arXiv:2301.01893v1 [cs.CV] 5 Jan 2023

GIVL：通过预训练方法提升视觉语言模型的地理包容性

Da Yin[1] Feng Gao[2] Govind Thattai[2] Michael Johnston[2] Kai-WeiChang[1,2] 加利福尼亚大学洛杉矶分校 [2] 亚马逊 Alexa AI{da.yin,kwchang}@cs.ucla.edu, {fenggo,thattg,mjohnstn}@amazon.com



图 1. 世界各地的场景，包括节日和婚礼。即使是相同的场景，在不同地区也具有鲜明不同的视觉特征（也称为地理多样性）。与以往的视觉-语言预训练模型（VLPs）相比，GIVL 在 GD-VCR 的非西方数据上表现出色 [48]。GIVL 还可以显著缩小西方与非西方案例之间的差距。

摘要

人工智能发展的一个关键目标，是开发不仅满足某一群体需求、而且服务于所有社区的技术，不论这些社区位于哪个地理区域。事实上，相当一部分知识由特定地区的人们共同分享，但由于文化差异，这些知识未必同样适用于其他地区。如果模型不了解区域特征，就可能导致不同地区之间出现性能差异，并对代表性不足的群体造成偏见。我们提出了 *GIVL*，即 ***Geographically Inclusive Vision-and-Language*** 预训练模型。地理多样化视觉概念具有两项有助于学习地理多样化知识的属性：*1*）相似类别下的概念具有独特的知识和视觉特征；*2*）视觉特征相似的概念可能属于截然不同的类别。受这些属性启发，我们设计了新的预训练目标——图像-知识匹配（*IKM*）和图像编辑检查（*IEC*），用于预训练 *GIVL*。与使用相近规模数据预训练的同等规模模型相比，*GIVL* 在地理多样化视觉与语言任务上取得了最先进的（*SOTA*）且更加均衡的表现。

1. 引言

视觉-语言预训练模型（VLPs） [9, 23, 24,29, 52] 在视觉与语言（V&L）任务中取得了显著的表现，包括视觉问答 [11, 12, 15], 图文检索 [22], 和图像描述 [19, 27]。这些模型使用大规模图文对语料进行预训练，例如 COCO [27], 和 OpenImages [21]。VLPs 能够学习多模态表征，并可在下游视觉与语言任务上进行有效微调。

虽然 VLPs 能够解决广泛的视觉与语言任务，但要将 VLPs 部署于现实应用中，必须考虑其地理包容性[1]。由于地理差异，不同地区的图像承载着大量仅在当地共享、却无法应用于其他地区的知识，也就是具有地理多样性。例如，在图 1中，不同地区的节日各不相同。

理想情况下，具有地理包容性的视觉语言预训练模型应能够对所有图像取得相当的表现，而不受图像来源的影响。然而，目前的视觉语言预训练模型

[1]我们使用地区作为衡量视觉与语言模型包容性的代理指标。同一地区的人们可能拥有不同的文化和传统。

在不同地区的数据上表现并不相同。例如，先前的研究 [28,48] 表明，在地理多样化的视觉与语言任务中，应用当前视觉语言预训练模型时，西方图像与东亚图像之间的表现差距接近 20%。为应对这种地理偏见，我们旨在设计相关方法，使视觉语言预训练模型在不同地区的表现更加均衡。

缓解偏见的一种方法是，为每个地区获取多样化的任务专属标注，并使用新的标注对视觉语言预训练模型进行微调。然而，根据 [17], ，亚马逊 Mechanical Turk 上的大多数标注人员来自美国和印度，可能不熟悉其他地区的文化。因此，即使在这样一个热门的众包平台上，获取大规模地理多样化标注也不现实。

使用大规模的统一、未经标注的地理多样化图像及其对应知识对视觉语言预训练模型进行预训练，可以使其成为基础模型，提供更具泛化能力的表征，并更轻松地迁移到理解来自不同地区的图像上。本文提出**GIVL**，即**地**理**包**容性**视**觉-**语**言预训练模型。我们重点研究**如何在预训练阶段促使 GIVL 更好地学习图像中蕴含的地理多样化知识**。

我们观察到地理多样化视觉概念具有两个有助于学习地理多样化知识的属性：

A1：相似类别下的概念具有独特知识和视觉特征。例如，西方和中国的传统节日，例如圣诞节和春节（见图1），其庆祝仪式各不相同，装饰风格也有所区别。GIVL 必须学习它们对应知识之间的差异，并准确区分这些视觉概念。另一方面，圣诞节和春节都属于节日。学习视觉概念的共性（例如，图1中的两幅图像都属于同一“节日”类别）有助于模型关联西方与非西方概念，并更有效地迁移到地理多样化图像上。

A2：视觉特征相似的概念可能属于完全不同的类别。在图2中，中国剪纸与红色飞盘具有相似的视觉特征（例如颜色、形状）。同样，甘蔗和长笛也具有相似的视觉特征。然而，这些概念彼此无关。由于地理多样化的图像涵盖了更广泛的视觉
概念，因此，在视觉上下文中区分视觉上相似的概念也至关重要。

为此，除了视觉语言预训练模型常用的预训练目标——掩码语言建模（MLM）和图文匹配（ITM）之外，我们还提出了两个额外的预训练目标：**图像-知识匹配（IKM）**和**图像编辑检查（IEC）**。IKM 用于学习图像与维基百科中相应文本知识之间的对齐。它要求 GIVL 不仅要判断



图2：中国剪纸与红色飞盘（左），以及甘蔗与长笛（右）的示例。不同的视觉概念可能具有相似的视觉特征，但其功能可能截然不同。

输入文本知识是否与输入图像匹配，还要识别输入知识中描述的视觉概念是否属于与输入图像中的概念相似的类别。这促使 GIVL 学习知识与图像之间相应的关系，并识别地理多样化视觉概念之间的相似性。IEC 旨在判断输入图像中的某个视觉概念是否被另一个视觉上相似、但类别无关的概念所替换（示例见图3）。借助视觉上下文，IEC 使 GIVL 能够捕捉替换前后视觉上相似概念之间的细微差异。

我们的贡献和实证结果如下：

- 考虑到地理多样化视觉概念的属性，我们提出了两种新颖的视觉与语言预训练目标：图像-知识匹配（IKM）和图像编辑检查（IEC），它们能够显著提升视觉语言预训练模型的地理包容性。

- 与采用相近规模数据进行预训练的同等规模视觉语言预训练模型相比，GIVL[2] 在包括 MaRVL 在内的地理多样化视觉与语言任务上，取得了最先进（SOTA）的表现，并且在不同地区的表现更加均衡 [28], GD-VCR [48] 以及 WIT 图文检索 [37]。在 Dollar Street 数据集上的地理多样化零样本图像分类任务中[3]，GIVL 的表现优于 VinVL [52] 26%。

2. 相关工作

视觉-语言预训练模型（VLPs）。VLPs[4, 9, 22–24, 26, 29, 35, 41, 52] 旨在应对需要同时理解图像和文本的任务。遵循 [8,30,32], 通常采用的语言模型预训练范式，VLPs 使用 Transformer [42] 作为骨干网络，并利用大规模图像-标题对对其进行预训练。常用的图文平行数据来自多个来源，包括 COCO [27], 、Flickr30K [49],、Conceptual Captions [34]和 OpenImages [21] 数据集。目前，VLPs 已在各种视觉与语言任务上取得了显著性能，包括视觉问答 [12, 15], 、视觉推理 [39], 和图像描述 [27],

[2]代码和模型检查点将会发布。[3]Dollar Street 数据集可在以下网址获取：https://github.com/ greentfrapp/dollar-street-images。

以及图文检索 [27, 49]。最近的研究大多聚焦于扩大 VLPs 的规模；本文则研究一个正交但重要的问题——如何利用多样化知识来提升 VLPs 的包容性。

地理偏差。 地理偏差 [7, 33, 43, 47] 是人工智能应用中的一个严重问题。以往研究 [28, 48] 揭示，在地理多样化视觉与语言任务中，使用视觉语言预训练模型时，非西方图像与西方图像之间存在显著的性能差距。同样，物体识别模型在非西方图像上的性能也会大幅下降 [7, 33]。研究人员 [7, 33, 43] 发现，地理偏差的一个成因是训练数据在地理位置分布上的不平衡。他们 [7] 观察到，COCO 和 OpenImages 这两个视觉语言预训练模型广泛使用的预训练语料库，都以美洲和欧洲为中心。性能下降的另一个原因是，视觉语言预训练模型能够理解来自不同地区图像中的基本视觉信息，但较难利用地理多样化知识并进行推理 [48]。

地理多样化视觉与语言任务。 GD-VCR [48] 研究模型能否理解地理多样化图像中的常识。视觉与语言模型需要根据涉及地理多样化常识的文本问题及相应图像，从四个答案选项中选出正确答案。MaRVL [28] 是另一项视觉与语言任务，要求模型结合非西方地区图像中的文化知识进行视觉推理。该任务被设计为一个二元分类问题，模型需要判断某个句子是否正确描述了两张来自非西方地区的图像。WIT 图像文本检索 [3, 37] 是一项针对维基百科地理多样化图像的标准多模态检索任务。

3. 方法

本节将详细介绍 GIVL 的预训练方法。第 3.1 节介绍 GIVL 预训练方法的基础内容，包括视觉概念和类别的定义。第 3.2节描述四种预训练目标。第 3.3 和 3.4 节阐述获取必要信息的过程，这些信息用于构建图像-知识匹配（IKM）和图像编辑检查（IEC）目标的输入内容。具体而言，第 3.3节介绍如何从图像描述中提取视觉概念名称，并从维基百科中获取其类别信息。第 3.4节介绍如何在输入图像中为检测到的物体定位对应的视觉概念。

3.1. 基础知识

视觉概念与类别的定义。 视觉概念 是指图像主要涉及的对象或场景。例如，图3展示了中国剪纸这一视觉概念。每种具体的视觉概念都对应一个通用的 类别。每个类别涵盖具有特定共同特征的各种视觉概念。例如，视觉概念 中国剪纸 所属的类别是 艺术。 艺术 类别还包括其他视觉概念，例如犹太剪纸。视觉概念及其类别信息的提取流程将在第3.3节介绍。

预训练语料库。 为提升视觉语言预训练模型的地理包容性，我们使用维基百科图文（WIT）数据集作为地理多样性图像的主要来源。 [37]WIT 共包含 295 万张图像[4]。我们还纳入了 22 万张常用的视觉与语言预训练图像，来自 COCO、 [27], Flickr30K [49], 和 GQA。WIT 数据集中的图像附有相应的维基百科章节，其中包含与 WIT 图像相关的知识。这些知识[5]，（例如风俗和历史）通常与文化相关，却未在图像中明确呈现。这些知识对于帮助视觉语言预训练模型更全面地理解地理多样性图像中的视觉概念至关重要。

预训练输入。我们按如下方式组织 GIVL 预训练的输入：

$$[\text{CLS}]\ \mathbf{c}\ [\text{SEP}]\ \mathbf{k}\ [\text{SEP}]\ \mathbf{t}\ [\text{SEP}]\ \mathbf{v}, \tag{1}$$

其中，$\mathbf{c}$ 可以是图像描述，也可以是 GQA 问题；$\mathbf{k}$ 表示输入图像$\mathbf{I}$中视觉概念对应的知识；$\mathbf{t}$ 可以是检测到的物体的标签，也可以是 GQA 答案；$\mathbf{v}$ 是由 ResNeXt-152 C4 检测模型从输入图像$\mathbf{I}$ 生成的视觉嵌入列表 [52]。 p_v是图像$\mathbf{I}$中所含视觉概念的名称。

3.2. GIVL 的预训练目标

我们使用四种目标对 GIVL 进行预训练：掩码语言建模（MLM）、图文匹配（ITM）、图像-知识匹配（IKM）和图像编辑检查（IEC）。下面分别介绍各项预训练目标。

3.2.1 掩码语言建模与图文匹配目标

掩码语言建模（MLM）是视觉与语言预训练中普遍采用的一种学习目标。给定模型输入的上下文，GIVL 需要还原被 [MASK] 掩盖的词元。MLM 损失 $\mathcal{L}_{MLM}$ 是针对词汇表中正确掩码词元预测概率计算的所有交叉熵损失的平均值。

图文匹配（ITM）是另一种常用的目标，使 GIVL 能够学习文本与图像之间的对齐关系。遵循 VinVL [52], , 给定输入图像$\mathbf{I}$，我们构造三种类型的输入内容

[4]GIVL 专注于仅使用英语的视觉与语言任务。我们只考虑带有英文图像描述的图像，这些图像仅占 WIT 全部图像的 30%。[5]COCO、Flickr30K 和 GQA 图像所涉及的知识，是图像描述中提及的物体对应的维基百科页面中的第一句话。

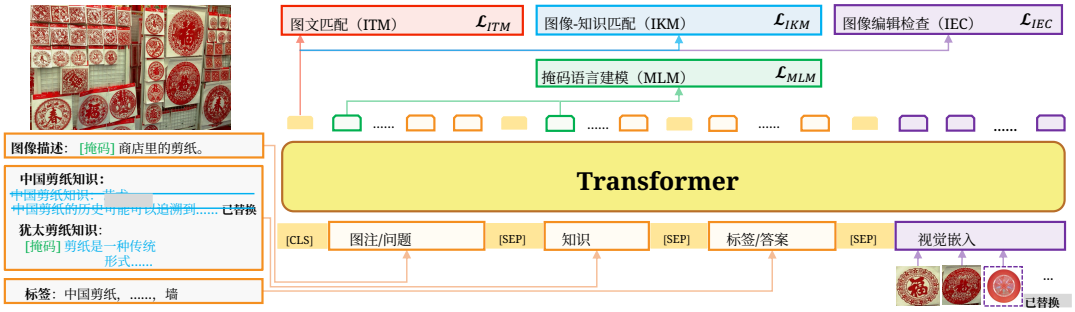


图3。GIVL采用包含四个预训练目标的预训练方法。输入图像涉及视觉概念中国剪纸。输入知识涉及犹太剪纸，而非中国剪纸，但它同样是描述一种与中国剪纸属于相似类别的视觉概念的知识。因此，对于图像-知识匹配（IKM）目标，输入内容属于类型3。此外，视觉概念中国剪纸 被替换为视觉上相似的概念红色飞盘。因此，对于图像编辑检查（IEC）目标，输入内容属于类型2。

用于**c** 和**t**。该任务被表述为一个三分类任务，$y^{c,t} \in \{0,1,2\}$，其中 0 表示**c** 和**t** 均与输入图像**I**匹配；1 表示**t** 与图像**I** 匹配，而**c** 与图像**I**不匹配；2 表示**c** 与**I** 匹配，但**t** 与**I**不匹配。图像-文本匹配损失是关于根据 [CLS]表征预测输入内容类型的概率的交叉熵损失。

3.2.2 图像-知识匹配（IKM）

我们提出图像-知识匹配（IKM），以帮助GIVL学习地理多样化视觉概念的知识。借助IKM，我们鼓励GIVL学习与视觉概念相对应的知识，并发现地理多样化视觉概念之间的联系。

尽管 GIVL 预训练语料库中的地域多样化视觉概念在视觉特征上可能大相径庭，但它们可以归入相似的类别。例如，在图1中，传统的西方与非西方节日的视觉特征各不相同，但这些场景都属于同一节日类别。学习识别类别相似性，可以将不同的视觉概念归入相似类别，并更轻松地推广到对不同地区其他相关概念的理解。另一方面，相似类别中的每个视觉概念都关联着独特知识。因此，GIVL 还必须准确判断输入知识是否与输入图像相符。

为此，我们构建了三种输入内容，并将 IKM 定义为一项三分类任务，以促使 GIVL 识别输入类型：

- 类型 1：**k** 与输入图像**I**相符；
- 类型 2：**k** 与输入图像**I** 不匹配，并且**k** 所描述的视觉概念不属于**NOT** 与 p_v 中**I**的视觉概念相似的类别；
- 类型 3：**k** 与输入图像**I** 不匹配，但**k** 所描述的视觉概念属于与**I**中 p_v 的视觉概念相似的类别。

要为 IKM 中的类型 3 输入选择知识 **k**，我们需要执行两个步骤：(i) 从输入图像 p_v 的图像描述中提取视觉概念**I** 的名称（对于 GQA 数据，参见补充材料）；(ii) 查找属于相似类别的视觉概念。关于从图像描述中提取 p_v 及其类别信息的更多细节将在第3.3节介绍。从图像描述中提取视觉概念名称 p_v 后，为了找到与 p_v 所属类别最相关的视觉概念，我们随机抽取 200 个视觉概念作为候选项。然后，选择类别与 p_v的类别在语义上最相似的候选概念。具体而言，根据候选项类别名称的文本嵌入[6] 与 p_v类别名称的嵌入之间的余弦相似度，对抽取的候选项进行排序。选择最相关视觉概念 p^* 的过程如公式（2）所示，

$$p^* = \arg\max_{p_i} \mathbf{CosineSim}(z_{p_i}, z_{p_v}), \tag{2}$$

其中，z_{p_i} 是第 i个采样视觉概念 p_i的类别嵌入，z_{p_v} 是 p_v的类别嵌入，**余弦相似度**表示用于量化两个嵌入之间余弦相似性的函数。p^* 对应的知识可视为类型 3 输入中的**k** 。

为准备 Type 2 输入内容中的**k**，我们首先为类别名称（$\tau = 0.3$）嵌入之间的余弦相似度设定阈值，以筛选出与 p_v 相关的视觉概念。然后，我们从保留的视觉概念中随机选取一个。所选视觉概念表示与 p_v 的类别信息无关的概念。其对应的知识可用作 Type 2 输入中的**k** 。

[6]我们使用 FastText [2] 在维基百科上预训练的嵌入。短语嵌入是对短语中各个词的嵌入取平均池化得到的。



图4. 挖掘视觉概念类别信息的步骤。中心名词（“大门”，句法树的根节点）与其修饰语（“传统的日本”，与“大门”存在“amod”关系的词）的组合可视为鸟居的类别（“传统的日本大门”）。

IKM损失是一个交叉熵损失，用于衡量基于 [CLS]表征预测输入图像与知识之间关系类型的概率，

$$\mathcal{L}_{IKM} = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\log p(y_i^k|\mathbf{c},\mathbf{k},\mathbf{t},\mathbf{v}), \tag{3}$$

其中 $\mathcal{D}$ 表示整个预训练语料库[7] 而 y_i^k则是 IKM 中输入类型的标签。

3.2.3 图像编辑检查（IEC）

为了更好地区分视觉上相似但无关的概念，我们提出了另一种预训练目标——图像编辑检查（IEC）。在地理多样化设置中，视觉概念很可能具有相似的视觉特征，却属于完全不同的类别。例如，在图2中，中国剪纸呈红色且呈圆形，这与红色飞盘的视觉特征相符。IEC 旨在识别输入图像中的特定视觉概念 p_v 输入图像 **I** 是否被替换成了另一个视觉上相似、但属于无关类别的概念。

我们为 IEC 考虑两类输入内容：

- 类型 1：输入图像 **I**保持不变；
- 类型 2：输入图像 **I** 中的视觉概念p_v 的视觉嵌入被替换为另一个在视觉上相似、但属于与 p_v无关类别的概念的嵌入。

在图 3中，由于视觉概念中国剪纸被替换为红色飞盘，因此输入类型为类型 2[8]。

为准备类型2数据的输入内容，我们需要完成两个步骤：（i）根据图像描述，查找视觉概念 p_v 在输入图像**I** 中对应的检测到的物体；（ii）寻找视觉上相似的概念进行替换。定位视觉概念 p_v 的流程见第3.4节。定位视觉概念 p_v 后，为类型2输入选择合适的替换视觉概念，

[7]IKM 中类型 1、2 和 3 输入的比例为 $\mathcal{D}$ 2 : 1 : 1。[8]IEC 中类型 1 和 2 输入的比例为 $\mathcal{D}$ 1 : 1。

对于类型2输入，我们随机抽取20张图像，然后将抽取图像中所有检测到的物体的视觉嵌入和标签名称收集为候选项。替换视觉概念根据以下两个标准选定：（i）其类别与概念 p_v 的类别信息不同[9]；（ii）其视觉嵌入与p_v的视觉嵌入最接近。我们使用 p_v 选取不相关的视觉概念，以确保这种替换结合图像上下文来看是不合理的。

IEC损失是关于根据输入图像在 [CLS]表征下是否被修改进行预测的概率的二元交叉熵损失，

$$\mathcal{L}_{IEC} = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\log p(y_i^v|\mathbf{c},\mathbf{k},\mathbf{t},\mathbf{v}), \tag{4}$$

其中 y_i^v是 IEC 中输入类型的标签。最终损失 L是上述所有损失的总和：

$$\mathcal{L} = \mathcal{L}_{MLM} + \mathcal{L}_{ITM} + \mathcal{L}_{IKM} + \mathcal{L}_{IEC}. \tag{5}$$

3.3. 获取视觉概念的类别

获取视觉概念的类别是为 IKM 和 IEC 构建 GIVL 输入的前置步骤。我们首先需要从图像描述中提取输入图像 p_v **I** 的视觉概念名称。我们通过使用 [31] 解析图像描述来实现这一点。 p_v 是句法树中中心名词及其修饰语的组合。例如，给定图像描述“商店里的中国剪纸”， p_v 是，它由中心名词“剪纸”及其句法树中的修饰语“中国”组成。

要获取 p_v的类别，我们随后使用关键词 p_v搜索维基百科。如果 p_v 是维基百科中的一个词条，我们会发现，类别信息通常可以在维基百科介绍段落的第一句中找到。如图4所示，鸟居（即传统日本鸟居）的类别出现在第一句：“鸟居……是一种传统的日本鸟居，最常见于……”接着我们注意到，类别名称是由第一句中的中心名词及其修饰语组成的短语。在这个例子中，第一句的中心名词是“鸟居”，修饰词是“传统”和“日本”。最终组合成的“传统日本鸟居”就是的类别。尽管用这些简单启发式方法挖掘出的类别信息并不完善，但这种提取方法易于实现，并且能够高效获取大量视觉概念的类别。

3.4. 在图像中定位视觉概念

在目标类别标签数量有限的情况下，当前的目标检测器很难检测到具有地理多样性的视觉概念

[9]我们使用候选视觉概念类别与 p_v 类别的嵌入之间的余弦相似度。相似度低于 0.3 的候选概念均视为不相似。

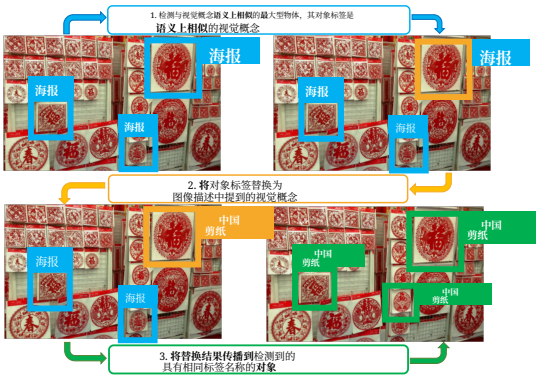


图5. 在输入图像中定位新颖视觉概念的步骤。

。 p_v因此，我们提出了一种简单的方法，以便根据视觉概念高效定位对应的物体 p_v。我们发现，视觉概念 p_v 通常（i）会被归类为与 p_v的类别语义相近的标签名称；（ii）其图像块在图像中占据较大面积。为此，我们根据经验发现设计了一套启发式方法，用于定位新的视觉概念。首先，只考虑每张图像中面积最大的前 10 个检测到的物体。其次，我们计算这些物体的对象标签与 p_v的类别之间的相似度。相似度分数最高的物体将被视为与 p_v对应的物体。 以图5 为例。要定位的视觉概念p_v 是中国剪纸。假设某个中国剪纸（右上角的物体）位于面积最大的前 10 个检测到的物体之中。此外，它原先检测到的对象标签是海报，其与中国剪纸的类别在语义上最为相似。因此，我们可以将其原有的对象标签替换为中国剪纸，因为它正是我们要寻找的对应物体。

上述方法每张图像仅定位一个视觉概念。然而，一张图像中可能包含多个相同的视觉概念。例如，图5中有几个中国剪纸。为解决这一问题，我们只需将中国剪纸的视觉概念名称传播到其他具有相同原始检测标签的对象上。

4. 实验

我们通过两组实验来评估 GIVL。首先，我们在多种地理多样化视觉与语言任务上评估 GIVL，包括零样本图像分类、视觉与语言推理以及图文检索。这有助于验证 GIVL 在地理多样化环境下的有效性。另一方面，我们还在常见的视觉与语言任务上进行实验，以证明 GIVL 预训练方法的泛化能力。

4.1. 消融研究的基线方法

下面介绍五种基线方法。为确保比较公平，预训练语料库、预训练步数和

模型 先前的视觉语言预训练模型	参数量	Acc.	西方/非西方
VinVL*	112M	1.21	1.77/1.01
VinVL [52]	112M	1.29	1.25/1.30
Ours			
GIVL 不使用 $\mathcal{L}_{IKM}$	112M	21.37	25.31/20.37
不含 $\mathcal{L}_{IEC}$ 的 GIVL	112M	12.96	12.71/13.02
使用 CLIP 的 GIVL	199M	18.04	22.89/16.82
GIVL-B	112M	20.35	23.93/19.45
GIVL	112M	**27.25**	**31.65/26.15**

表 1. Dollar Street 数据集上地理多样化零样本图像分类的结果。我们还展示了西方图像和非西方图像各自的表现。

超参数均与 GIVL[10] 完全相同。由于视觉与语言预训练极其耗时，消融研究中的所有基线方法均采用 500K 步进行预训练。

GIVLw/o $\mathcal{L}_{IKM}$ & GIVL w/o $\mathcal{L}_{IEC}$. GIVLw/o $\mathcal{L}_{IKM}$ 和 GIVL w/o $\mathcal{L}_{IEC}$ 分别是未使用图像-知识匹配（IKM）目标和图像编辑检查（IEC）目标进行预训练的模型。我们通过这两个基线方法展示了所提出的预训练目标的有效性。

VinVL*. VinVL* 仅使用掩码语言建模和图文匹配目标进行预训练，如 VinVL [52]。它与 GIVL 使用相同的预训练语料库。GIVL 和 VinVL* 之间唯一的区别在于目标。GIVL 使用图像-知识匹配（IKM）和图像编辑检查（IEC）进行预训练，而 VinVL* 则没有。比较 GIVL 和 VinVL*可以体现引入 IKM 和 IEC 目标对地理多样化视觉与语言任务带来的提升。对于 GIVL 和 VinVL 在常见视觉与语言任务上的预训练方法，这种比较也公平。

采用 CLIP 的 GIVL。 近期的一些视觉语言预训练模型使用 CLIP [32] 作为视觉编码器。我们将 GIVL 中的对象级视觉编码器替换为 CLIP，以检验其能否进一步提升性能。CLIP 提供的是网格级视觉表示，而不是对象级视觉表示。因此，移除了图像编辑检查目标，因为该目标涉及对象级替换。

GIVL-B。 GIVL 与 GIVL-B 之间的唯一区别在于，GIVL-B 的图像-知识匹配目标是二分类目标，而不是三分类。对于图像-知识匹配目标，GIVL-B 需要识别输入知识是否与图像内容相匹配。GIVL-B 无需判断输入知识描述的视觉概念是否与输入图像中的概念属于相似类别。比较 GIVL 和 GIVL-B，可以展示融入类别信息对于学习地理多样化视觉概念知识的作用。

4.2. 地理多样化基准测试结果

地理多样化零样本图像分类。 地理多样化零样本图像分类是一项下游地理

[10]实验设置的详细信息见附录A。

模型	数据/步骤	参数量	Acc.	Δ
先前的视觉语言预训练模型				
ViLBERT [29]	3.3M/-	274M	66.53	10.87
VinVL [52]	5.65M/2M	112M	72.48	8.55
VinVL*	3.17M/500K	112M	69.66	8.27
X-VLM† [51]	16M/-	216M	73.02	11.39
ALBEF† [23]	14M/-	210M	73.17	9.37
METER† [9]	4M/-	352M	73.47	8.86
Ours				
不含 GIVL $\mathcal{L}_{IKM}$	3.17M/500K	112M	72.11	-
不含 GIVL $\mathcal{L}_{IEC}$	3.17M/500K	112M	68.58	-
GIVL 与 CLIP	3.17M/500K	199M	71.78	-
GIVL-B	3.17M/500K	112M	70.26	-
GIVL	3.17M/500K	112M	72.50	**6.56**
GIVL	3.17M/900K	112M	**72.70**	7.17

表 2. MaRVL 测试集上的结果。我们还展示了性能差异 Δ，即 NLVR2 与 MaRVL 之间的差异。† 表示报告于 [54]。

模型	参数量	Acc.	非西方	Δ
先前的视觉语言预训练模型				
VisualBERT [29]	135M	53.95	-	10.42
ViLBERT [29]	274M	59.99	-	7.28
VinVL*	112M	69.07	66.45	8.46
VinVL [52]	112M	70.20	66.78	11.04
Ours				
不含 $\mathcal{L}_{IKM}$ 的 GIVL	112M	69.56	65.96	-
不含 $\mathcal{L}_{IEC}$ 的 GIVL	112M	69.89	66.92	-
GIVL（使用 CLIP）	199M	70.43	67.25	10.20
GIVL-B	112M	69.56	65.96	-
GIVL	112M	70.32	68.41	6.14
GIVL（100 万）	112M	**72.01**	**70.4**	**4.97**

表 3. GD-VCR 上的结果。我们还展示了 GD-VCR 中所有非西方图像上的结果，以及西方与非西方图像之间的差异 Δ。

多样化的视觉与语言任务，可直接评估预训练方法的有效性。我们在 Dollar Street 数据集上评估模型[11]。该数据集标注了 127 个类别，每个类别都包含来自世界各地的图像。对于单张图像的分类，我们构建 127 个输入，每个输入由一个类别名称、该类别对应的知识[12]，标签名称以及检测到的对象的视觉嵌入拼接而成。我们通过图文匹配目标，比较所有 127 个类别名称与输入图像匹配的预测概率。概率最高的类别即为最终分类结果。

如表1所示，GIVL 的表现显著优于 VinVL 和 VinVL*，优势约为 26%。在消融研究中，GIVL 的表现提升了 6%–20%，证明了所提出的 IKM 和 IEC 目标的有效性。我们还发现，GIVL 的表现优于采用强大视觉编码器的 GIVL w/ CLIP。这进一步证明，物体级视觉表示和物体级预训练目标 IEC 有助于学习地理多样化视觉概念。

多元文化视觉推理（MaRVL）。遵循 NLVR2 [39]，MaRVL [28] 是一项视觉与语言任务，要求模型判断一个句子是否准确描述了两张输入图像的内容。MaRVL 图像涵盖了非西方地区多样的视觉概念。由于 MaRVL[13] 仅为测试集，因此遵循 [28]，在 NLVR2 训练集上微调模型，然后根据 NLVR2 开发集选择最佳检查点，以评估模型在 MaRVL 上的表现。

从表 2可以看出，GIVL 的表现优于分别在未采用我们提出的 IKM 和 IEC 目标的情况下进行预训练的消融基线方法。此外，与在 Dollar Street 数据集上的观察结果类似，与 VinVL 相比*，在与 GIVL 相同的语料库上进行预训练的 VinVL，GIVL 的表现更优。这进一步表明，GIVL 的预训练目标有助于视觉语言预训练模型比 VinVL 更好地学习地理多样化的视觉概念。

我们还将 GIVL 与以下视觉语言预训练模型进行了比较：(i) 3×规模更大的模型（METER）；以及 (ii) 在 $2-5\times$ 更大语料库上进行预训练的模型（VinVL、X-VLM 和 ALBEF）。GIVL 使用少得多的数据和更小的模型规模，取得了有竞争力的表现。此外，我们特别关注 GIVL 在 NLVR2 与 MaRVL 上的表现对比。[28] 结果表明，NLVR2 数据集中的视觉概念以西方为中心。NLVR2 与 MaRVL 之间的性能差距越小，意味着对非西方地区的偏见越小。我们观察到，GIVL 在两个数据集上都能取得更加均衡的表现，而包括 METER、X-VLM 和 ALBEF 在内的其他视觉语言预训练模型则存在更大的性能差异。

地理多样化视觉常识推理（GD- VCR）。GD-VCR 是一个用于评估多模态模型理解图像中地理多样性常识知识能力的测试集。这是一项需要进行地理多样化视觉常识推理的多项选择问答任务。我们在 VCR [50]训练集上对模型进行微调，并选择 VCR 开发集上的最佳检查点，在 GD-VCR 上进行评估。

如表 3所示，GIVL 的表现优于所有使用数量相近的图像进行训练、规模相近的以往视觉语言预训练模型。GIVL 的表现也优于除 GIVL w/ CLIP 之外的所有消融基线方法；后者使用了更强大的视觉编码器，但仅取得了 0.1% 的细微提升。此外，我们强调了 GD-VCR 中西方数据与非西方数据之间的性能差距。GIVL 的性能差距显著小于任何消融基线方法。尽管 GIVL w/ CLIP 相较于 GIVL 略有提升，但 GIVL 的性能差距比 GIVL w/ CLIP 小 4.06%。

维基百科图像-文本检索（WIT）。WIT 图像-文本检索是地理多样化数据集上的标准检索任务

[11]Dollar Street 中的图像均标注了国家信息。用于划分西方国家和非西方国家的代理指标基于 [16]。[12]各类别的知识来源于维基百科和 Wordhoard。

[13]我们使用了 MaRVL 数据集的英文翻译版本，见 [54]。

模型	Data	参数量	I/R	T/R
先前的视觉语言预训练模型				
LXMERT [41]	-	240M	14.28	14.86
VisualBERT [24]	-	135M	15.36	15.75
UNITER [4]	-	-	15.43	16.01
VL-BERT [38]	3.3M	-	15.11	16.09
ViLBERT [29]	3.3M	274M	15.40	16.93
VinVL [52]	5.65M	112M	27.78	28.65
VinVL*	3.17M	112M	25.44	25.50
Ours				
不含 $\mathcal{L}$ 的 GIVL	3.17M	112M	26.21	26.97
不含 $\mathcal{L}_{IEC}$ 的 GIVL	3.17M	112M	28.08	28.18
使用 CLIP 的 GIVL	3.17M	199M	27.94	28.17
GIVL-B	3.17M	112M	29.97	29.86
GIVL	3.17M	112M	28.00	28.79
GIVL（1M）	3.17M	112M	**29.98**	**30.79**

表 4. WIT 图像文本检索任务的结果。I/R 和 T/R 分别表示图像检索和文本检索。评估指标为 Recall@1。1M 表示预训练步数。

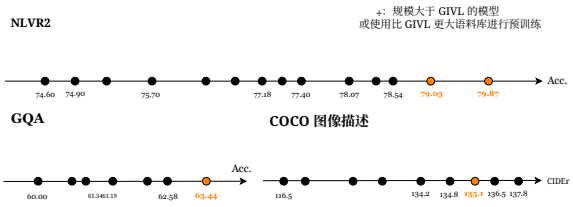


图6. GIVL在常见视觉与语言任务上的表现。完整结果见附录B。

维基百科图像[14]。表4 显示，除 GIVL-B 外，GIVL 的性能均优于基线方法。经过 100 万步预训练后，GIVL 在 WIT 图像-文本检索任务上取得了 SOTA 性能。

4.3. 常见视觉与语言基准测试的结果

除了在地理多样化视觉与语言任务上测试GIVL外，我们还在常见视觉与语言任务上对GIVL进行基准测试，以考察GIVL的预训练方法在现有视觉语言预训练模型中是否具有竞争力。我们并不预期GIVL在这些视觉与语言基准测试中胜过所有最先进的视觉语言预训练模型，因为这些基准测试使用以西方为中心的数据进行标注，而最先进方法也使用规模大得多的类似数据进行训练。我们的目标是回答两个问题。问题一：*GIVL*能否达到与使用规模相近的数据进行预训练的视觉语言预训练模型相当的表现？问题二：使用相同语料库进行预训练时，*GIVL*能否达到与最先进的视觉语言预训练模型同样出色的表现？

为回答问题一，我们在常见的视觉与语言基准测试上评估 GIVL，包括 NLVR2、GQA 和 COCO 图像描-述。在 NLVR2 上，GIVL 能够胜过 11 个参数规模大得多、且使用更多数据进行预训练的视觉语言预训练模型。在 GQA 上，GIVL 的表现优于大多数视觉语言预训练模型。在 COCO 图像描述任务上，其表现甚至接近 SimVLM-base；后者是一个使用 18 亿张图像进行预训练的视觉语言预训练模型。总体而言，尽管 GIVL 使用的语料库

[14]我们使用经过翻译的英文WIT检索数据，见 [3]。



图7. GIVL和VinVL在与地理多样化类别相关的非西方图像和西方图像上的表现。

的领域与常见的视觉与语言基准测试并不相似，它仍能取得有竞争力的结果。这证明了 GIVL 预训练方法的有效性。

针对问题二，我们以 VinVL 为目标。VinVL 是一种强大的视觉语言预训练模型，曾在多个视觉与语言任务的排行榜上名列前茅。为保证比较公平，我们使用 GIVL 预训练语料库复现了 VinVL 的预训练过程。如第4.1节所述，我们将复现得到的预训练模型记为 VinVL*。在上述三个视觉与语言数据集上，GIVL 与 VinVL* 之间的性能差异很小。我们认为，如果使用 VinVL 的语料库进行预训练，GIVL 在常见视觉与语言基准测试上也能达到与 VinVL 同样出色的表现。

4.4. 地理多样化类别的定性研究

我们展示了 GD-VCR 和 Dollar Street 数据集中的示例，以更好地说明 GIVL 的优势。如图 7所示，非西方节日、仆人和宗教与西方地区的情况截然不同。我们发现，在 GD-VCR 中，GIVL 在涉及这些类别的图像上的性能差距明显小于 VinVL。此外，GIVL 在非西方图像上的表现比 VinVL 高 5–8%。对于 Dollar Street 数据集，尽管 GIVL 的整体表现约为 30%，但在识别蔬菜和晾晒衣物时，准确率可达到 55% 以上；这些类别在全球各地的数据中差异很大。GIVL 在这些类别上的表现甚至比 VinVL 高出 50%。GIVL 在这些地理多样化程度很高的类别上的出色表现，进一步证明了其有效性。

5. 结论

我们提出了 GIVL，这是一种地理包容性视觉语言预训练模型。GIVL 在多个地理多样化视觉与语言任务上取得了出色且更加均衡的结果，在常见的视觉与语言任务上也表现出色。通过提出 GIVL，我们呼吁研究人员设计能够进一步提升视觉语言预训练模型地理包容性的方法，并让其应用惠及所有人。

参考文献

[1]Peter Anderson、Xiaodong He、Chris Buehler、Damien Teney、Mark Johnson、Stephen Gould 和 Lei Zhang。用于图像描述与视觉问答的自下而上和自上而下注意力。载于《IEEE计算机视觉与模式识别会议论文集》，第6077–6086页，2018年。12，13[2] Piotr Bojanowski、Edouard Grave、Armand Joulin 和 Tomas Mikolov。利用子词信息丰富词向量。《计算语言学协会汇刊》，5：135–146，2017年。4[3] Emanuele Bugliarello、Fangyu Liu、Jonas Pfeiffer、Siva Reddy、Desmond Elliott、Edoardo Maria Ponti 和 Ivan Vuli´c。IGLUE：跨模态、任务和语言的迁移学习基准。ICML，2022年。3，8，12[4] Yen-Chun Chen、Linjie Li、Licheng Yu、Ahmed El Kholy、Faisal Ahmed、Zhe Gan、Yu Cheng 和 Jingjing Liu。UNITER：通用图文表征学习。载于《欧洲计算机视觉会议论文集》，第104–120页。施普林格，2020年。2，8，12[5] Jaemin Cho、Jie Lei、Hao Tan 和 Mohit Bansal。通过文本生成统一视觉与语言任务。载于《国际机器学习会议论文集》，第1931–1942页。PMLR，2021年。12，13[6] Kyunghyun Cho、Bart van Merri¨enboer、Caglar Gulcehre、Dzmitry Bahdanau、Fethi Bougares、Holger Schwenk 和 Yoshua Bengio。使用 RNN 编码器—解码器学习用于统计机器翻译的短语表征。载于《2014年自然语言处理实证方法会议论文集（EMNLP）》，第1724–1734页，多哈，卡塔尔，2014年10月。计算语言学协会。12[7] Terrance de Vries、Ishan Misra、Changhan Wang 和 Laurens van der Maaten。物体识别能惠及所有人吗？载于《IEEE/CVF计算机视觉与模式识别研讨会论文集》，第52–59页，2019年。3[8] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向 Transformers 预训练。载于《2019年北美计算语言学协会分会年会：人类语言技术论文集》第一卷（长篇与短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州，2019年6月。计算语言学协会。2[9] Zi-Yi Dou、Yichong Xu、Zhe Gan、Jianfeng Wang、Shuohang Wang、Lijuan Wang、Chenguang Zhu、Pengchuan Zhang、Lu Yuan、Nanyun Peng 等。端到端视觉与语言变换器训练的实证研究。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第18166–18176页，2022年。1，2，7[10] Zhe Gan、Yen-Chun Chen、Linjie Li、Chen Zhu、Yu Cheng 和 Jingjing Liu。用于视觉与语言表征学习的大规模对抗训练。《神经信息处理系统进展》，33：6616–6628，2020年。12[11] Feng Gao、Qing Ping、Govind Thattai、Aishwarya Reganti、Ying Nian Wu 和 Prem Natarajan。变换-检索-生成：以自然语言为中心、利用外部知识的视觉问答。在《IEEE/CVF计算机视觉与模式识别会议论文集》中，第5067–5077页，2022年。1[12] Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra和Devi Parikh。让VQA中的V发挥作用：提升图像理解在视觉问答中的作用。载于《计算机视觉与模式识别会议》（CVPR），2017年。1, 2[13] Zhicheng Huang、Zhaoyang Zeng、Yupan Huang、Bei Liu、Dongmei Fu和Jianlong Fu。跳出框架看问题：面向视觉语言表征学习的端到端预训练。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第12976–12985页，2021年。12[14] Zhicheng Huang、Zhaoyang Zeng、Bei Liu、Dongmei Fu和Jianlong Fu。Pixel-BERT：通过深度多模态变换器对齐图像像素与文本。ar Xiv预印本 arXiv:2004.00849，2020年。12[15] Drew A Hudson和Christopher D Manning。GQA：用于真实世界视觉推理和组合式问答的新数据集。《计算机视觉与模式识别会议》（CVPR），2019年。1, 2[16] Samuel P Huntington。文明冲突？载于《文化与政治》，第99–118页。施普林格，2000年。7[17] Panagiotis G Ipeirotis。Mechanical Turk的人口统计特征。2010年。2[18] Aishwarya Kamath、Mannat Singh、Yann LeCun、Gabriel Synnaeve、Ishan Misra和Nicolas Carion。MDETR：用于端到端多模态理解的调制检测。载于《IEEE/CVF国际计算机视觉会议论文集》，第1780–1790页，2021年。12[19] Andrej Karpathy和Li Fei-Fei。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，2015年。1[20] Wonjae Kim、Bokyung Son和Ildoo Kim。ViLT：无需卷积或区域监督的视觉-语言变换器。载于《国际机器学习会议论文集》，第5583–5594页。PMLR，2021年。12[21] Alina Kuznetsova、Hassan Rom、Neil Alldrin、Jasper Uijlings、Ivan Krasin、Jordi Pont-Tuset、Shahab Kamali、Stefan Popov、Matteo Malloci、Alexander Kolesnikov等。Open Images数据集V4。《国际计算机视觉期刊》，128(7):1956–1981，2020年。1, 2[22] Gen Li、Nan Duan、Yuejian Fang、Ming Gong和Daxin Jiang。UniCoder-VL：通过跨模态预训练实现视觉与语言通用编码器。载于《AAAI人工智能会议论文集》，第34卷，第11336–11344页，2020年。1, 2[23] Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong和Steven Chu Hong Hoi。先对齐再融合：通过动量蒸馏进行视觉与语言表征学习。《神经信息处理系统进展》，34:9694–9705，2021年。1, 2, 7[24] Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh和Kai-Wei Chang。VisualBERT：一种简洁的视觉与语言基线。

arXiv预印本 arXiv:1908.03557，2019年。1, 2, 8[25] Wei Li、Can Gao、Guocheng Niu、Xinyan Xiao、Hao Liu、Jiachen Liu、Hua Wu和Haifeng Wang。Unimo：通过跨模态对比学习实现统一模态理解与生成。arXiv预印本 arXiv:2012 15409，2020 年 13 . . [26] Xiujun Li 、 Xi Yin 、Chunyuan Li、Pengchuan Zhang、Xiaowei Hu、Lei Zhang、Lijuan Wang、Houdong Hu、Li Dong、Furu Wei等。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论文集》，第121–137页。施普林格，2020年。2, 12, 13[27] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar和C Lawrence Zitnick。Microsoft COCO：情境中的常见物体。载于《欧洲计算机视觉会议论文集》，第740–755页。施普林格，2014年。1, 2, 3[28] Fangyu Liu、Emanuele Bugliarello、Edoardo Maria Ponti、Siva Reddy、Nigel Collier和Desmond Elliott。跨语言与文化的视觉基础推理。载于《2021年自然语言处理实证方法会议论文集》，第10467–10485页，线上及多米尼加共和国蓬塔卡纳，2021年11月。计算语言学协会。2, 3, 7[29] Jiasen Lu、Dhruv Batra、Devi Parikh和Stefan Lee。ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表征。《神经信息处理系统进展》，32，2019年。1, 2, 7, 8, 12[30] Matthew E. Peters、Mark Neumann、Mohit Iyyer、Matt Gardner、Christopher Clark、Kenton Lee和Luke Zettlemoyer。深度上下文化词表示。载于《2018年计算语言学协会北美分会年会论文集：人类语言技术》，第1卷（长篇论文），第2227–2237页，新奥尔良，路易斯安那州，2018年6月。计算语言学协会。2[31] Peng Qi、Yuhao Zhang、Yuhui Zhang、Jason Bolton和Christopher D. Manning。Stanza：适用于多种人类语言的Python自然语言处理工具包。载于《第58届计算语言学协会年会论文集：系统演示》，第101–108页，线上，2020年7月。计算语言学协会。5[32] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever等。语言模型是无监督的多任务学习者。OpenAI博客，1(8):9，2019年。2, 6[33] Shreya Shankar、Yoni Halpern、Eric Breck、James Atwood、Jimbo Wilson和D Sculley。没有表征就没有分类：评估发展中世界开放数据集中的地理多样性问题。arXiv预印本 arXiv:1711.08536，2017年。3[34] Piyush Sharma、Nan Ding、Sebastian Goodman和Radu Soricut。Conceptual Captions：经过清洗并添加上位词的图像替代文本数据集，用于自动图像描述。载于《第56届计算语言学协会年会论文集》（第1卷：长篇论文），第2556–2565，澳大利亚墨尔本，2018年7月。计算语言学协会。 2[35] Sheng Shen、Liunian Harold Li、Hao Tan、Mohit Bansal、Anna Rohrbach、Kai-Wei Chang、Zhewei Yao 和 Kurt Keutzer。CLIP 能在多大程度上助益视觉与语言任务？arXiv 预印本 arXiv:2107.06383，2021。2，12，13[36]Mustafa Shukor、Guillaume Couairon 和 Matthieu Cord。利用视觉概念和层级对齐进行高效的视觉语言预训练。arXiv 预印本 arXiv:2208.13628，2022。12[37] Krishna Srinivasan、Karthik Raman、Jiecao Chen、Michael Bendersky 和 Marc Najork。WIT：基于维基百科的多模态多语言机器学习图像文本数据集。载于第44届国际 ACM SIGIR 信息检索研究与发展会议论文集，第2443–2449页，2021。2，3[38] Weijie Su、Xizhou Zhu、Yue Cao、Bin Li、Lewei Lu、Furu Wei 和 Jifeng Dai。VL-BERT：通用视觉语言表征的预训练。载于国际学习表征会议，2019。8[39] Alane Suhr、Stephanie Zhou、Ally Zhang、Iris Zhang、Huajun Bai 和 Yoav Artzi。基于照片的自然语言推理语料库。载于第57届计算语言学协会年会论文集，第6418–6428页，意大利佛罗伦萨，2019年7月。计算语言学协会。2，7[40] Ilya Sutskever、Oriol Vinyals 和 Quoc V Le。利用神经网络进行序列到序列学习。神经信息处理系统进展，27，2014。12[41] Hao Tan 和 Mohit Bansal。LXMERT：从 Transformers 中学习跨模态编码器表征。arXiv 预印本 arXiv:1908.07490，2019。2，8，12[42] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。神经信息处理系统进展，30，2017。2[43] Angelina Wang、Alexander Liu、Ryan Zhang、Anat Kleiman、Leslie Kim、Dora Zhao、Iroha Shirai、Arvind Narayanan 和 Olga Russakovsky。REVISE：一种用于测量和缓解视觉数据集偏见的工具。《国际计算机视觉杂志》，第1–21页，2022。3[44] Teng Wang、Wenhao Jiang、Zhichao Lu、Feng Zheng、Ran Cheng、Chengguo Yin 和 Ping Luo。VLMixer：通过跨模态 CutMix 进行非配对视觉语言预训练。载于国际机器学习会议论文集，第22680–22690页。PMLR，2022。12[45] Zirui Wang、Jiahui Yu、Adams Wei Yu、Zihang Dai、Yulia Tsvetkov 和 Yuan Cao。SimVLM：利用弱监督进行简单的视觉语言模型预训练。载于国际学习表征会议，2021。13[46] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、R´emi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander M. Rush。Transformers：最先进的自然语言处理。在

2020年自然语言处理实证方法会议：系统演示论文集，第38–45页，线上，2020年10月。计算语言学协会。12[47] Da Yin、Hritik Bansal、Masoud Monajatipoor、Liu-nian Harold Li 和 Kai-Wei Chang。Geomlama：针对多语言预训练语言模型的地理多样化常识探测。arXiv 预印本 arXiv:2205.12247，2022。3[48] Da Yin、Liunian Harold Li、Ziniu Hu、Nanyun Peng 和 Kai-Wei Chang。拓宽视觉视野：地理多样化视觉常识推理。载于2021年自然语言处理实证方法会议论文集，第2115–2129页，线上及多米尼加共和国蓬塔卡纳，2021年11月。计算语言学协会。1，2，3[49] Peter Young、Alice Lai、Micah Hodosh 和 Julia Hockenmaier。从图像描述到视觉指称：用于对事件描述进行语义推理的新型相似度度量。《计算语言学协会会刊》，2：67–78，2014。2，3[50] Rowan Zellers、Yonatan Bisk、Ali Farhadi 和 Yejin Choi。从识别到认知：视觉常识推理。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第6720 6731页，2019 7 – 。[51] Yan Zeng、Xinsong Zhang 和 Hang Li。多粒度视觉语言预训练：将文本与视觉概念对齐。arXiv 预印本 arXiv:2111.08276，2021。7[52] Pengchuan Zhang、Xiujun Li、Xiaowei Hu、Jianwei Yang、Lei Zhang、Lijuan Wang、Yejin Choi 和 Jianfeng Gao。VinVL：重新审视视觉语言模型中的视觉表征。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第5579–5588页，2021。1，2，3，6，7，8，12，13[53] Luowei Zhou、Hamid Palangi、Lei Zhang、Houdong Hu、Jason Corso 和 Jianfeng Gao。用于图像描述和视觉问答的统一视觉语言预训练。载于 AAAI人工智能会议论文集，第34卷，第13041–13049页，2020。13[54] Wangchunshu Zhou、Yan Zeng、Shizhe Diao 和 Xinsong Zhang。VLUE：用于评估视觉语言模型的多任务基准。arXiv 预印本 arXiv:2205.15237，2022。7

附录

A. 下游视觉与语言任务的预训练与微调细节

A.1. 预训练设置

GIVL 使用 BERT 基础模型的预训练参数进行初始化 [46]。模型最多进行 1M 步预训练，批量大小为 720。学习率为 $1e-4$，并采用线性衰减。输入文本和视觉对象中的最大令牌数分别为 70 和 50。GIVL 及消融基线方法的所有预训练实验均使用 8 个配备 40GB GPU 内存的 A100 GPU 完成。

A.2. 微调设置

MaRVL 和 NLVR2。我们在 NLVR2 上对 GIVL 进行 20 个周期的微调，批量大小为 72，学习率为 $3e-5$。输入文本和视觉对象中的最大令牌数为 55。由于每个样本包含两张输入图像，我们在模型输入中最多包含 80 个视觉对象，其中每张图像最多包含 40 个输入视觉对象。如第 4.2 节所述，由于 MaRVL 是一个遵循 NLVR2 构建方式的测试集，其微调结果基于此处讨论的微调方法。

GD-VCR。我们在原始 VCR 数据集上对 GIVL 进行微调，共训练 5 个周期，批次大小为 128，学习率为 $3e-5$。对于模型输入，我们将问题和四个答案选项拼接在一起，并加入输入图像的视觉嵌入。词元和视觉对象的最大数量分别为 100 和 50。

WIT 图像文本检索。我们在 WIT 图像文本检索训练集上对 GIVL 进行微调，共训练 20 个周期，批次大小为 128，学习率为 $2e-5$。输入文本中的词元和视觉对象的最大数量分别为 70 和 70。我们使用 IGLUE 提供的翻译后英文训练集和开发集 [3]。

COCO 图像描述。我们在 COCO 图像描述数据集上对 GIVL 进行微调，共训练 60 个周期，批次大小为 256，学习率为 $3e-5$，并采用序列到序列目标 [6, 40]。输入文本中的词元和视觉对象的最大数量分别为 70 和 50。随后，我们使用 CIDEr 指标进一步优化 GIVL 75 个周期，批次大小为 64，学习率为 $2e-6$。我们使用束搜索（束宽为 5 [1]）对生成结果进行采样，生成图像描述的最大长度为 20 个词。

GQA。我们在 GQA 上对 GIVL 进行微调，共训练 5 个 epoch，批量大小为 128，学习率为 $5e-5$。输入文本和视觉对象的最大 token 数分别为 165 和 45。

B. 常见视觉与语言任务的详细结果

如第4节所述，我们还针对常见的视觉-语言（V&L）任务开展了实验。表5、6 和7 分别展示了 GQA、NLVR2 和 COCO图像描述的详细实验结果。

在表5中，我们展示了 GIVL 在 GQA 上的表现优于许多先前的视觉-语言预训练模型（VLP）。值得强调的是，与大多数先前的 VLP 相比，GIVL 使用的数据显著更少，同时参数量也更少。为确保公平比较，VinVL* 使用与 GIVL 相同的预训练数据。

模型 先前的视觉语言预训练模型	参数量	Data	Acc.
LXMERT [41]	240M	-	60.00
Oscar [26]	-	4.1M	61.19
CLIP-ViL [35]	178M	-	61.34
MDETR [18]	-	-	62.48
VinVL* [52]	112M	3.17M	62.58
Ours			
GIVL	112M	3.17M	**63.44**

表5. GQA 测试开发集上的结果。

我们还在 NLVR2 数据集上评估了所提出的 GIVL。与 GQA 的结果类似，根据表 6，GIVL 的表现也优于所有列出的先前视觉语言预训练模型，且所需的预训练数据更少，模型规模更小。

模型 先前的视觉语言预训练模型	参数量	Data	Acc.
VL-T5 [5]	224M	-	74.60
LXMERT [41]	240M	-	74.90
VLMixer [44]	-	4M	75.28
ViLT [20]	87M	4M	75.70
PixelBERT [14]	114M	-	76.73
SOHO [13]	-	-	76.37
UNITER [4]	300M	4M	77.18
ViCHA [36]	-	-	77.27
ViLBERT [29]	274M	3.3M	77.40
Oscar [26]	-	4.1M	78.07
VILLA [10]	-	4M	78.39
VinVL* [52]	112M	3.17M	78.54
Ours			
GIVL	112M	317万	**79.03**
GIVL（90万）	112M	317万	**79.87**

表6。NLVR2 测试开发集上的结果。

图像描述是一项经典任务，可用于评估视觉语言预训练模型的性能。如表 7所示，GIVL 在不同的评估指标上表现与先前的视觉语言预训练模型相当。

模型	参数量	Data	BLEU@4	CIDEr	METEOR	SPICE
先前的视觉语言预训练模型						
VL-T5 [5]	224M	-	-	116.5	-	-
自下而上自上而下 [1]	-	-	36.3	120.1	27.7	21.4
视觉语言预训练 [53]	-	-	39.5	129.8	29.3	22.4
Unimo-Large [25]	300M	-	39.6	127.7	-	-
Oscar [26]	-	4.1M	40.5	137.6	29.7	22.8
CLIP-ViL [35]	178M	-	40.2	134.2	29.7	23.8
SimVLM-base [45]	-	1.8B	39	134.8	32.9	24
VinVL* [52]	112M	3.17M	39.6	136.5	30.4	24.4
Ours						
GIVL	112M	3.17M	39.6	135.1	30.3	24.3

表7. COCO图像描述结果。



图8. VQAv2 与 GD-VCR 的图像及相应问答对的对比。

大多数先前用于图像描述的视觉语言预训练模型使用的数据都远多于 GIVL，例如 SimVLM-base。以上三项实验均证明了 GIVL 的有效性和数据效率。

C. 定性示例

C.1. 常见与地理多样化视觉与语言任务

由于计算机视觉（CV）领域对地理多样化视觉-语言（V&L）任务的研究尚不广泛，读者可能难以直观地理解常见视觉与语言任务与地理多样化视觉与语言任务之间的区别。本节将通过一些示例来说明这一点。

在讨论示例之前，我们先介绍地理多样化视觉与语言任务的设定。首先，地理多样化视觉与语言任务（如 GD-VCR）只使用从不同地区和文化收集的图像。这确保图像背后的视觉概念与其所属地区和文化背景高度相关。其次，这些地理多样性数据集要求来自不同地区和文化的标注人员对数据进行标注，从而进一步赋予数据集地理多样性。第三，也是最重要的一点，地理多样性数据集中的问题或文本描述



图 9. NLVR2 与 MaRVL 的图像和陈述对比。

更关注来自不同地区和文化的视觉概念及其相关知识。

图8 展示了来自 VQA 和 GD-VCR 数据集的一些图像—问题对。VQA 数据集包含询问通用视觉概念的问题，例如颜色、天气、大小、等等。仅凭 VQA 数据集图像中的视觉信息就足以回答这些问题。另一方面，GD-VCR 提出的问题需要了解世界各地不同地区和文化的背景知识。例如，右侧的第一个示例描述了一个人在繁忙街道上准备早餐的场景。这在大多数西方国家并不常见，但在东亚和南亚地区的大多数地方却很常见。

C.2. GIVL 性能的实证分析

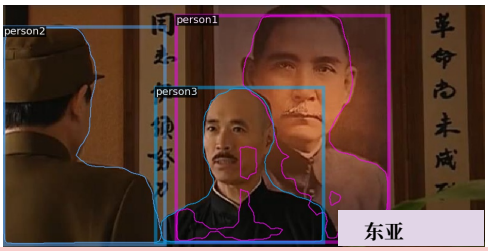
VQA 与 GD-VCR 的比较也能说明，为什么 GIVL 在常见视觉与语言任务上的表现与其他最先进方法相近，却能在地理多样性任务上大幅超越所有基线模型。对于常见的视觉与语言任务，尽管有些图像是在世界各地采集的，但它们并不具备地理多样性。无论图像中是否存在地理多样性因素，这些任务都只涉及常见的视觉概念及其基础视觉信息。例如，如图8所示，VQA 示例中的第二组图像与问题仅询问图像中大象的体型信息。但这个问题并没有询问与热带视觉概念隐含对应的知识。因此，在常见的视觉与语言任务上，GIVL 可能无法胜过那些使用规模大得多、且主要涵盖常见视觉概念的视觉与语言预训练语料库进行预训练的视觉语言预训练模型。

另一方面，GD-VCR 和 MaRVL 等地理多样化视觉与语言任务要求模型运用与图像所涉地区和文化背景相关的知识来完成任务。如图9右侧所示，模型需要识别地理多样化视觉概念，并利用图像内容之外的文化知识进行预测。由于先前的视觉语言预训练模型并未经过预训练来理解地理多样化视觉概念所蕴含的知识，GIVL 在地理多样化视觉与语言任务上的表现能够优于大多数最先进的视觉语言预训练模型。

C.3. GD-VCR 和 MaRVL 的案例研究

我们在图10中展示了 GD-VCR 的一些案例，以及 GIVL 和 VinVL 给出的预测。对于 GD-VCR 中的一些案例，VinVL 无法给出正确答案，而 GIVL 却能答对。在展示的大多数案例中，VinVL 的预测都不合情理。这些案例与文化背景密切相关，例如右下角的例子。图中人物身穿中国古代王室服饰。他们的姿势看起来像是在列队并半蹲。在中国古代，这是一种王室谢罪礼仪。图11展示了更多 MaRVL 案例，以及 GIVL 和 VinVL 给出的预测。

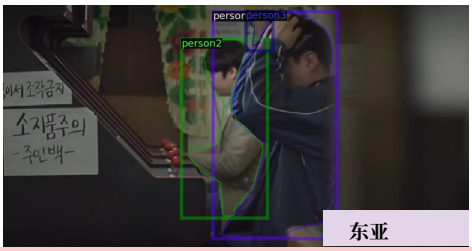
GD-VCR



问题：[人2] 和[人3] 在[illegible]？

VinVL：他的一位客户，曾为扬基队效力。

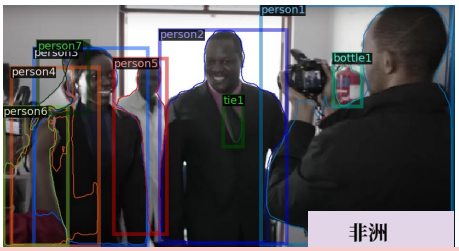
GIVL：他们在谈论战争。



问题：在哪里[人3]？

VinVL：在餐厅的柜台旁。

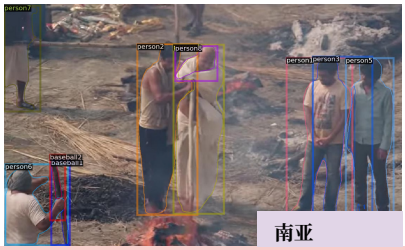
GIVL：[illegible]游戏室里。



问题：为什么[这位[illegible]这里?

VinVL：[illegible]物2] 正在参与恩典活动。

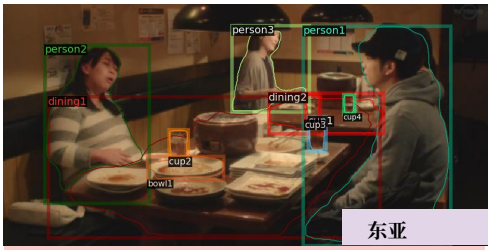
GIVL：[illegible]物2] 前来进行检查。



问题：[这位人物8] 在[illegible]

VinVL：.....正在给[illegible]物8] 一些鼓励。

GIVL：[illegible]物8] 正在火化遗体。



问题：这个人可能[这个[illegible]如何？为什么？

VinVL：[illegible]] 现在不太饿。

GIVL：[illegible]] 看起来很饱，因为[illegible]吃了太多肉。



问题：这个人[这个人[illegible]][这个人1]？

VinVL：[illegible]] 正在考虑点什么。

GIVL：[illegible]1] 想道歉。

图10。GD-VCR 案例研究。



图11. MaRVL 案例研究。