arXiv:2304.01961v1 [cs.IR] 4 Apr 2023

ATOMIC：支持多媒体内容创作的图像/文本检索测试集

Jheng-Hong Yang 滑铁卢大学 加拿大
Carlos Lassance Naver Labs Europe，法国
Rafael S. Rezende Naver Labs Europe，法国
Krishna Srinivasan Google Research，美国
Miriam Redi 维基媒体基金会 英国
Stéphane Clinchant, Naver Labs Europe，法国
Jimmy Lin，滑铁卢大学，加拿大

摘要

本文介绍了 ATOMIC（**A**uthoring **To**ols for **M**ulti**m**edia**C**ontent）数据集，旨在推动图像/文本跨模态检索研究。尽管视觉-语言预训练变换器显著提升了检索效果，现有研究仍依赖图像-标题数据集，这类数据集仅包含简单的图像—文本关系，且对检索任务的用户模型描述不足。为弥合这些过度简化的设置与多媒体内容创作现实应用之间的差距，我们提出一种构建检索测试集的新方法。我们利用层级结构、多样化的文本领域、风格和图像类型，以及维基百科中包含的大规模图像-文档关联。我们基于现实的用户模型制定了两项任务，并通过使用基线模型开展检索实验来验证数据集。ATOMIC 为可扩展、多样化且可复现的多媒体检索研究提供了测试平台。最后，我们的数据集为 2023年文本检索大会（TREC）专门赛道奠定了基础，并已公开发布于 https://github.com/TREC-ATOMIC/ATOMIC。

1 引言

多媒体内容创作需要理解以不同模态编码的元素之间的联系。视觉元素（如照片、图形和示意图）常用于补充文本信息，以实现装饰、补充或改变内容含义等不同目的 [21, 28]。然而，近期有关多媒体信息检索的研究通常依赖于假设图像与文本之间关系简单的数据集。

自对比语言—图像预训练（CLIP）取得巨大成功以来 [32],，大多数近期关于视觉—语言预训练 Transformer 的研究都使用图像描述数据集进行检索基准测试 [22, 23, 41, 48, 50],，例如 COCO [24] 或 Flickr30k [49]。通常，这些数据集中的文本片段会以笼统且直白的方式描述图像内容，例如取自 COCO 的标题“一名男子骑着自行车，行驶在一列火车旁边”；而多媒体内容创作则需要对上下文有更丰富、更准确的理解，例如“Larry 在唐宁街10号门口，他的官方住所和工作场所”，其中 Larry 实际上指的是一只猫。[1] 为了推动该领域的发展，研究界需要能够捕捉图像与文本在自然场景中各种关联方式的数据集。我们介绍<sty

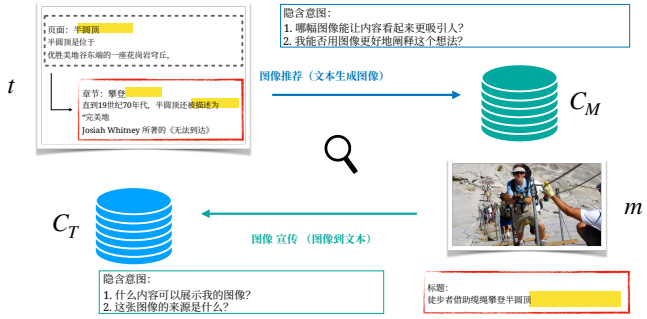


图1：我们的两项检索任务：图像推荐和图像推广。我们使用维基百科中现有的图像-文本关联作为相关性标注的代理。

工具，用于研究维基百科语境下图像与文本之间这一差距的**多媒体内容**（ATOMIC）数据集。

图像与文本之间的联系错综复杂，且用途各异。近期研究提供了多种图像-文本集合，以探究作者将图像与文本关联起来的时间和原因，以及其中微妙的意图。例如，Vempala和Preoţiuc-Pietro [45] 确定了Twitter推文中文本与图像之间信息重叠的四种类别，而Kruk等人[20] 发现，Instagram帖文背后的意图可能更加复杂，例如表达反讽。Lommatzsch等人 [26] 强调了图像与其相关联的新闻报道之间存在描绘差距：图像往往只与报道所述事件略有联系。关于图像—文本关系的研究日益增多，这凸显了需要更多资源来推动我们对这一领域的理解。

ATOMIC 数据集旨在满足一项真实的用户需求：协助作者创作多媒体内容，提升以文本为主的网页的吸引力，例如旅游博客文章 [1]或维基百科页面。[2] 作者通常会通过挑选合适的图像（或视频等其他多媒体内容）来补充文本 [4, 12, 34, 47]。此外，多媒体内容的信息质量往往与所附图像有关 [43]。然而，这一编辑过程既繁琐又耗时。同样，作者可能不清楚图像应该呈现什么内容（即不了解相关的关键词、概念或主题），因此可能需要借助系统解读文本并推荐优质的图像

[1]https://en.wikipedia.org/wiki/Larry_(猫)

[2]相关维基媒体倡议的短链接：https://tinyurl.com/3dasup3r

候选项。开发能够高效访问多媒体资源库的辅助工具，有助于更有效地创作和迭代内容。

为满足上述对图像/文本检索任务所需的多样且贴近现实的测试集的需求，我们推出了 AToMiC 数据集。该数据集基于维基百科的文本和图像集合，以及维基百科编辑整理的二者关联。AToMiC 数据集为研究多媒体百科文档中文本和图像的自然共现提供了宝贵的评估资源、基线和社区。我们展示了相关实验，分析并证明了 AToMiC 数据集在现实场景中的潜力，同时也指出了其局限性，例如标签偏差和流行度偏差。具体而言，我们定义了两项检索任务，并使用公开可用的基于 Transformer 的模型进行了初步评估。AToMiC 数据集已公开发布，并将在 2023 年文本检索会议（TREC）设立的专门赛道中亮相。

总之，我们的贡献如下：
（1）我们根据真实用户需求定义了两项多媒体检索任务：图像推荐和图像推广。（2）我们提供了一个易于使用的数据集，以支持这些任务。该数据集可免费获取，并将用于 TREC 2023 年评测，同时还将提供一个标注更丰富的测试集。（3）我们开展了初步实验和分析，将其与广泛使用的 COCO 数据集进行比较，并提供基线，以展示 AToMiC 在所提任务中的能力。

2 相关工作

我们首先概述其他图像/文本评估资源，并说明 AToMiC 在这一更广泛的上下文中的定位：

ImageCLEF： *2012* ~ 如今。 </style}多年来，ImageCLEF 挑战赛一直在维基百科上组织多模态检索任务 [29, 30, 44], ，最近一次是在 2011 年。此后，ImageCLEF 的研究范围转向医学、社交媒体和自然应用等特定领域 [17]。我们的数据集也使用维基百科，但重点是为能够处理多种领域数据的系统开发通用基准。

2012 年之前的 *INEX* 和 *ImageCLEF*。 我们提出的评估与 INEX即席 赛道 [11] 以及 ImageCLEF 主办的维基百科挑战赛 [31]有关。这些评估以维基百科为资源构建测试集，并设计任务，使其紧密遵循 TREC即席 检索任务的“经典”形式。INEX 赛道致力于开发能够处理结构化文档的 XML 检索系统，而维基百科挑战赛则侧重于搜索图像，查询由多种模态构成。AToMiC 提供了一个与时俱进且规模更大的测试集，融合了这些早期举措的理念，即利用维基百科中的文档结构，并开发能够处理多种模态的系统。

FIRE：自动故事插图。由 FIRE 研讨会举办的自动故事插图（ASI）赛道旨在为图像文档扩展任务构建一个测试集，即通过检索与文本片段相关的图像来为故事配图 [13]。ASI 赛道也使用维基百科作为其图像和文本集合。不过，它只关注少数几个主题，内容涉及儿童童话。此外，这个测试集目前已不再公开提供。

图像-标题数据集。 继 Karpathy 和 Li [18], 的工作之后，近期关于预训练视觉-语言 Transformer 的研究通常依赖现成的图像描述数据集，例如 COCO [24] 或 Flickr30k [49]，来评测模型 [5, 36] 在检索任务中的表现。这些数据集包含复杂图像，描绘了丰富背景中的多个物体或动作。它们不同于以单一特定类别物体为焦点的数据集，例如 CUB Captions [46], 中的鸟类、Fashion200 [15], 中的服装，或 CheXpert [52] 中的医学图像。尽管图像描述数据集涵盖丰富的视觉领域，其标题往往只是对图像中物体或动作及其属性的单句描述。相比之下，我们的数据集提供了丰富的文本集合，为图像提供外部上下文。

维基百科图像-说明文字数据集。 AToMiC 基于 WIT 数据集 [42] 构建。AToMiC 与 WIT 的主要区别在于评测设置和文本集合的范围。WIT 数据集提供可公开获取的图像（URL）-文本元组，可用于各种任务。例如，维基媒体基金会主办的 Kaggle 挑战赛侧重于图像-说明文字匹配，[3] 而 Kreiss 等人 [19] 则优先研究图像描述生成。相比之下，我们的目标是构建一个支持多媒体内容创作的检索数据集。因此，我们关注图像与文档（章节）之间的关联，并围绕 即席检索任务构建 AToMiC。AToMiC 与 WIT 之间的关系总结如下：

- **图文关联**。 我们聚焦于章节级别的图文关联，并遵循 WIT 的训练集/验证集/测试集划分（见第 4.1 节）。
- **文本**。 我们抓取英文维基百科条目，并选取未关联图像的章节，从而构建更贴近实际的场景，因为并非每个文本片段都应该与图像关联。更多详情见第 4.2 节。
- **图像**。 我们沿用 WIT 中的图像及其元数据，但以统一格式提供图像像素值。更多详情见第 4.3 节。

3 任务表述

在标准的 即席 检索中，我们假设存在一个语料库 C，由一组文档 $\{d_1, d_2 \ldots d_n\}$ 组成。对于以查询 q 表示的用户信息需求，系统的目标是返回文档的前 k 个排序列表，并使其在 nDCG、平均倒数排名或 MAP 等质量指标上达到最优。

AToMiC 作为一个图像/文本多模态检索测试集，其设置略有不同。它由两个独立的集合组成：文本集合 $\mathbf{C}_T = \{t_1, t_2 \ldots t_n\}$（$t$表示文本）和图像集合 $\mathbf{C}_M = \{m_1, m_2 \ldots m_n\}$（$m$ 表示媒体）。图1展示了一个来自 $\mathbf{C}_T$ 的示例 t，以及一个来自 $\mathbf{C}_M$ 的示例 m。每个 m 都由图像本身（即像素值）以及元数据（若有）组成，其中可能包含描述图像的标题。值得注意的是，维基百科的标题旨在关联图像及其所属章节，如 Kreiss 等人 [19] 的研究和风格指南所示。[4] 因此，我们预期

[3] https://www.kaggle.com/c/wikipedia-image-caption [4] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Captions

这些标题在语义上与其关联的文档相关。在下文定义的两项任务中，我们利用已知的图文关联作为标签。

图像推荐任务（T2M）。信息需求（为方便起见，称为查询），记作 q，就是从 C_T 中抽取的一段文本 t，即 $q \in C_T$。也就是说，维基百科的一位编辑查看文章中的某个特定章节，并希望添加一张合适的图像。给定 q，系统的任务是从 C_M 中返回一个包含排名前 k 位图像的列表，使某项质量指标最大化。在此上下文中，相关性被定义为“这张图像适合附加到该章节”。因此，对于由文本 t 表示、涉及半圆顶热门活动的查询，图1中人们攀登山峰的图像 m 就是相关的。请注意，在本例中，这张图像已关联到维基百科“半圆顶”条目中的“攀登”章节。

图像推广任务（M2T）。这项任务与图像推荐任务相反：查询 q 是来自 C_M 的一张图像，而待检索的集合是 C_T。目标是返回一个排名前 k 位的文本列表，其中相关性被定义为“这是维基百科文章中适合附加这张图像的一个章节”。虽然这不是我们工作的重点，但为此任务构建的模型可以支持创作工具，为图像撰写说明文字。例如，用户可以写下标题：“准备徒步攀登半圆顶”，使其与图1所示的文本章节相关；也可以写下“登山者沿着山脊前行”，使这张图像与维基百科上的登山文章相关。[5] 图像推广任务要求检索系统提供相关上下文，以支持描述性图像说明生成 [25] 和开放知识视觉问答 [27, 40]等下游任务。

4 数据集构建

本章节介绍 AToMiC，这是一个旨在支持多媒体内容创作领域图像/文本跨模态检索研究的数据集。该数据集的开发基于以下目标：(1) 确保自然图像与文本之间存在多样化的关联；(2) 提供可公开获取且采用标准化评估流程的数据集；(3) 涵盖广泛的模态内多样性。现有数据集主要聚焦于简短的文本字幕和面向对象的图像，而现实世界中的文本涵盖多个字段、层级关系和不同长度，图像-标题数据集则往往只考虑照片等特定风格；(4) 提供充足的数据，用于训练和评估先进的神经网络。

4.1 相关性判断

基于 WIT 构建 AToMiC，我们利用基于维基百科的图像文本（WIT）数据集 [42],[6]。该数据集包含3760万个条目，每个条目由图像、文本和上下文数据组成，这些数据提取自279种语言的维基百科页面，最近一次整理约在2020年8月30日。上下文信息包括章节和页面标题及描述，以及每张图像的字幕、替代文本和署名说明。我们使用与

[5]https://en.wikipedia.org/wiki/Mountaineering [6] https://github.com/google-research-datasets/wit

数据划分	训练集	验证集	Test	其他
# 文本	3,002,458	17,173	9,873	7,105,240
# 图像	3,386,183	16,131	8,605	7,608,283
# 相关性判定集	4,401,903	17,801	9,873	-
# M/T	1.47 (± 2.72)	1.03 (± 0.43)	1.00 (± 0.00)	-
# T/M	1.30 (± 0.82)	1.10 (± 0.43)	1.15 (± 0.53)	-

表 1：AToMiC 数据集统计信息。文本/图像数量根据相关性标签定义。# M/T（T/M）表示每条文本对应的相关图像数量（每幅图像对应的文本数量）。

章节级图像-文本关联相对应的 WIT 条目，其中图像与页面的某个章节相关联。为简化评估，我们仅关注英文维基百科中的关联。

定义稀疏相关性标注。我们数据集中的相关性标注（按 TREC 术语称为 qrels）基于英文维基百科中的图像–章节关联。如图1所示，在“半圆顶”维基百科页面的示例片段中，相关图文对由一个章节及其关联图像组成。相比之下，我们将所有其他图文对定义为不相关，即使这可能导致假负例（参见第 6 节）。我们从 WIT 中提取这些图文对，构建 AToMiC 相关性判定集，并与 WIT 数据集使用的训练集、验证集和测试集划分保持一致。[7] 这些图文对（相关性判定集）还经过以下三步筛选：(1) 删除图像网址已无法在互联网上访问的关联。(2) 如果重复文本的章节标题、页面标题和描述相同，则将其重新映射到同一 ID。由于 WIT 图像–文本元组以图像为中心，文本可能重复，因此这一步是必要的。(3) 如果某个图文对中的图像或文本同时出现在训练集、验证集和测试集中，则将其删除，以确保各数据集之间没有重叠。这样做是为了模拟搜索未知图像或文本的任务。

统计信息。 表 1 根据上述相关性判定集展示了 AToMiC 数据集的统计信息。请注意，各数据划分中的“查询”数量分别列于训练集、验证集和测试集对应的列中。“其他”列表示未纳入这三个数据划分的文本和图像。在进行图像推荐和图像推广任务时，索引集合和检索集合应分别对应第 4.4 节所定义三种设置中的文本总数 $|C_T|$ 或图像总数 $|C_M|$。请注意，表 1 所示的 M/T 和 T/M 值体现了图像–章节关联的多对多特性。

4.2 文本集合

我们构建 AToMiC 数据集，旨在形成一种能够更真实地反映现实世界中文档多样性的数据分布。为此，我们采用英文维基百科页面集合作为文本数据来源。维基百科旨在涵盖人类知识的全部内容，因此其页面主题广泛、内容多样，包括传记、政治、体育、教育、地理等。

文本集合扩展。 考虑到编辑方面的因素，并非每段文本都需要配有图像，因此我们扩展

[7]https://github.com/google-research-datasets/wit/blob/main/DATA.md

字段	空值率	CLIP	BERT
		AToMiC-Texts	
标题（P）	0%	4.79 (± 2.52)	4.68 (± 2.23)
标题（S）	0.02%	2.53 (± 3.17)	2.32 (± 2.78)
层级	0%	2.30 (± 8.05)	2.15 (± 7.71)
上下文（P）	19.21%	146.15 (± 159.54)	137.81 (± 150.99)
上下文（S）	0%	195.93 (± 312.87)	184.68 (± 294.76)
		AToMiC-Images	
参考	42.28%	17.46 (± 23.65)	13.98 (± 17.48)
替代文本	74.79%	6.67 (± 10.85)	5.90 (± 8.90)
属性	2.57%	47.69 (± 112.87)	41.63 (± 90.42)
		COCO-Captions	
标题	0%	11.54 (± 2.70)	11.81（± 2.81）

表 2：词元数量统计。我们使用 CLIP 和 BERT（uncased）的分词器。统计词元数量时，不计入特殊词元和空实例。

文本集合，使其同时包含配有图像和未配有图像的文本，而 WIT 数据集未涵盖后者。为此，我们采用 Zhang 等人 [51] 工作中使用的 BM25 匹配方法，通过将 WIT 数据集映射到最新的维基百科文本集合来执行投影步骤，从而扩展 WIT 数据集。为此，我们使用最新的 XML 转储解析维基百科，并为检索准备其 Lucene 索引，以便构建 AToMiC 数据集。[8] 随后，我们将 WIT 中的文本（约 300 万个实例）用作查询，在最新的维基百科文本集合中进行搜索，并使用 Anserini 工具包执行 BM25 匹配，参数为（k_1，b） =（0。9，0。4）。最后，我们选取排名第一的章节及同一维基百科页面中的相关章节作为候选项。因此，投影后的文本集合在训练集、验证集、测试集和其他（无相关性标注）集合中共包含约 1000 万个候选项，如表 1 所示。这种方法使我们能够纳入种类多样的文本，无论其是否配有图像，并扩大数据集的范围，使其更能反映现实场景。

词元统计。 在本章节中，我们考察五个字段中关键文本输入的统计情况：页面（章节及其层级）标题，记为标题（P）、标题（S）和层级；以及页面（章节）上下文描述，记为上下文（P）和上下文（S）。由于我们主要考虑预训练 Transformer，例如 CLIP 或 BERT，因此文本输入需要进行分词，即映射为它们接受作为输入的单词/子词。这一点很重要，因为某些模型的最大词元长度有限，例如 CLIP 仅接受 77 个词元，超出部分将被截断。词元统计结果见表 2。该表显示

不同字段中的词元数量差异很大。上下文（S）和上下文（P）等描述性字段包含的词元最多，CLIP 和 BERT 中均约为 140 个，标准差约为 300，这可能会给 CLIP 带来问题。相比之下，标题（P）、标题（S）和层级等抽象性字段仅包含约 10 个词元，且标准差较小，其统计特征与 COCO 更为相似。

维基百科页面类别。为展示文本集合的多样性，我们采用分类器从

[8]我们从 enwiki-20221101 XML 转储中解析文本。

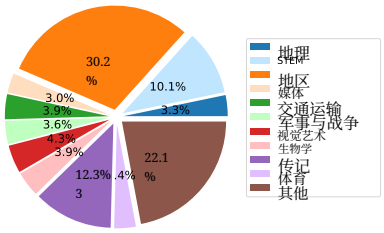


图 2：维基百科页面类别组成。

维基媒体基金会。[9] 我们的文本集合所涵盖的英文维基百科页面的类别构成见图 2。共有 34 个预定义类别，我们将占比低于 2.5% 的类别在图中合并为“其他”。从图中可以看出，维基百科语料库主要涉及地区（30.2%）、体育（12.3%）和科学、技术、工程与数学（10.1%）类条目。其他类别的分布则更为均衡：每类占 3~4%。此外，其他类别还呈现长尾分布，合并后的类别总占比为 22.1%。

4.3 图像集

我们的图像集利用维基百科上高质量、大规模的图像资源构建而成。该资源涵盖了照片之外的多种图像风格，例如艺术作品、地图、射线影像等。与现有的图像-标题数据集相比，它是一个多样且贴近现实的资源。

图像集准备。我们从 WIT 数据集中提取所有图像 URL，并使用 img2dataset 工具[10] 抓取图像。无效的 URL 会被筛除，图像以 WebP 格式存储，质量参数设为 100。我们的图像集包含 WIT 数据集中的所有图像，不受语言限制；对于我们的任务而言，下载维基共享资源数据库中的所有图像可能过于激进。

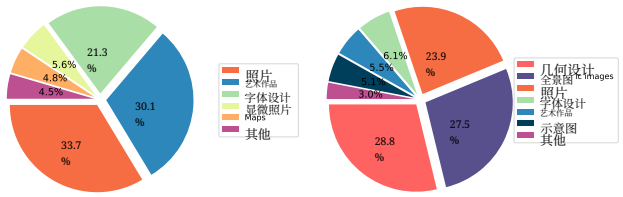
除了图像之外，我们还使用 WIT 数据集中的其他元数据，包括参考字段中的上下文化图像说明、替代文本和署名字段。表 2 展示了这些字段的词元统计。将这些图像元数据添加到文章后，可以使图像与其所属章节的上下文更加吻合。研究人员可以利用这些文本数据，使用文本匹配模型构建参考系统。不过，值得注意的是，这些图像说明的空值率相当高：参考字段约为 42%，替代文本字段约为 75%。图像说明的高空值率表明，依赖这些特征的匹配系统可能泛化能力有限，因为这些特征并非总是可用，尤其是在图像尚未关联到章节的现实场景中。

图像风格。我们使用 CLIP 对 AToMiC 数据集中的图像风格进行分类和分析。按照 Radford 等人 [32]，的方法，我们为 15 个预定义的图像风格文本模板分别编码类别嵌入，并使用 CLIP 编码的图像嵌入进行线性分类。[11] 我们构建图像风格模板时，将类别名称置于开头，后接其描述

[9]数据短链接：https://tinyurl.com/pfjfa2 5 10 e
https://github.com/rom1504/img2dataset [11]
https://huggingface.co/openai/clip-vit-large-patch14

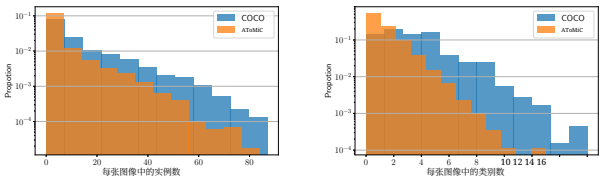
，并采用我们的启发式方法。[12] 我们对图像集以及 COCO 数据集中的图像进行零样本分类，并将占比低于 1% 的长尾风格合并为其他。如图 3 所示，与 COCO 相比，AToMiC 包含更多艺术作品和地图。请注意，由于我们依赖 CLIP 的零样本图像分类能力，图像风格的分类可能不完全准确；未来可通过提供更多图像风格标签或使用更强大的模型来改进。

目标检测分析。除了图像类别等高层信息，探索场景中的底层对象也很重要。尽管 AToMiC 图像集中包含许多非摄影类图像，但其中仍有一定数量的摄影相关图像，如图 3 所示。为进行初步分析，我们利用预训练的 DETR [7] 模型，对 AToMiC 和 COCO 图像集执行目标检测任务。[13] 我们统计 DETR 检测到的实例数和类别数，并以直方图呈现结果，如图 4 所示。AToMiC 中排名前五的目标类别依次为：(1) 人，(2) 领带，(3) 汽车，(4) 椅子和 (5) 书。有趣的是，从图 (a) 可以看出，AToMiC 图像集中每张图像的实例数范围很广；而从图 (b) 可以看出，与 COCO 相比，AToMiC 数据集中的目标类别分布更集中于少数几类。请注意，DETR 是在 COCO 上训练的，因此我们的评估可能存在一定偏差。



(a) AToMiC 图像风格　　(b) COCO 图像风格

图 3：CLIP 零样本图像风格分析



(a) 实例数量　　(b) 类别数量

图 4：DETR 零样本目标检测分析

4.4 评测设置

我们根据语料库规模设置了三种评测方案：

• **小规模**：候选项仅限于数据集中定义的数据划分。这与图像—字幕任务 [18],类似，在这些任务中，系统会在验证集或测试集中搜索候选项。语料库规模约为 20K，适合系统开发。

• **基线**：候选项扩展至包括训练集、验证集和测试集。这通过添加来自其他数据划分的候选项扩大了语料库，使系统难以将相关项与干扰项区分开来。此设置下的语料库规模约为 3M，类似于即席 检索设置，但仅考虑关联的图文对。

• **大规模**：检索所有候选项，包括表 1 中“其他”集合里的非关联图像和文本。图像和文本检索任务的语料库规模均约为 1000 万。这种设置更贴近大规模即席检索的实际情况，也使系统更难判断某张图像是否应关联到某个章节，反之亦然。

我们设计这些评测方案，旨在体现检索任务的难度。目前的图像–标题检索数据集语料库规模有限，难以复现现实中的检索场景。正如 Chun 等人所指出的，这一局限使得识别最新模型之间的差异颇具挑战。 [8, 9]针对这一问题，研究人员尝试将任意图像–标题数据集组合起来，以引入“干扰项”。 [14]然而，使用任意干扰项可能产生不理想的影响，也会使系统效果难以比较、复现和解读。我们的设置专为解决这一问题而设计，旨在为大规模跨模态搜索引擎提供更贴近现实的评测环境。

鉴于相关性标注较为稀疏，我们使用平均倒数排名（MRR@10）和召回率（R@10、R@1K）作为评估指标。需要注意的是，Recall@k 一词可能引起混淆。本文所指的是 $\frac{1}{|Q|}\sum_{q\in\{Q\}}$ 检索到的正例数@k 正例总数，这与计算机视觉领域采用的定义不同：$\frac{1}{|Q|}\sum_{q\in\{Q\}}$ 至少检索到一个正例@k [10, ~~16,~~18],，我们将其记为 Success @k。我们使用 trec_eval 和 ranx [3] 工具包计算这些指标。

5 检索实验

本章节介绍旨在回答以下研究问题的检索实验：

RQ1：最先进的预训练视觉-语言变换器在 AToMiC 测试集上的检索任务中表现如何？RQ2：其检索效果对索引规模的变化是否具有稳健性？RQ3：能否利用所提供的训练数据有效微调预训练 Transformer？RQ4：预训练视觉-语言变换器是否能够捕获文本特征未能捕获的有用信息？

5.1 基线结果

为回答第一个研究问题，我们使用公开可用的实现和模型检查点开展零样本检索实验。为便于在数百万候选项中进行端到端检索（包括索引和搜索），我们考虑了能够独立编码图像和文本的基线。对于级联排序等更复杂检索流程的研究，将留待未来开展。

基线。 我们考察了具有代表性的、可公开获取且能够独立编码图像和文本表示的视觉-语言预训练变换器，即双编码器。

[12]模板的短链接：https://tinyurl.com/3fa8z6fn [13] https://huggingface.co/facebook/detr-resnet-101

	基线（验证集） 条件模型	T2M MRR@10	T2M R@10	T2M R@1K	M2T MRR@10	M2T R@10	M2T R@1K
	预言机方法（将上下文化图像说明作为图像）						
(a)	BM25	0.285bcdefghijk	42.16bcdefghijk	70.76bcdefghijk	0.431bcdefghijk	57.90bcdefghijk	80.73bcdefghijk
	噪声图像-文本对比预训练						
(b)	CLIP ViT-B-32	0.030ijk	5.64dijk	30.83deijk	0.026ijk	4.57ijk	27.39dijk
(c)	CLIP ViT-L-14	0.069bdeijk	12.25dijk	47.34deijk	0.044bdijk	7.72bdeijk	34.67bdeijk
(d)	OpenCLIP ViT-B-32†	0.030ijk	4.57ijk	22.78ijk	0.031bijk	5.26bijk	24.25ijk
(e)	OpenCLIP ViT-B-32	0.038bdijk	6.61bdijk	29.63dijk	0.041bdijk	6.90bdijk	31.82bdijk
(f)	OpenCLIP ViT-L-14	0.067bdeijk	11.37bdeijk	42.42bdeijk	0.067bcdeijk	11.46bcdeijk	42.54bcdeijk
(g)	OpenCLIP ViT-H-14	0.076bcdefijk	12.95bdefijk	46.15bdefijk	0.082bcdefijk	13.61bcdefijk	47.17bcdefijk
(h)	OpenCLIP ViT-G-14	**0.087**bcdefgijk	**14.95**bcdefgijk	**50.06**bcdefgijk	**0.096**bcdefgijk	**15.84**bcdefgijk	**51.34**bcdefgijk
	图像-描述多任务预训练						
(i)	BLIP ViT-B-32	0.008k	1.42	12.46	0.010k	1.84	14.95
(j)	BLIP ViT-L-16	0.011ik	2.20ik	16.15i	0.013ik	2.44ik	18.15i
(k)	FLAVA ViT-B-32	0.004	1.14	19.73ij	0.007	1.62	21.24ij

† LAION-400m

表 3：在 AToMiC-Base 基础设置下，图像推荐和图像推广任务的零样本结果，分别记为 T2M 和 M2T。使用配对 t 检验（$p < 0.05$）进行显著性检验，并采用 Bonferroni 校正 $\alpha = 55$。

此外，我们还将使用字幕作为参考系统的 BM25 纳入考虑。我们的基线如下：

• 预言机方法（BM25）：这是我们的参考系统，它使用上下文化图像说明来表示图像，而不是处理像素值。我们只考虑英语字幕，并将图像的三个字段（参考、替代文本和署名信息）中的文本串联起来。对于文本，我们将表 2 中五个字段的内容串联起来。[14]

• 对比式预训练：CLIP 是一种广为人知的视觉语言预训练变换器，能够理解并生成图像的自然语言描述，还能将图像分类到大量类别中。我们同时考察 HuggingFace 库提供的实现，以及开源复现库 OpenCLIP。[15] OpenCLIP 库提供了基于开源 LAION 数据集 [38, 39] 预训练的 CLIP 变体；相比之下，OpenAI 发布的 CLIP 模型是在其专有数据集上预训练的。考虑到未来工作中还将进行进一步分析，我们以 OpenCLIP 作为主要配置。我们考察了基于 LAION-2B 数据集预训练的 OpenCLIP 变体；该数据集仅包含英语图文对，且未作进一步说明。

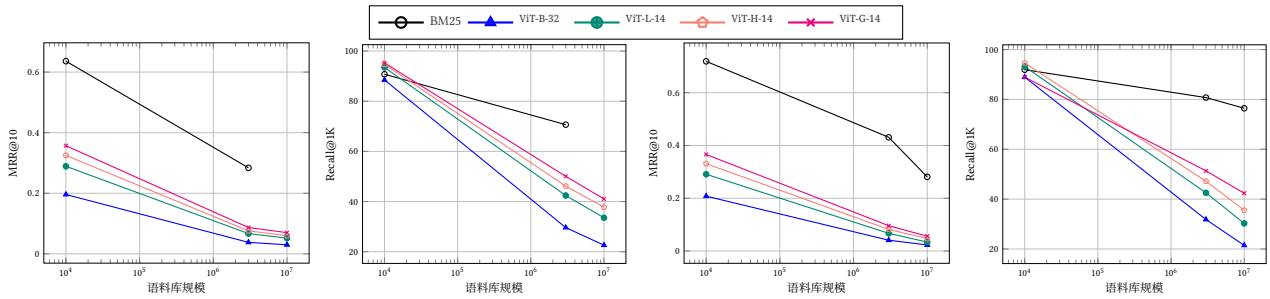
• 多任务预训练：基于 Radford 等人提出的对比目标 [32]，图文数据的多模态学习取得了显著进展。BLIP 模型 [22] 通过联合学习对比、交叉注意力排序和语言建模目标，进一步拓展了这一方向。相比之下，FLAVA 模型的作者 [41]对对比学习之外的其他预训练目标进行了全面研究，包括掩码图像建模、掩码语言建模和掩码多模态建模。我们使用 HuggingFace 上提供的 FLAVA 和 BLIP 模型开源实现及检查点；这两个模型均在 COCO 数据集上进行了微调。[16,17,18]

我们使用 Anserini（用于 BM25）和 FAISS（用于基于 Transformer 的模型），以零样本方式编码图像和文本表示，执行端到端检索。所有基于 Transformer 的基线均使用由五个字段拼接而成的相同文本（与 BM25 相同），但图像则使用像素值。除非另有说明，我们使用验证集分析所有基线。

观察结果。零样本检索实验结果见表 3。第 (a) 行中的 BM25（预言机方法）表明，我们的稀疏相关性标注与经上下文处理的字幕之间具有很高的相关性；对于图像推荐和推广任务，其 MRR@10 分别约为 (0.284, 0.431)，R@1K 分别约为 (70.69, 80.74)。尽管这种方法比基于 Transformer 的模型更有效，但需要注意的是，现实世界中图像的“预言机”字幕十分稀少。比较第 (d) 行和第 (e) 行可以发现，当视觉编码器的规模受限于 ViT-B-32 时，将数据规模从 LAION-400m 扩大到 LAION-2B 带来的收益微乎其微。然而，将视觉编码器从第 (e) 行的 ViT-B-32 扩大到第 (h) 行的 ViT-G-14 后，R@1K 的效果有所提升，得分为 (50.06, 51.34)。在使用图像-标题数据集的零样本检索设置中，多任务预训练 Transformer 通常优于 CLIP；但第 (j)–(l) 行的结果显示，其效果不及第 (b)、(d) 和 (e) 行采用 ViT-B-32 作为视觉编码器的基础 CLIP 设置，

[14] [ ] 标题（P）、标题（S）、
[15] 层级、上下文（S）、上下文（P）
https://github.com/mlfoundations/open_clip

[16] https://huggingface.co/facebook/flava-full [17] https://huggingface.co/Salesforce/blip-itm-base-coco [18] https://huggingface.co/Salesforce/blip-itm-large-coco



（a）文本到音乐检索　　（b）音乐到文本检索

图 5：语料库规模对检索鲁棒性的影响。

条件	基线（验证集） 标题†	上下文（S）	上下文（P）	T2M MRR@10	T2M R@1K	M2T MRR@10	M2T R@1K
(a)	✓	✓	✓	0.067	42.42	0.067	42.54
(b)	✓		✓	0.053	32.76	0.052	34.52
(c)		✓		0.054	38.66	0.057	39.11
(d)			✓	0.017	15.70	0.025	20.86

† 拼接标题（P）、标题（S）和层级

表 4：使用 OpenCLIP ViT-L-14 模型评估不同字段效用的消融研究

鉴于这些模型是在 COCO 数据集上微调的（正如我们所见，该数据集与 AToMiC 截然不同），这一结果是符合预期的。

不同文本字段的消融研究。在此，我们研究了不同文本字段对图像-章节检索任务的效用。我们使用 OpenCLIP ViT-L-14 模型开展消融研究，并在表 4 中展示结果。我们发现，第 (c) 行的上下文（S）提供了最有用的信息，其 R@1K= (38.66,39.11) 优于采用所有可用信息的第 (a) 行中的 R@1K= (42.42,42.54)。鉴于其 R@1K < 10，章节和层级标题等其他字段提供的信息有限。我们认为，可以开发更复杂的技术来融合文本字段，这将留待未来工作。

5.2 检索鲁棒性与语料库规模

为回答关于系统鲁棒性以及检索效果受语料库规模影响的第二个研究问题，我们对 Small、基线和大规模设置下的 OpenCLIP 模型与 BM25 进行了比较分析，结果见图 5。结果表明，在 Small 设置下评估时，OpenCLIP 模型表现出色，两项任务的 R@1K 均约为 90，与使用预言机特征的 BM25 相当。然而，虽然现成的 ViT-G-14 模型在不使用预言机特征时，MRR@10 可达 0.35，但随着语料库规模扩大，其效果会迅速下降。我们也观察到 BM25 的效果呈现类似的下降趋势，但其下降幅度小于 OpenCLIP 模型。值得注意的是，报告 BM25 在大规模设置下的 T2M 任务得分可能会造成误导，因为 BM25 模型未考虑非英语图注，即没有将干扰项编入索引。我们仅将 BM25 用作未来可能改进的参考。最终目标是使自动化系统无需借助人工制作的预言机特征，就能理解图像。

5.3 微调与融合

接下来，我们将探讨最后两个研究问题，分别涉及通过微调和融合实验评估 (a) 训练数据和 (b) 基于 Transformer 的模型的效用。由于这些实验需要选择超参数，我们提供测试集结果作为参考。为便于比较，我们在表 5 的 (a) 至 (d) 行中报告基线运行结果。

微调设置。我们选择了两个经过预训练的 OpenCLIP 模型：ViT-B-32 和 ViT-L-14。我们采用对比损失和批内负样本，遵循 Radford 等人 [32] 的做法，但使用 AToMiC 训练数据。所有模型的批量大小均设为 1024，并使用 AdamW 优化器训练 9075 步，学习率为 10^{-5}。遵循 Zhai 等人 [50], 的做法，我们在微调 CLIP 时尝试了三种设置：(1) 仅微调文本编码器（记作文本）；(2) 仅微调视觉编码器（视觉）；(3) 同时微调两者（文本、视觉）。

融合设置。融合技术，例如早期融合、后期融合或跨媒体融合，有助于提升多媒体检索效果 [2]。我们采用简单的后期融合方法，将 BM25 和 ViT-G-14 检索得到的排序列表进行混合：使用最小-最大归一化的加权和（WSUM）以及倒数排名融合（RRF）。针对验证集上的 MRR@10 优化超参数：WSUM 的权重设为 (BM25, ViT-G-14) $= (0.6, 0.4)$，并将 $k = 30$ 设为 RRF 的参数。

观察结果。表格概览显示，验证集与测试集的评估结果一致且高度相关，证实了前几节中的观察结果。对第 (e) 行至第 (j) 行的分析表明，训练数据在 R@1K 方面提供了有价值的信息。将第 (e) 行至第 (g) 行与第 (c) 行进行比较可见，微调后的 ViT-B-32 模型在零样本设置下的 R@1K 表现与 ViT-L-14 模型相近。第 (h) 行至第 (j) 行中的微调后的 ViT-L-14 模型也呈现出相同趋势。关于我们的微调方法，我们指出两项关键观察结果：(1) 它不会提升 MRR@10；(2) 对两个编码器都进行微调，在 R@1K 方面始终能取得更好的结果。

	Base		验证集				Test			
			T2M		M2T		T2M		M2T	
条件	方法	模型	MRR@10	R@1K	MRR@10	R@1K	MRR@10	R@1K	MRR@10	R@1K
(a)	零样本	BM25	0.285	70.76	0.431	80.73	0.359	78.12	0.447	83.29
(b)		ViT-B-32	0.038	29.61	0.041	31.80	0.038	29.15	0.044	31.91
(c)		ViT-L-14	0.067	42.42	0.067	42.54	0.066	42.40	0.067	42.85
(d)		ViT-G-14	0.087	50.06	0.096	51.34	0.089	51.19	0.095	52.73
微调；用 $\alpha = 6$ 表示 (e,f,g,b) 和 (h,i,j,c)										
(e)	Text	(b)	0.035	38.44bf	0.044bf	41.38bf	0.037	39.66bf	0.049bf	44.06bf
(f)	视觉		0.038e	36.77b	0.036	36.90b	0.040	38.38b	0.039	39.20b
(g)	文本，视觉		0.037	42.31bef	0.043b	43.80bef	0.043be	44.683bef	0.0483b	46.813bef
(h)	Text	(c)	0.066	49.36ci	0.074cij	51.95ci	0.063	50.13c	0.075cij	53.38ci
(i)	视觉		0.066	47.96c	0.059	47.21c	0.065	49.73c	0.059	48.16c
(j)	文本、视觉		0.066	52.67chi	0.068i	53.63chi	0.065	54.84chi	0.071i	56.05chi
融合；with $\alpha = 6$ for (a,d,k,l)										
(k)	RRF	(a) + (d)	0.276d	**79.01**广告	0.371天	**86.36**广告	0.323天	**84.46**广告	0.384天	**88.42**广告
(l)	WSUM		**0.329**adk	78.95ad	**0.468**adk	86.29ad	**0.402**adk	84.45ad	**0.483**adk	88.33ad

表 5：微调模型及融合结果。采用配对t检验进行显著性检验（$p < 0.05$），使用 Bonferroni 修正 $\alpha = 6$。

实验结果表明，尽管 BM25 基线利用了预言机特征，通过后期融合技术纳入基于 Transformer 的模型能够带来显著改进。这一发现说明，基于 Transformer 的模型仍能捕获有价值的信息。采用 WSUM 方法时，MRR@10 和 R@1K 均有所提升；尽管 RRF 在 MRR@10 方面不及 WSUM，但在 R@1K 方面表现相当。我们还比较了 BM25 与 ViT-G-14 的检索运行覆盖率，发现 BM25 覆盖的候选结果仅约为 ViT-G-14 检索结果的 30%。据此我们推测，预训练 OpenCLIP 模型能够捕获图像中的缺失信息，因为图像所包含的信息十分丰富，而上下文化图像说明可能不足以涵盖图像的所有方面。

5.4 案例研究

在本小节中，我们通过考察一个案例研究中的检索候选项进行定性分析。图6展示了检索到的图像。文本查询取自 Nessos Painter 页面上的“作品示例”章节，并显示在图像下方。原始附加图像，即“真实标注”，显示在面板 (a) 中。我们从 BM25 和 OpenCLIP ViT-L-14 的检索结果列表（共 1K 项）中，分别选取排名前 10 和后 10 的候选项，如面板 (b) 和 (c) 所示。

我们的发现表明，OpenCLIP 模型倾向于检索相似的艺术品，而在本例中，BM25 模型已经开始关注排序列表底部的人物。这个例子说明，图像推荐任务需要更多关于文本中场景的背景知识，例如章节“In the namevase amphora depicting Nessos fighting Heracles.”中的关键线索。OpenCLIP 或许能够识别描绘相关人物的花瓶，但在结合上下文做出精确判断时仍有困难。

6 局限性与未来工作

使用 AToMiC 数据集时，了解其局限性非常重要。尽管如此，我们认为这些局限性为研究社区开展讨论和进一步探索提供了机会。通过这样做，我们可以共同努力，提高该数据集在未来研究中的质量和实用性。

稀疏判断。 目前，AToMiC 因使用稀疏代理相关性标签而存在一项重要局限。相关性判断集合基于对所有可能图像—文本关系的高度偏倚抽样。我们的实验表明，即使存在大量空字幕，基于字幕的 BM25 在 MRR@10 和 R@1K 指标上仍表现出色。然而，这种数据偏差与其在多媒体内容创作情境中的相关性之间有何联系，目前尚不明确。同时，仅依赖字幕可能会引发安全问题，例如将与相应字幕配对的随机图像排在最前面，因为像素到字幕的关系被忽略了。此外，数据集中稀疏的二元标注限制了 MAP 和 nDCG 等指标在进一步分析中的使用。为缓解这些问题，需要采用更成熟的评估方法，例如 TREC 评估。

流行度偏差。 AToMiC 数据集包含多样化的文本和图像数据，这些数据取自维基百科页面的自然分布。然而，这也可能引入偏差，因为使用宏平均评估系统时，若对所有文本和图像“查询”进行宏平均，该数据集就会受到流行度偏差的影响。在对文本和图像数据的分析中，我们发现了许多长尾案例，需要自动化系统提升其质量或效用。考虑所有文本和图像的宏平均指标可能无法充分反映这些长尾案例。因此，未来的研究应致力于开发和评估能够满足这些长尾案例特定需求的自动化系统，



(a) 原始 at- 附加图像。

（b）BM25 检索出的前10个和后10个（共1K个）候选项

（c）OpenCLIP ViT-L-14 检索出的前10个和后10个（共1K个）由 OpenCLIP ViT-L-14 检索得到

'text_id': 'projected-06900598-003', 'page_title': 'Nessos Painter', 'section_title': '作品示例', 'context_section_description': "描绘 Nessos 与 Heracles 战斗的命名花瓶双耳瓶中，画家运用了图像志，例如描绘了一位留着小胡子的形象。这不同于通常将 Heracles 画成留着胡须，并身着其惯常装束——狮皮斗篷和狮面具的艺术品。Nessos 和 Heracles 的名字都写在他们上方，表明....."

图6：比较检索候选项的案例研究。

同时也应考虑流行度偏差对这些系统有效性的潜在影响。

静态测试集。 AToMiC 数据集记录了维基百科的一个快照，其中包含数据集整理时的图像、文本及其关联。然而，需要承认的是，维基百科的一个关键属性是其时变特性，而我们在评估中未考虑这一点。随着时间推移，编辑者可能会选择替换现有的文本和图像，从而影响底层数据。此外，如果编辑者在页面中添加或删除图像，我们在数据集中捕获的图文关联也可能发生变化。因此，必须意识到维基百科的动态特性，以及底层数据的变化可能对未来研究产生的影响。未来的工作可以考虑维基百科的时变特性，以便开发和评估依赖该数据集的自动化系统。

缺乏对生成模型评估的支持。 近期，文本生成图像 [35, 37] 和文本生成文本生成模型 [6, 33] 的发展展现出了巨大潜力，有助于提升创意数字内容创作任务。然而，AToMiC 数据集对评估此类模型的支持有限。这是因为数据集中的所有图像都需要预先定义并建立索引，以用于搜索，这可能不利于评估能够采样现实世界中不存在的图像的生成模型。在某些场景下，生成的图像或文本可能会提升数据集在数字内容创作任务中的质量和实用性。未来的研究应考虑将此类生成模型纳入评估，以便更好地衡量其在现实场景中的有效性。

7 结论

在本文中，我们介绍了 AToMiC 数据集，这是一个真实的测试集，旨在促进可扩展、多样且可复现的多媒体检索研究。我们介绍了一种构建该数据集的方法，其中包含一个基于维基百科文本和图像集的真实用户模型。AToMiC 数据集提供了评估资源、基线以及一个社区，用于研究文本和图像的自然共现。

我们的实验展示了 AToMiC 数据集的能力，并将其与广泛使用的 COCO 数据集进行了对比。我们制定了两项检索任务，并使用基于 Transformer 的模型进行了初步评估。结果表明，AToMiC 数据集更真实地呈现了自然环境中图像与文本之间的关系。

总的来说，我们将 AToMiC 提供为一个测试集，用于开发旨在辅助多媒体内容创作的图文检索系统。我们认为，该数据集有助于分析图像—文本关系，并推动大规模图文检索基准的发展。我们鼓励研究人员在未来的多媒体检索研究中使用 AToMiC 数据集，并欢迎进一步的反馈与贡献。随着该数据集的开放，我们期待研究界在理解图像-文本关系以及开发多媒体内容创作工具方面取得重大进展。

致谢

本研究部分由加拿大自然科学与工程研究委员会（NSERC）资助。

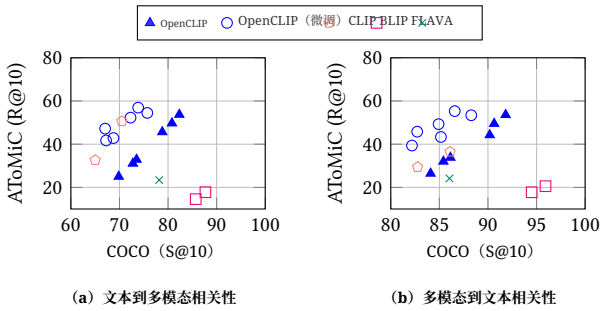
参考

[1] Julien Ah-Pine、Marco Bressan、Stephane Clinchant、Gabriela Csurka、Yves Hop- penot 和 Jean-Michel Renders。2009年。《在不同应用场景中交织文本与视觉内容》。*Multimedia Tools Appl.* 42, 1（2009年3月），31–56页。[2] Julien Ah-Pine、Gabriela Csurka 和 Stéphane Clinchant。2015年。《在基于内容的医学图像检索中使用基于图的方法进行无监督视觉与文本信息融合》。*ACM Trans. Inf. Syst.* 33, 2，文章9（2015年2月），共31页。[3] Elias Bassani。2022年。《ranx：用于排序评估与比较的极速 Python 库》。载于*ECIR* 论文集（计算机科学讲义，第*13186*卷），259–264页。

[4] Sarah Bibi、Dian Shinta Sari 和 Muhammad Iqbal Ripo Putra。2020年。《多媒体叙事的设计》。载于《英语教学杂志》第9卷，16–26页。[5] Maurits Bleeker 和 Maarten de Rijke。2022年。《度量学习的经验能否推广到图像字幕检索？》。载于 ECIR 论文集，535–551页。[6] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。《语言模型是少样本学习器》。NeurIPS 论文集 33（2020年），1877–1901页。[7] Nicolas Carion、Francisco Massa、Gabriel Synnaeve、Nicolas Usunier、Alexander Kirillov 和 Sergey Zagoruyko。2020年。《使用 Transformer 进行端到端目标检测》。载于 ECCV 论文集，213–229页。[8] Sanghyuk Chun、Wonjae Kim、Song Park、Minsuk Chang Chang 和 Seong Joo n Oh。2022年。《ECCV Caption：通过收集经机器与人工验证的 MS-COCO 图像—字幕关联来纠正假负例》。载于 ECCV 论文集，1–19页。[9] Sanghyuk Chun、Seong Joon Oh、Rafael Sampaio De Rezende、Yannis Kalantidis 和 Diane Larlus。2021年。《用于跨模态检索的概率嵌入》。载于 IEEE/CVF CVPR 论文集，8415–8424页。[10] Fartash Faghri、David J Fleet、Jamie Ryan Kiros 和 Sanja Fidler。2018年。《VSE++：利用困难负例改进视觉语义嵌入》。（2018年）。[11] Norbert Fuhr、Jaap Kamps、Mounia Lalmas、Saadia Malik 和 Andrew Trotman。2007年。《INEX 2007年即席检索赛道综述》。载于 INEX 论文集，1–23页。[12] Kazuki Fujii、Hidetsugu Nanba、Toshiyuki Takezawa 和 Aya Ishino。2016年。《通过旅行博客条目和已回答问题档案丰富旅行指南》。载于 ENTER eTourism 论文集，157–171页。[13] Debasis Ganguly、Iacer Calixto 和 Gareth JF Jones。2015年。《FIRE 2015年自动故事插图任务综述》。载于 FIRE Workshops，63–66页。[14] Gregor Geigle、Jonas Pfeiffer、Nils Reimers、Ivan Vulić 和 Iryna Gurevych。2022年。《快速检索，智能重排序：改进跨模态检索的协作与联合方法》。《计算语言学协会汇刊》10（2022年），503–521页。[15] Xintong Han、Zuxuan Wu、Phoenix X Huang、Xiao Zhang、Menglong Zhu、Yuan Li、Yang Zhao 和 Larry S Davis。2017年。《自动发现空间感知时尚概念》。载于 ICCV 论文集，1463–1471页。[16] Micah Hodosh、Peter Young 和 Julia Hockenmaier。2013年。《将图像描述构建为排序任务：数据、模型与评估指标》。JAIR 47（2013年），853–899页。[17] Bogdan Ionescu、Henning Müller、Renaud Péteri、Johannes Rückert、Asma Ben Abacha、Alba G. Seco de Herrera、Christoph M. Friedrich、Louise Bloch、Raphael Brüngel、Ahmad Idrissi-Yaghir、Henning Schäfer、Serge Kozlovski、Yashin Di- cente Cid、Vassili Kovalev、Liviu-Daniel Ştefan、Mihai Gabriel Constantin、Mihai Dogariu、Adrian Popescu、Jérôme Deshayes-Chossart、Hugo Schindler、Jon Cham- berlain、Antonio Campello 和 Adrian Clark。2022年。《ImageCLEF 2022年综述：医学、社交媒体与自然应用中的多媒体检索》。载于《实验信息检索与多语言、多模态及交互》，Alberto Barrón-Cedeño、Giovanni Da San Martino、Mirko Degli Esposti、Fabrizio Sebastiani、Craig Macdonald、Gabriella Pasi、Allan Hanbury、Martin Potthast、Guglielmo Faggioli 和 Nicola Ferro（编），541–564页。[18] Andrej Karpathy 和 Fei-Fei Li。2015年。《用于生成图像描述的深度视觉语义对齐》。载于 IEEE/CVF CVPR 论文集，3128–3137页。[19] Elisa Kreiss、Fei Fang、Noah Goodman 和 Christopher Potts。2022年。《Concadia：迈向有目的的基于图像的文本生成》。载于 EMNLP 论文集。阿联酋阿布扎比，4667–4684页。[20] Julia Kruk、Jonah Lubin、Karan Sikka、Xiao Lin、Dan Jurafsky 和 Ajay Divakaran。2019年。《整合文本与图像：判定 Instagram 帖子的多模态文档意图》。载于 EMNLP-IJCNLP 论文集，4622–4632页。[21] Joel R Levin 和 Alan M Lesgold。1978年。《散文中的图片》。ECTJ 26, 3（1978年），233–243页。[22] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。2022年。《BLIP：用于统一视觉语言理解与生成的自举式语言图像预训练》。载于 ICML 论文集。[23] Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。2021年。《先对齐后融合：通过动量蒸馏进行视觉与语言表征学习》。NeurIPS 论文集 34（2021年），9694–9705页。[24] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。2014年。《微软 COCO：上下文中的常见物体》。载于 ECCV 论文集，740–755页。[25] Fuxiao Liu、Yinghan Wang、Tianlu Wang 和 Vicente Ordonez。2021年。《视觉新闻：新闻图像字幕的基准与挑战》。载于 EMNLP 论文集，6761–6771页。[26] Andreas Lommatzsch、Benjamin Kille、Özlem Özgöbek、Yuxiao Zhou、Jelena Tešić、Cláudio Bartolomeu、David Semedo、Lidia Pivovarova、Mingliang Liang 和 Martha Larson。2022年。《NewsImages：弥合描绘差距》

一项用于文本-图像重新匹配的在线新闻数据集。载于 ACM多媒体系统会议论文集。227–233.[27] Kenneth Marino、Mohammad Rastegari、Ali Farhadi 和 Roozbeh Mottaghi。2019年。OK-VQA：一项需要外部知识的视觉问答基准。载于 IEEE/CVF CVPR 论文集。3195–3204。[28] Emily E Marsh 和 Marilyn Domas White。2003年。图像与文本关系分类体系。《文献学杂志》59卷第6期（2003年），647–672。[29] Henning Müller、Paul D. Clough、Thomas Deselaers 和 Barbara Caputo（编）。2010年。《ImageCLEF：视觉信息检索的实验评估》。施普林格。[30] Vivien Petras 和 Paul D. Clough。2011年。CLEF 2011实验室简介。载于 CLEF（笔记论文/实验室/研讨会），Vivien Petras、Pamela Forner 和 Paul D. Clough（编）。[31] Adrian Popescu、Theodora Tsikrika 和 Jana Kludas。2010年。ImageCLEF 2010维基百科检索任务概述。载于 CLEF（笔记论文/实验室/研讨会）。[32] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sand-hini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021年。通过自然语言监督学习可迁移的视觉模型。载于 ICML 论文集。8748–8763。[33] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。2020年。探索统一文本到文本转换器的迁移学习极限。《机器学习研究杂志》21卷第140期（2020年），1–67。[34] Daniele Rama、Tiziano Piccardi、Miriam Redi 和 Rossano Schifanella。2022年。维基百科上读者与图像交互的大规模研究。《EPJ数据科学》11卷第1期（2022年），1。[35] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Rad-ford、Mark Chen 和 Ilya Sutskever。2021年。零样本文本到图像生成。载于 ICML 论文集，第139卷。8821–8831。[36] Jun Rao、Fei Wang、Liang Ding、Shuhan Qi、Yibing Zhan、Weifeng Liu 和 Dacheng Tao。2022年。性能提升来自何处？——关于图像-文本检索的可复现性问题。载于 信息检索研究大会论文集。2727–2737。[37] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022年。利用潜在扩散模型进行高分辨率图像合成。载于 IEEE/CVF CVPR 论文集。10684–10695。[38] Christoph Schuhmann、Ro main Beaumont、Richard Vencu、Cade W Gordon、Ross Wightman、Mehdi Cherti、Theo Coombes、Aarush Katta、Clayton Mullis、Mitchell Wortsman、Patrick Schramowski、Srivatsa R Kundurthy、Katherine Crowson、Ludwig Schmidt、Robert Kaczmarczyk 和 Jenia Jitsev。2022年。LAION-5B：用于训练下一代图像文本模型的大规模开放数据集。载于 NeurIPS：数据集与基准分会场。[39] Christoph Schuhmann、Richard Vencu、Romain Beaumont、Robert Kaczmarczyk、Clayton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev 和 Aran Komatsuzaki。2021年。LAION-400m：由 CLIP 筛选的4亿图像文本对开放数据集。arXiv:2111.02114（2021年）。[40] Dustin Schwenk、Apoorv Khandelwal、Christopher Clark、Kenneth Marino 和 Roozbeh Mottaghi。2022年。A-OKVQA：运用世界知识进行视觉问答的基准。arXiv:2206.01718（2022年）。[41] Amanpreet Singh、Ronghang Hu、Vedanuj Goswami、Guillaume Couairon、Woj-ciech Galuba、Marcus Rohrbach 和 Douwe Kiela。2022年。FLAVA：一种基础语言与视觉对齐模型。载于 IEEE/CVF CVPR 论文集。15638–15650。[42] Krishna Srinivasan、Karthik Raman、Jiecao Chen、Michael Bendersky 和 Marc Najork。2021年。WIT：用于多模态多语言机器学习的基于维基百科的图像文本数据集。载于 信息检索研究大会论文集。2443–2449。[43] Besiki Stvilia、Michael B Twidale、Les Gasser 和 Linda C Smith。2005年。社区协作百科全书中的信息质量。载于《知识管理：培育文化、创新与技术》。101–113。[44] Theodora Tsikrika 和 Jana Kludas。2010年。维基百科图像检索任务。载于《ImageCLEF：视觉信息检索的实验评估》，Henning Müller、Paul D. Clough、Thomas Deselaers 和 Barbara Caputo（编）。163–183。[45] Alakananda Vempala 和 Daniel Preoţiuc-Pietro。2019年。推文中文本与图像关系的分类与推断。载于 ACL 论文集。2830–2840。[46] P. Welinder、S. Branson、T. Mita、C. Wah、F. Schroff、S. Belongie 和 P. Perona。2010年。加州理工学院-加州大学圣迭戈分校鸟类200。技术报告 CNS-TR-2010-001。加州理工学院。[47] Qian Xu 和 S Shyam Sundar。2014年。灯光、摄影、音乐、互动！电子商务中的互动式说服。《传播研究》41卷第2期（2014年），282–308。[48] Lewei Yao、Runhui Huang、Lu Hou、Guansong Lu、Minzhe Niu、Hang Xu、Xiao-dan Liang、Zhenguo Li、Xin Jiang 和 Chunjing Xu。2022年。FILIP：细粒度交互式语言-图像预训练。载于 国际学习表征会议论文集。[49] Peter Young、Alice Lai、Micah Hodosh 和 Julia Hockenmaier。2014年。从图像描述到视觉指称：事件描述语义推理的新相似度度量。《计算语言学协会汇刊》2卷（2014年），67–78。[50] Xiaohua Zhai、Xiao Wang、Basil Mustafa、Andreas Steiner、Daniel Keysers、Alexan-der Kolesnikov 和 Lucas Beyer。2022年。LiT：通过锁定图像的文本调优实现零样本迁移。载于 IEEE/CVF CVPR 论文集。18123–18133。

[51] Xinyu Zhang、Nandan Thakur、Odunayo Ogundepo、Ehsan Kamalloo、David Alfonso-Hermelo、李晓光、刘群、Mehdi Rezagholizadeh 和 Jimmy Lin。2022年。构建 MIRACL：跨越语言连续体的多语言信息检索。*arXiv:2210.09984*（2022年）。[52] 张宇豪、蒋航、三浦康秀、Christopher D Manning 和 Curtis P Langlotz。2022年。基于配对图像和文本的医学视觉表征对比学习。载于*MLHC* 论文集。2–25。



（a）文本到多模态相关性　　（b）多模态到文本相关性

图 7：效果对比。我们报告 AToMiC 上的 R@10 和 COCO 上的 S@10，比较基于 Transformer 的基线模型以及在 AToMiC 上微调的 OpenCLIP 模型。

A ATOMIC 与 COCO

在本小节中，我们扩展了对模型在 AToMiC 和 COCO 数据集上的检索效果的分析。评估 COCO 数据集时，我们遵循以往研究采用的常见设置，使用 Small 设置下测试集上的 Success@10（S@10）指标。同样，对于 AToMiC，我们报告 Small 设置下 AToMiC 验证集上的 R@10。由于 COCO 数据集采用图像–标题检索设置，不包含预言机特征，因此我们仅比较基于 Transformer 的模型。我们在图 7 中以散点图呈现 T2M 和 M2T 的结果。结果表明，FLAVA 和 BLIP 模型的效果更偏向 COCO，而在 AToMiC 上表现较差。相比之下，CLIP 和 OpenCLIP 模型在两个数据集上的效果较为均衡。我们推测，这是因为 FLAVA 和 BLIP 模型是在汇总的图像–标题数据集上训练的，这与我们的任务设置有很大不同。BLIP 模型的效果偏向性尤为明显，因为作者提出在模型自身生成的自举生成的图像–标题对上进一步训练模型。最后，我们的微调方案仍有改进空间。如图所示，经过微调的 OpenCLIP 模型在 COCO 上表现较差。这可能归因于持续学习设置中的灾难性遗忘现象。

B 图像分析

图像以 Pillow 对象的字节形式存储在图像列中。为便于在第 4.3 节进行图像分析，用户可以使用 Transformers 库。[19] 我们还提供了使用 CLIP 和 DETR 等预训练模型分别进行图像风格分类和目标检测任务的示例。

B.1 图像风格分类

对于图像风格分类任务，我们采用了 Radford 等人提出的方法 [32]，并将 CLIP 用作零样本线性分类器。我们整理了 15 种图像风格，格式为 <图像风格>: <描述>，并在每个图像风格类别前添加了两个提示词："一张……的图像"和"一幅……的渲染图"。我们使用 CLIP 文本编码器对每个类别的权重进行编码，其中，每个权重向量是同一图像风格类别在使用两个提示词后得到的 L2归一化向量的平均值。有关图像风格分类任务的详细示例，请参阅笔记本。[20]

B.2 目标检测

为执行目标检测任务，我们采用了预训练DETR模型。我们使用该模型处理图像，并获取检测到的实例数量和类别。代码仓库中的笔记本详细介绍了这一流程。[21]作为设计多阶段图像–文本检索流程的预处理步骤，用户可以利用边界框提取区域特征，并在此基础上构建匹配模型 [14]。

[19]文档见 https://huggingface.co/docs/transformers/index

[20]https://github.com/TREC-AToMiC/AToMiC/blob/f9c6ac86dc0baac5e8c84f3fdf75a37e0022af1a/notebooks/02_图像_风格.ipynb [21] https://github.com/TREC-AToMiC/AToMiC/blob/f9c6ac86dc0baac5e8c84f3fdf75a37e0022af1a/notebooks/03_对象_检测.ipynb