本描述。正如我们所观察到的，当前基准数据集中的图像和文本粒度较粗，无法提供细粒度的图文检索评估。考虑到从头构建全新的基准数据集成本高昂且耗时，我们提出改进现有基准数据集 B，以满足更高的细粒度评估要求。接下来的第 3.1 节首先介绍如何构建更具挑战性、粒度更细的图像池。第 3.2 节介绍我们将粗粒度文本优化为细粒度文本的半自动方法。

3.1 改造图像候选池

由于当前图像检索池中的图像差异较大，模型无需进行细粒度理解，仅根据文本查询就能轻松选出正确的目标图像。为了改进现有基准数据集，使其支持细粒度语义评估，我们利用相似图像来扩充图像池。将原始图像池记为 Iold，并将池中的每张图像（目标图像）记为 I_t (t∈[0,N])，我们为 I_t 搜索相似图像，通过以下三个步骤将 Iold 扩展为新的图像池 Inew：(i) 准备一个无重叠的图像池 Ican，以提供候选图像；(ii) 从 Ican 中为每个 I_t 搜索一个相似图像集 S_t；(iii) 将所有相似图像集 $\{S_t\}$ 汇集到一个新的图像池 Inew 中。

3.1.1 准备候选图像。在搜索相似图像以扩充原始图像池之前，我们应首先确定哪些图像可以作为相似图像的候选图像。除了原始图像池中的图像外，我们还需要更多候选图像。我们根据以下三个原则选择候选图像：

- 候选图像不能与图文检索模型通常使用的训练集和验证集重叠，以防止数据泄漏。
- 候选图像应与目标图像属于同一领域，以提高其成为相似图像的可能性。
- 候选图像的数量应足够大，远远超过原始候选池的规模，从而为相似图像提供多样化的候选项。

由于两个广受欢迎的图文检索数据集 MSCOCO 和 Flickr30K 都包含来自图片分享网站 Flickr 的图像，我们采用从 Flickr 收集的其他图像，作为相似图像的额外候选来源。具体而言，我们采用 COCO[4] (unlabeled2017) 中 120,000 张未标注图像，并将其与原始图像池 Iold 中的图像合并，作为候选图像 Ican。

3.1.2 搜索相似图像。基于候选图像 Ican，我们通过图像到图像和文本到图像匹配，为每个目标图像 I_t 建立相似图像集 S_t，如图 3 所示。具体而言，搜索过程包括以下三个主要组成部分：

[3]https://www.flickr.com/ [4] https://cocodataset.org/

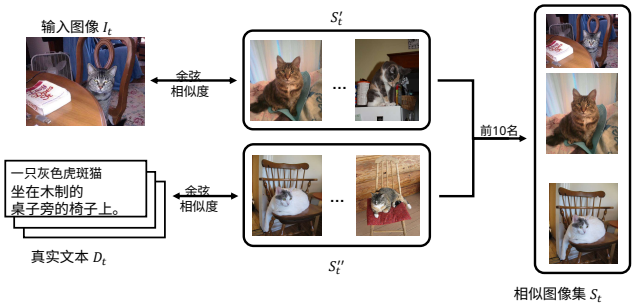

图 3：通过图像到图像和文本到图像匹配，为目标图像构建相似图像集的概览。

(1) 首先，我们使用视觉语言模型分别从图像和文本中提取视觉特征和文本特征。具体来说，我们使用预训练的 CLIP [32] 模型，将 Ican 中的所有候选图像（包括 Iold 中的目标图像）以及 D 中的所有文本嵌入到共享语义空间中。(2) 然后，对于 Iold 中的每个 I_t，我们计算其与 Ican 中所有候选图像的余弦相似度分数，并选出前-k_I 张图像作为 S_{It}。对于其对应的 M 文本描述 $D_t = \{d_i\}_{i=1}^{M}$，我们也计算它们与所有候选图像的相似度分数，并仅保留前-k_{II} 张图像作为 S_{IIt}。(3) 具体而言，我们结合 S_{It} 和 S_{IIt}，选出前10张图像组成相似图像集 S_t，其中包括目标图像 I_t 和与 I_t 最相似的9张图像。

3.1.3 组装相似图像集。最后，我们汇总前述步骤中构建的所有相似图像集 $S = \{S_i\}_{i=1}^{N}$，并从 S 中移除重复图像，以获得新的图像池 Inew，其中包含完整的 Iold 以及大量补充相似图像。新图像池 Inew 的规模更大，图像之间的相似度也更高，因此图文检索需要更强的细粒度语义理解能力。

3.2 改进粗粒度文本

现有基准数据集中存在大量粗粒度文本（如图 2(c) 所示），这可能导致歧义问题，使得不同的非目标图像都与该文本匹配。为解决这一问题，我们提出了一个包含四个步骤的半自动改进流程，从粗粒度文本 d_f 中提取细粒度文本 d_c，如图 4 所示。

3.2.1 查找粗粒度文本。当前基准数据集中既有粗粒度文本，也有细粒度文本，只有粗粒度文本需要改进。由于人工检查成本高且耗时较长，我们利用先进的视觉语言预训练模型筛选粗粒度文本。CLIP [32] 在包括图像-文本匹配在内的多种零样本图文任务上表现出有竞争力的性能。因此，我们在此步骤中使用预训练的 CLIP。如果 CLIP 根据文本描述 d_0 能够将目标图像 I_0 排在首位，那么我们认为该文本 d_0 的细粒度足以让模型将目标图像与其他图像区分开来；否则，我们将其视为粗粒度文本 d_0。经过筛选，原始基准数据集 B 中超过一半的文本被判定为粗粒度文本。随后，我们将这些文本传入后续步骤，进一步将其细化为更细粒度的文本。

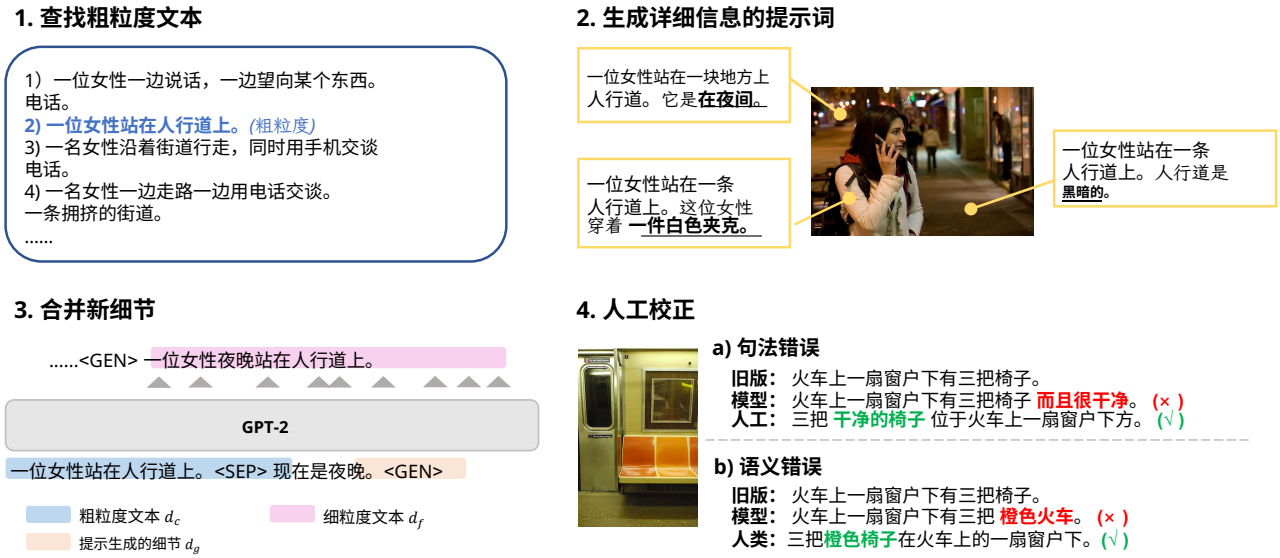

图 4：用于将粗粒度文本改进为细粒度文本的半自动改进流程。

表 1：五个示例提示，用于启发 VinVL 从不同方面生成细节。X 是文本中出现的名词（对象）的占位符，例如“人行道”

Row	提示词	细节
1	这是	新对象/地点/天气
2	存在	新对象
3	X 是	属性
4	X 的颜色	颜色
5	男性/女性穿着/在	穿着

3.2.2 生成细节的提示词。 为了细化第一步筛选出的粗粒度文本，我们需要从图像中提取更多新细节，并将其补充到原始粗粒度文本中。由于预训练视觉-语言模型在图像描述生成方面已取得出色表现，我们采用一个预训练模型来挖掘图像中的更多细节。具体而言，我们在基准数据集上微调 VinVL [46]，并使用表 1 所示的不同类型提示词，促使 VinVL 根据原始文本和图像生成各种细节。具体流程如下：

(1) 给定粗粒度文本 d_c，我们首先使用自然语言处理工具 Stanford CoreNLP[5] 对其进行分析，并从中提取 m 名词（对象），记作 $X = \{x_i\}_{i=1}^{m}$。 (2) 根据模板，利用 X 中的名词创建不同的提示词 $P = \{p_i\}_{i=1}^{n'}$，并在 d_c 之后分别应用这些提示词，以获得输入文本 $\{d_c^i\}_{i=1}^{n'}$，其中 $d_c^i = d_c, p_i$。 (3) 将每个 d_c^i 与对应的图像 I_c 一起输入 VinVL，并获取新的文本 $d_n = d_c, d_g$。 (4) 筛选结果，仅保留包含新信息的结果。具体而言，我们应用 CLIPScore [11] 来衡量图文兼容性，并丢弃任何 CLIPScore 低于 d_c 的 d_n。

3.2.3 合并新细节。 注意，d_c 和 d_n 的格式不同，其中 d_c 仅用一个句子描述图像（例如，一位女性站在人行道上），而 d_n 采用两个句子的格式（例如，一位女性站在人行道上。现在是夜晚）。为了使新文本保持相同的格式，我们使用自然语言模型将 d_n 的两个句子合并为一个句子的格式（例如，一位女性夜晚站在人行道上）。具体而言，我们微调 GPT-2 [33] 来执行句子合并：

(1) 首先，我们需要构建一些与 $d_n = d_c, d_g$ 格式相同的训练数据，用于 GPT-2。我们首先使用自然语言处理工具 Spacy[6] 为每个细粒度文本创建解析树。然后，我们根据解析树中的词性标签和依存关系信息，将内容拆分为两部分，分别对应细节部分 $d'g$ 和其余部分 $d'c$。具体而言，我们主要使用两种结构来构建 $d'g$：形容词-名词结构 和介词短语。我们首先提取文本中所有形容词及其相关对象的配对，以及所有介词短语。然后，我们遵循生成细节句子 d_g 的相似格式来构建 $d'g$ 作为训练数据。例如，我们将“一个小男孩在山顶玩飞盘。”拆分为“一个小男孩玩飞盘。”和“它在山顶上。”
(2) 然后，我们使用根据上述步骤构建的训练数据 $d' = d'c, d'g$ 对 GPT-2 进行微调。具体来说，给定文本 $d'c, d'g$

，

模型需要重建 $d\prime$。我们留出部分训练数据，以检验模型是否学会准确合并文本。(3) 将 d_n 输入微调后的 GPT-2，并以单句形式生成细粒度文本 d_f。如图 4 第 3 步所示，我们将原始粗粒度文本 “一位

[5]https://stanfordnlp.github.io/CoreNLP/

[6]https://spacy.io/

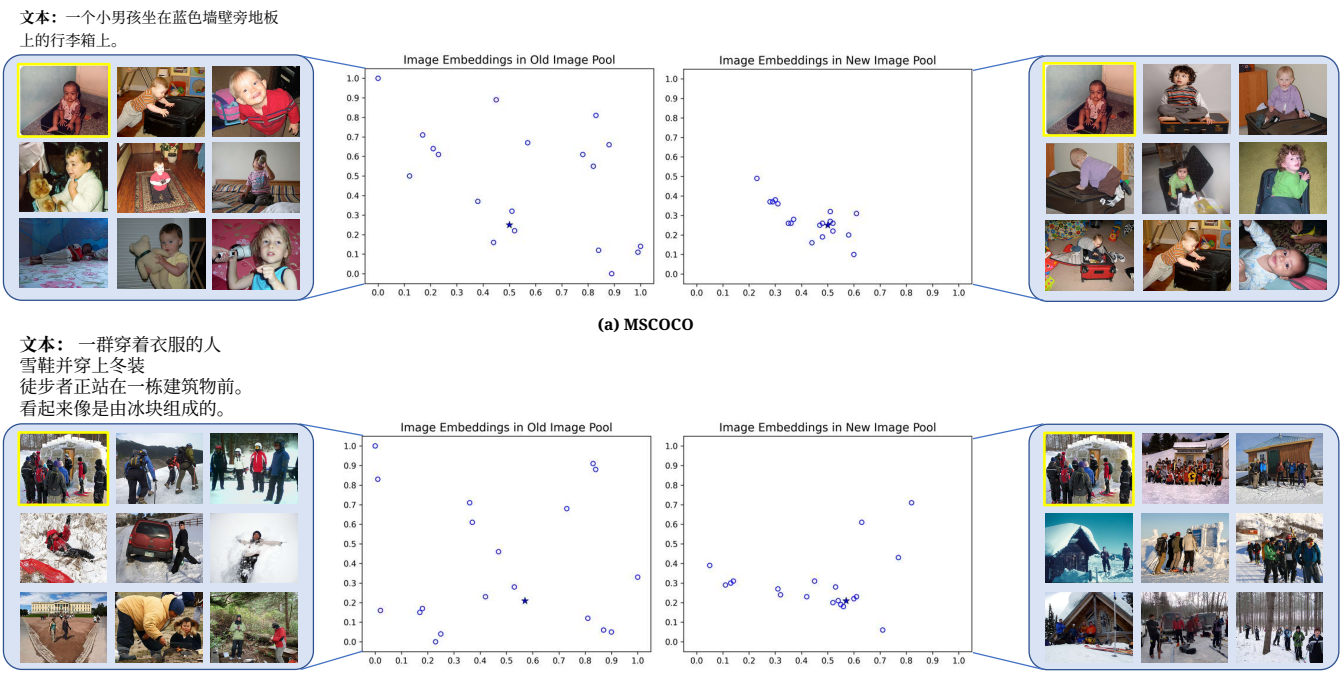

图5：MSCOCO和Flickr30K上原始检索图像池与我们构建的新检索图像池示意图。我们使用 t-SNE 分别可视化旧图像池和新图像池中每张目标图像(⋆)及其前20张最相似图像(○)的 CLIP特征。我们还展示了目标图像（以黄色边框标出）及其前8张相似图像。

站在人行道上的女性”和新的细节“现在是夜晚”一同输入 GPT-2，并生成更细粒度的文本“一位女性夜晚站在人行道上。”

(4) 我们再次使用 CLIPScore 筛选结果，并剔除得分低于原始文本的结果。对于包含多个新文本的情况，我们选择 CLIPScore 最高的文本 d_{best} 作为最终改写文本。

3.2.4 人工校正。即使经过严格筛选，一些新文本仍然存在错误，通常是语法错误和语义错误。我们随机抽取了100条新文本，发现约40%存在一些轻微错误。考虑到这些错误通常很容易修正，我们招募了20名大学生来人工检查并纠正这些错误。对于每个包含图像和生成文本的案例，我们要求他们检查生成文本是否准确描述了图像，并提供原始文本，帮助他们快速找到新增的细节。如果发现错误，他们需要尽可能少用词语进行修正。最终，我们通过上述步骤改写粗粒度文本，得到新的细粒度文本。

4 实验

我们通过改进图像池和文本，对 MSCOCO 和 Flickr30K 上现有的基准数据集进行了改造，并构建了相应的新基准数据集，分别命名为 MSCOCO-FG 和 Flickr30K-FG。本节首先介绍改进后的新基准数据集的更多细节，随后在这些新基准数据集上评估几种经典的图文检索模型。此外，我们还将详细分析模型借助新基准数据集理解细粒度语义的能力。

表 2：MSCOCO 和 Flickr30K 上原始图像池与我们新建图像池的规模比较。

图像池	MSCOCO	Flickr30K
原文	5,000	1,000
Our	**31,244**	**6,867**

4.1 实验设置

实现细节。 除非另有说明，我们使用 CLIP$_{ViT-B/32}$ 作为预训练模型，以帮助搜索相似图像并筛选粗粒度文本。我们选择 VinVL$_{base}$ 并使用提示词生成细节。为了将新细节与粗粒度文本合并，我们采用 GPT-2$_{small}$。我们在两个数据集上训练 GPT-2，学习率为 3e-4，批量大小为 16，训练轮数为 10。我们使用 PyTorch 实现了该方法，并在单张 NVIDIA GeForce RTX 1080 Ti GPU 上进行了所有实验。

评估指标。 我们遵循先前的工作 [7, 18, 20]，使用 Recall@K（R@K）评估图文检索性能，该指标定义为能够在前 K 个检索结果中正确检索到真实标注结果的查询所占的百分比。K 的取值为 1、5 和 10。

4.2 新基准数据集的统计情况

我们新建基准数据集的统计信息及改进细节如下所示。

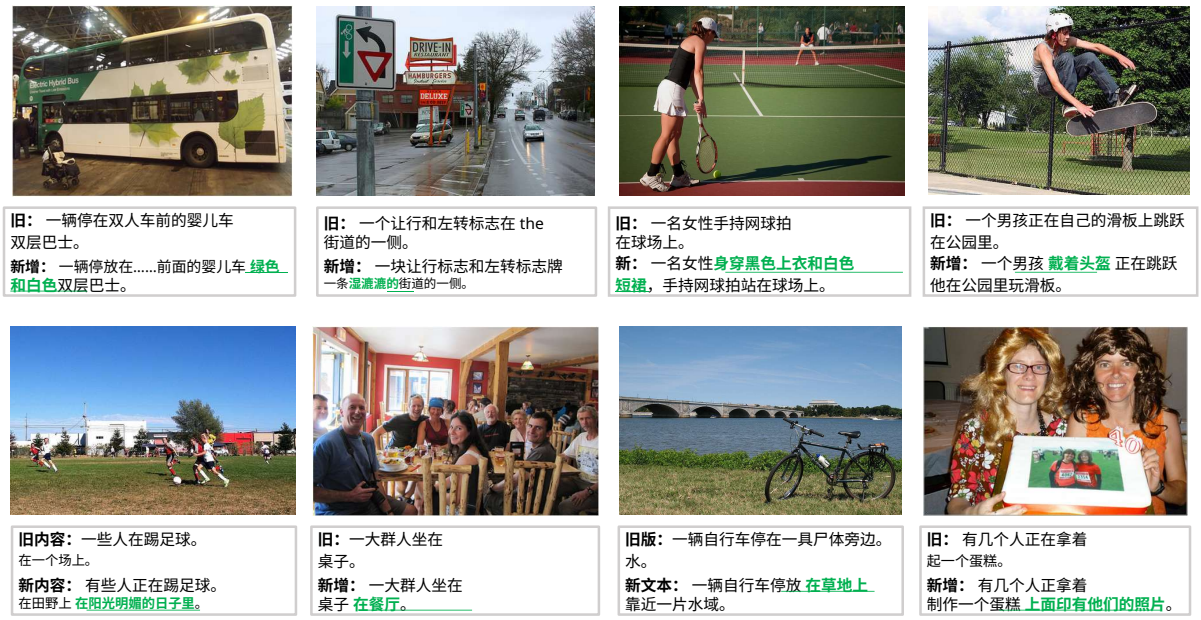

图6：同一图像对应的旧文本和改进后文本示例。 绿色 文字表示通过我们的改进流程补充的细节。

4.2.1 更大且语义密度更高的图像池。 如表2所示，我们新基准数据集中的图像池显著大于旧基准数据集中的图像池，MSCOCO和Flickr30K分别包含31k+ 和6k+ 张图像。为了检验新图像池中图像的语义密度提升情况，我们使用t-SNE [27] 分别将每个目标图像及其在旧图像池和新图像池中的前20张最相似图像的CLIP特征可视化，如图5所示。我们可以看到，与旧图像池中的图像相比，新图像池中的图像在语义上更加接近彼此（即图像池的语义密度更高）。此外，我们还展示了旧图像池和新图像池中与同一目标图像最相似的前8张图像。新图像池包含了与目标图像具有更多相似对象或场景的图像。

4.2.2 更详细、更细粒度的文本。 我们从不同方面比较改进前后的文本质量。我们统计所有文本中的名词和形容词数量、每个句子中名词和形容词的平均数量，以及平均句长，以衡量文本的粒度。表3中的结果表明，在两个数据集上，粗粒度文本经过改进后语义都更加丰富。此外，我们在图6中展示了一些改进后的新文本与原始粗粒度文本的对比示例。我们的方法能够从多个方面改进粗粒度文本，描述图像的更多细节，包括新增属性（“绿色和白色”）、详细外观（“黑色上衣”和“白色裙子”）、天气（“晴天”）、地点（“餐厅”），以及未曾提及的新对象（“一幅画”）等。

4.3 在新基准数据集上的评估

我们在新建的基准数据集上评估了不同的现有图文检索模型，以证明我们的改进能够实现更细粒度评估的有效性。具体而言，我们选择了三种

表3：粗粒度文本与细化文本的统计数据。

(a) MSCOCO

文本类型	#文本	#名词	#Adj	Avg. Noun	Avg. Adj	Avg. 长度
粗粒度	11,050	38,944	6,893	3.52	0.62	10.23
优化后的		**45,075**	**14,176**	**4.08**	**1.28**	**12.59**

(b) Flickr30K

文本类型	#文本	#名词	#Adj	Avg. Noun	Avg. Adj	Avg. 长度
粗粒度	2,156	7,587	1,906	3.52	0.88	10.93
精细		**9,497**	**2,726**	**4.40**	**1.26**	**13.91**

非预训练模型和两个多模态预训练模型。这些模型的更多详细信息如下：

- **VSE++** [7] 使用 CNN 将图像编码为网格特征，并在损失函数中纳入困难负样本。
- **SCAN** [18]首先从图像中检测对象，并关注文本中的词语，以学习潜在的视觉-语义对齐关系。
- **VSRN** [20] 提出结合GCN [16] 以增强局部视觉语义推理，并结合GRU [1] 以执行全局语义推理。
- **UNITER** [2] 提出了一种基于 Transformer 的编码器，通过在四个常见的图文数据集上设计的四种预训练任务，学习联合视觉和文本嵌入表示。
- **X-VLM** [44] 通过多粒度视觉语言预训练，在给定相关文本的情况下定位图像中的视觉概念，并使文本与这些概念对齐，在包括图文检索在内的许多视觉与语言任务上取得了最先进的性能。

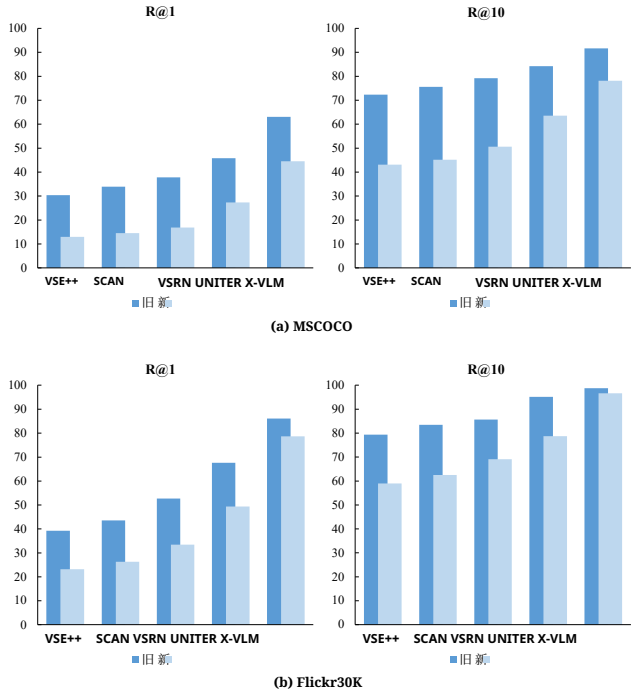

图 7：给定新的精炼文本查询，对比不同模型在旧图像池与新图像池上的文本到图像检索性能。

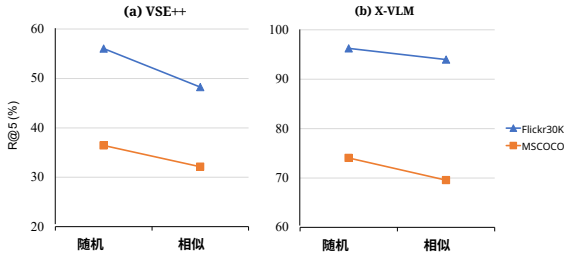

图 8：对比两个模型在含随机图像的扩展图像池上的性能，以及在含相似图像的新图像池上的性能。

对于非预训练模型，我们按照相同的设置在两个基准数据集上训练 VSE++、SCAN 和 VSRN。对于预训练模型，我们在两个基准数据集上微调 UNITER$_{base}$ 和 X-VLM$_{base}$。

4.3.1 不同图像池上的实验。 为验证使用相似图像扩大图像池的有效性，我们在两个基准数据集上进行文本到图像（T2I）检索实验。给定细粒度查询文本，要求模型从旧图像池或新图像池中检索目标图像。图 7 展示了五个模型在旧图像池和新图像池上的 R@1 和 R@10 性能。可以看出，所有模型在新图像池上进行检索时，性能均明显下降。性能下降表明，我们的新图像池更具挑战性，要求模型具备更好的细粒度语义理解能力。

桌子 4：在新的候选池上，原始文本与我们细化后的新文本的 R@1 性能对比。

模型	细化文本	MSCOCO I2T	MSCOCO T2I	Flickr30K I2T	Flickr30K T2I
VSE++	✗	41.26	10.51	52.80	19.80
	✓	**45.44**	**12.93**	**57.20**	**23.18**
SCAN	✗	45.30	11.46	68.10	22.70
	✓	**50.90**	**14.55**	**72.30**	**26.28**
VSRN	✗	50.22	13.73	69.90	29.10
	✓	**55.82**	**16.83**	**75.30**	**33.42**
UNITER	✗	59.08	22.05	80.70	42.86
	✓	**64.82**	**27.32**	**85.00**	**49.38**
X-VLM	✗	80.98	38.77	96.80	73.10
	✓	**84.16**	**44.50**	**97.40**	**78.70**

为进一步验证新图像池中的相似图像对检索评估的影响，我们构建了一个扩展图像池的变体。我们没有使用相似图像进行扩展，而是从同一候选图像源中随机选取图像，将图像池扩展到相同的规模。随后，我们在随机扩展的图像池和新的图像池上进行 T2I 对比实验。图 8 展示了 VSE++ 和 X-VLM 在这两类图像池上的 R@5 性能。我们观察到，对于同一模型，在相似图像池上的性能明显更低，这表明将相似图像引入图像池确实增加了检索难度，要求检索模型具备更好的跨模态细粒度语义理解和匹配能力。

4.3.2 不同粒度文本上的实验. 我们还开展了文本到图像（T2I）检索和图像到文本（I2T）检索实验，以验证改进后文本的有效性。我们比较了同一模型在新的候选池上使用原始文本或我们新生成的细化文本时的检索表现。表 4 的结果显示，无论将新改写的文本用作查询还是候选项，模型的 R@1 表现都更好，这表明改进后的文本粒度更细，从而实现更好的跨模态匹配。

4.4 经典模型的实证研究

上述实验表明，我们的新基准数据集能够更好地评估模型在细粒度语义理解和对齐方面的能力。随后，我们尝试通过在新基准数据集上开展实验，对开篇提出的问题进行更深入的分析并提供见解：当前模型是否真的具备可与人类相媲美的跨模态语义知识，尤其是在细粒度语义理解方面?

如图 7 所示，所有模型在我们的新基准数据集上进行文本到图像检索时，召回率都明显降低；即使是最先进的模型 X-VLM 也出现了显著退化，尤其是在 MSCOCO-FG 上，R@1 仅为 44.5。这些结果表明，**当前的图文检索模型在细粒度跨模态语义理解方面仍远未达到完美。**



图 9：X-VLM 为给定图像选择正确文本的成功或失败示例。

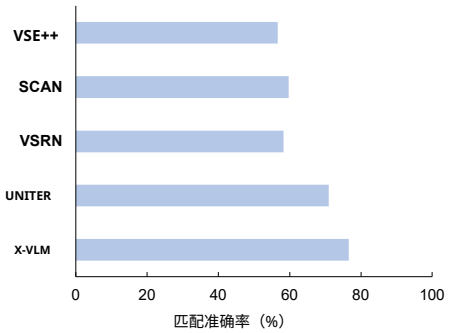

图 10：不同模型在图像与错误/正确文本对匹配任务中的性能比较。

此外，我们注意到，不同模型在新基准数据集上的性能下降程度各不相同。预训练模型（UNITER、X-VLM）在 MSCOCO-FG 和 Flickr30K-FG 上的表现仍远胜于非预训练模型（VSE++、SCAN、VSRN），这证明了预训练模型的优势：通过观察大量图像-文本对，并在预训练过程中执行特定的对齐任务，获取更多跨模态语义知识。不同的非预训练模型和预训练模型之间，表现也各有差异。配备了用于细粒度语义理解和对齐的额外模块的非预训练模型（如 VSRN），表现远胜于未配备此类模块的模型（即 VSE++、SCAN）；而通过特定细粒度预训练任务训练的预训练模型（如 X-VLM），往往比未采用此类任务训练的模型（即 UNITER）表现更好。

尽管当前模型已在提升图像与文本之间的细粒度语义理解方面做出了一些有效尝试，但仍有很大的改进空间。我们进一步开展了一些额外实验，以启发未来在细粒度语义理解方面的研究方向。具体而言，我们利用半自动改进流程中模型生成的错误文本，以及 MSCOCO-FG 中与之对应的经人工修正文本，进一步研究模型在细粒度语义理解方面的能力。根据我们在修正过程中的观察，错误文本与正确文本几乎完全相同，仅有少数词语不同。给定一张图像，以及一对错误文本和正确文本，模型需要为该图像选出最合适的文本，以此考察模型区分并对齐两种模态中细节的能力。图 10 中的结果证实，经过预训练的模型表现更好于未经预训练的模型，在细粒度语义理解方面表现更佳；然而，它们在选择正确文本时仍有一定困难，因为其匹配准确率仍低于 80%。

我们进一步研究了 X-VLM 的匹配成功与失败案例，并在图 9 中展示了一些示例。如图 9(a) 所示，X-VLM 等先进的图文检索模型已展现出在细粒度语义层面理解和区分对象及其属性的良好能力。然而，如图 9(b) 所示，这类模型仍难以区分图像中相邻对象的属性。这表明，未来的图文检索研究可以更加关注如何提升模型对细粒度对象间关系和对象-属性对齐的理解能力。例如，可以考虑将类似的图像与容易混淆的文本（如上文所述）作为困难负样本来训练模型，从而提升其细粒度语义理解能力。

5 结论

在这项工作中，我们回顾了图文检索的常用基准数据集，并指出由于图像和文本的粒度较粗，这些数据集不足以评估细粒度跨模态语义对齐的真实性能。因此，我们基于原有基准数据集构建了改进的新基准数据集 MSCOCO-FG 和 Flickr30K-FG。具体而言，我们通过纳入更多相似图像来扩大原始图像池，并提出了一套半自动化改造流程，只需少量人工投入，就能将粗粒度文本改进为更细粒度的文本。我们通过在新基准数据集上评估具有代表性的图文检索模型，展示了这套改造流程的有效性；同时，通过大量实验分析了模型的细粒度语义理解能力。结果表明，即使是最先进模型，在细粒度语义理解方面仍有很大的提升空间，尤其是在区分图像中相近对象的属性和关系方面。

值得注意的是，我们的改进流程可以利用不同的预训练语言模型和视觉-语言模型，通过集成来改进基准数据集。由于视觉-语言模型在提示生成新细节方面存在局限，本研究仅提供了基于特定模型的基础版本，这可能会在图像选择和文本改进方面引入偏差。随着大型预训练模型不断发展，我们可以进一步利用这些模型，并相应地改进基准数据集。我们希望这些新基准数据集能激发对跨模态检索的进一步深入研究。

致谢

本研究部分受到国家重点研发计划（编号：2020AAA0108600）和中国国家自然科学基金（编号：62072462）的资助。

参考文献

[1] Amulya Arun Ballakur 和 Arti Arya。2020。航空延误预测中门控循环神经网络架构的实证评估。载于 *2020*年第*5*届国际计算、通信与安全会议（*ICCCS*）。1–7。[2] Yen-Chun Chen、Linjie Li、Licheng Yu、Ahmed El Kholy、Faisal Ahmed、Zhe Gan、Yu Cheng 和 Jingjing Liu。2020。Uniter：通用图像文本表征学习。载于 欧洲计算机视觉会议。104–120。[3] Karan Desai、Gaurav Kaul、Zubin Aysola 和 Justin Johnson。2021。RedCaps：由大众为大众创建的网络策展图像文本数据。载于神经信息处理系统数据集与基准数据集专题论文集。[4] Haiwen Diao、Ying Zhang、Lin Ma 和 Huchuan Lu。2021。图像-文本匹配的相似性推理与过滤。载于 *AAAI*人工智能会议论文集（第*35*卷第*2*期：*AAAI-21*技术分会场*2*）。1218–1226。[5] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xi- aohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。2021。一张图像胜过16×16个词：用于大规模图像识别的Transformer。载于 第九届国际学习表征会议（*ICLR*）。[6] Zi-Yi Dou、Yichong Xu、Zhe Gan、Jianfeng Wang、Shuohang Wang、Lijuan Wang、Chenguang Zhu、Pengchuan Zhang、Lu Yuan、Nanyun Peng、Zicheng Liu 和 Michael Zeng。2022。端到端视觉语言Transformer训练的实证研究。载于 *2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。18145–18155。[7] Fartash Faghri、David J. Fleet、Jamie Ryan Kiros 和 Sanja Fidler。2018。VSE++：利用困难负样本改进视觉语义嵌入。载于 英国机器视觉会议论文集（*BMVC*）。[8] Andrea Frome、Greg S. Corrado、Jonathon Shlens、Samy Bengio、Jeffrey Dean、Marc'Aurelio Ranzato 和 Tomas Mikolov。2013。DeViSE：深度视觉语义嵌入模型。载于 第*26*届神经信息处理系统国际会议论文集——第*2*卷。2121–2129。[9] Zhe Gan、Yen-Chun Chen、Linjie Li、Chen Zhu、Yu Cheng 和 Jingjing Liu。2020。用于视觉语言表征学习的大规模对抗训练。载于 第*34*届神经信息处理系统国际会议论文集。6616–6628。[10] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016。用于图像识别的深度残差学习。载于 *2016*年*IEEE*计算机视觉与模式识别会议（*CVPR*）。770–778。[11] Jack Hessel、Ari Holtzman、Maxwell Forbes、Ronan Le Bras 和 Yejin Choi。2021。CLIPScore：一种无需参考的图像描述生成评估指标。载于*2021*年自然语言处理实证方法会议论文集。7514–7528。[12] Micah Hodosh、Peter Young 和 Julia Hockenmaier。2015。将图像描述构建为排序任务：数据、模型与评估指标。载于 第*24*届人工智能国际会议论文集。4188–4192。[13] Zhicheng Huang、Zhaoyang Zeng、Yupan Huang、Bei Liu、Dongmei Fu 和 Jianlong Fu。2021。跳出框框看世界：用于视觉语言表征学习的端到端预训练。载于 *2021*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。12971–12980。[14] Justin Johnson、Ranjay Krishna、Michael Stark、Li-Jia Li、David A. Shamma、Michael S. Bernstein 和 Li Fei-Fei。2015。使用场景图进行图像检索。载于 *2015*年*IEEE*计算机视觉与模式识别会议（*CVPR*）。3668–3678。[15] Wonjae Kim、Bokyung Son 和 Ildoo Kim。2021。ViLT：无需卷积或区域监督的视觉语言Transformer。载于 第*38*届国际机器学习会议。5583–5594。[16] Thomas N. Kipf 和 Max Welling。2017。基于图卷积网络的半监督分类。载于国际学习表征会议（*ICLR*）。[17] Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalantidis、Li-Jia Li、David A. Shamma、Michael S. Bernstein 和 Li Fei-Fei。2017。Visual Genome：利用众包密集图像标注连接语言与视觉。国际计算机视觉期刊 123，1（2017），32–73。[18] Kuang-Huei Lee、Xi Chen、Gang Hua、Houdong Hu 和 Xiaodong He。2018。用于图像-文本匹配的堆叠交叉注意力。载于 欧洲计算机视觉会议。Springer，212–218。[19] Gen Li、Nan Duan、Yuejian Fang、Ming Gong 和 Daxin Jiang。2020。Unicoder-VL：通过跨模态预训练实现视觉与语言的通用编码器。载于 *AAAI*人工智能会议论文集。11336–11344。[20] Kunpeng Li、Yulun Zhang、Kai Li、Yuanyuan Li 和 Yun Fu。2019。用于图像-文本匹配的视觉语义推理。载于《2019年IEEE/CVF国际计算机视觉会议（ICCV）》。4653–4661。[21] Wei Li、Can Gao、Guocheng Niu、Xinyan Xiao、Hao Liu、Jiachen Liu、Hua Wu 和 Haifeng Wang。2021。UNIMO：通过跨模态对比学习实现统一模态理解与生成。载于《计算语言学协会第59届年会暨第11届自然语言处理国际联合会议论文集》（第1卷：长篇论文）。2592–2607。[22] Xiujun Li、Xi Yin、Chunyuan Li、Pengchuan Zhang、Xiaowei Hu、Lei Zhang、Lijuan Wang、Houdong Hu、Li Dong、Furu Wei、Yejin Choi 和 Jianfeng Gao。2020。Oscar：面向视觉语言任务的对象-语义对齐预训练。载于《欧洲计算机视觉会议》。121–137。[23] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。2014。Microsoft COCO：情境中的常见对象。载于《欧洲计算机视觉会议》。740–755。[24] Chunxiao Liu、Zhendong Mao、An-An Liu、Tianzhu Zhang、Bin Wang 和 Yong- dong Zhang。2019。聚焦注意力：用于图像-文本匹配的双向焦点注意力网络。载于《第27届ACM国际多媒体会议论文集》。3–11。[25] Chunxiao Liu、Zhendong Mao、Tianzhu Zhang、Hongtao Xie、Bin Wang 和 Yongdong Zhang。2020。用于图像-文本匹配的图结构网络。载于《2020年IEEE/CVF计算机视觉与模式识别会议（CVPR）》。10918–10927。[26] Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。2019。ViLBERT：面向视觉与语言任务的任务无关视觉语言表征预训练。载于《第33届神经信息处理系统国际会议论文集》。13–23。[27] Laurens van der Maaten 和 Geoffrey Hinton。2008。使用 t-SNE 可视化数据。《机器学习研究期刊》9，86（2008），2579–2605。[28] Hyeonseob Nam、Jung-Woo Ha 和 Jeonghee Kim。2017。用于多模态推理与匹配的双重注意力网络。载于《2017年IEEE计算机视觉与模式识别会议（CVPR）》。2156–2164。[29] Vicente Ordonez、Girish Kulkarni 和 Ta mara L Berg。2011。Im2Text：使用100万张带说明文字的照片描述图像。载于《第24届神经信息处理系统国际会议论文集》。1143–1151。[30] Leigang Qu、Meng Liu、Da Cao、Liqiang Nie 和 Qi Tian。2020。用于图像-文本匹配的上下文感知多视图摘要网络。载于《第28届ACM国际多媒体会议论文集》。1047–1055。[31] Leigang Qu、Meng Liu、Jianlong Wu、Zan Gao 和 Liqiang Nie。2021。用于图文检索的动态模态交互建模。载于《第44届ACM信息检索研究与开发国际会议论文集》。1104–1113。[32] 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。通过自然语言监督学习可迁移的视觉模型。载于《第38届国际机器学习会议论文集》。8748–8763。[33] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019。语言模型是无监督的多任务学习者。https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf，最后访问日期：2023-1-1。[34] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。2017。Faster R-CNN：借助区域提议网络实现实时对象检测。《IEEE模式分析与机器智能汇刊》39，6（2017），1137–1149。[35] Piyush Sharma、Nan Ding、Sebastian Goodman 和 Radu Soricut。2018。概念说明文字：一个经过清理并采用上位词处理的图像替代文本数据集，用于自动图像描述生成。载于《计算语言学协会第56届年会论文集》（第1卷：长篇论文）。2556–2565。[36] Botian Shi、Lei Ji、Pan Lu、Zhendong Niu 和 Nan Duan。2019。用于图像-文本匹配的知识感知语义概念扩展。载于《第二十八届人工智能国际联合会议论文集，IJCAI-19》。5182–5189。[37] Bart Thomee、David A. Shamma、Gerald Friedland、Benjamin Elizalde、Karl Ni、Douglas Poland、Damian Borth 和 Li-Jia Li。2016。YFCC100M：多媒体研究中的新数据。《ACM通讯》59，2（2016），64–73。[38] Sijin Wang、Ruiping Wang、Ziwei Yao、Shiguang Shan 和 Xilin Chen。2020。用于关系感知图文检索的跨模态场景图匹配。载于《2020年IEEE冬季计算机视觉应用会议（WACV）》。1497–1506。[39] Yaxiong Wang、Hao Yang、Xueming Qian、Lin Ma、Jing Lu、Biao Li 和 Xin Fan。2019。用于图像-文本匹配的位置聚焦注意力网络。载于《第二十八届人工智能国际联合会议论文集》。3792–3798。[40] Zihao Wang、Xihui Liu、Hongsheng Li、Lu Sheng、Junjie Yan、Xiaogang Wang 和 Jing Shao。2019。CAMP：用于文本-图像检索的跨模态自适应消息传递。载于《2019年IEEE/CVF国际计算机视觉会议（ICCV）》。5763–5772。

[41] Yiling Wu、Shuhui Wang、Guoli Song 和 Qingming Huang。2019。学习用于图像-文本匹配的片段自注意力嵌入。载于《第*27*届*ACM*国际多媒体会议论文集》。2088–2096。[42] Peter Young、Alice Lai、Micah Hodosh 和 Julia Hockenmaier。2014。从图像描述到视觉指称：用于事件描述语义推理的新相似度度量。《计算语言学协会会刊》2（2014），67–78。[43] Fei Yu、Jiji Tang、Weichong Yin、Yu Sun、Hao Tian、Hua Wu 和 Haifeng Wang。2021。ERNIE-ViL：通过场景图增强的视觉语言表征。载于《*AAAI*人工智能会议论文集》（第*35*卷第*4*期：*AAAI-21*技术专题*4*）。3208–3216。

[44] Yan Zeng、Xinsong Zhang 和 Hang Li。2022。多粒度视觉-语言预训练：将文本与视觉概念对齐。发表于第*39*届国际机器学习会议论文集。25994–26009。[45] Kun Zhang、Zhendong Mao、Quan Wang 和 Yongdong Zhang。2022。用于图像-文本匹配的负样本感知注意力框架。发表于*2022*年 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）。15640–15649。[46] Pengchuan Zhang、Xiujun Li、Xiaowei Hu、Jianwei Yang、Lei Zhang、Lijuan Wang、Yejin Choi 和 Jianfeng Gao。2021。VinVL：重新审视视觉-语言模型中的视觉表征。发表于*2021*年 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）。5575–5584。