arXiv:2303.05313v2 [cs.CV] 6 May 2023

面向细粒度多模态预训练的精细化视觉-语言建模

Lisai Zhang[1], Qingcai Chen[1], Zhijian Chen[2],Zhonghua Li[2],Yunpeng Han[1],Zhao Cao[2]

[1]哈尔滨工业大学（深圳）， [2]华为技术有限公司

摘要

基于物体标注的细粒度监督已广泛应用于视觉语言预训练（VLP）。然而，在现实应用场景中，对齐的多模态数据通常采用图像-字幕格式，只提供粗粒度监督。收集物体标注并针对不同场景构建物体标注预提取器，不仅成本高昂，也十分耗费计算资源。本文从语言学角度提出一种无需物体标注的细粒度视觉-语言预训练方案。首先，我们提出同音词句子改写（HSR）算法，以提供词元级监督。该算法将字幕中的动词、名词、形容词或量词替换为 WordNet 中的同音词。相应地，我们提出精细化视觉-语言建模（RVLM）框架，以利用词元级监督。为学习细粒度对齐，我们提出了三项精细化任务，即精细化图文对比（RITC）、精细化图文匹配（RITM）和替换语言建模（RLM）。在多个下游任务上的大量实验表明，所提方法性能优越。

1 引言

近年来，视觉与语言预训练（VLP）极大地增强了机器理解多模态数据的能力。VLP旨在像人类一样理解多模态语义，使其能够识别图像物体与词元之间的细粒度对齐。然而，在真实世界的应用场景中，对齐的多模态数据通常采用粗粒度的图像-字幕格式。如何从粗粒度监督中学习细粒度对齐是一项重要挑战。

大多数现有的细粒度视觉-语言预训练方法（Zeng、Zhang 和 Li，2021年；Li 等人，2022b）都通过图 1(a) 所示的对象标签或对象短语式监督，学习跨模态语义对齐。由于预训练语料库中没有这类细粒度标注，一种常见做法是使用 Faster-RCNN（Ren 等人，2015）等现成的对象检测模型，提取物体区域特征及其对应的标签（Su 等人，2019年；Qi 等人，2020年；Li 等人，2020b；Zhang 等人，2021年）。随后，将物体特征作为视觉词元输入 Transformer（Vaswani 等人（2017）），并采用掩码语言建模（MLM）等自监督任务

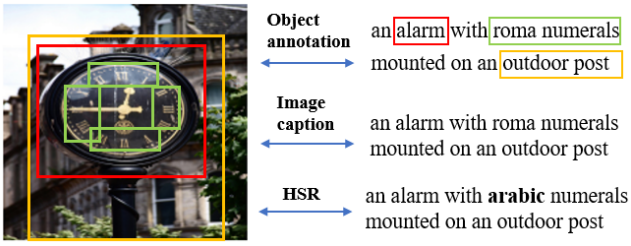


图1：对象标注监督、图像字幕监督以及我们提出的同音异义句子改写（HSR）监督方案示意图。结合真实文本，HSR无需人工标注即可提供词元级监督。

和图文匹配（ITM）来学习多模态表征。近期的一种方法（Li 等人，2022b）利用场景图解析器（Anderson 等人，2016年），将对象标签标注扩展为多粒度对象概念，并显著提升了性能。近期研究已证明（Zhang 等人，2021年），这些视觉语言预训练方法的效果在很大程度上得益于物体检测器质量。然而，这些现成的预提取器是针对特定场景、依赖昂贵的边界框标注构建的，这不可避免地限制了视觉语言预训练在更通用领域中的应用。另一方面，对象检测器和场景图解析器需要对高分辨率图像进行推理，而研究已证明（Li 等人，2021a），这会带来高昂的计算开销。

一种无需物体标注即可实现细粒度视觉语言预训练的弱监督范式（Gao 等人，2020年；Kim、Son 和 Kim，2021年），是将图像块视为视觉词元。然而，据报道，移除对象检测器会导致性能下降（Kim、Son 和 Kim，2021年）。近期工作（Li 等人，2021a、2022a；Yu 等人，2022年）在融合多模态特征之前，先对单模态编码器采用图像文本对比（ITC）目标。结合多模态编码器上的掩码语言建模与图文匹配目标，这些模型取得了出色的性能。然而，据报道（Li 等人，2022b），在使用同一数量级的训练语料时，这些方法的表现仍不及细粒度方法。

视觉模态的语义，例如纹理、物体，

以及动作，并不会像其他内容那样自然地分割开来，因此很难从视觉模态中构建明确的细粒度监督。幸运的是，语言符号天然离散，其基本语言语义单元是词元。因此，无需人工标注，也可以从语言模态中创建细粒度的自监督信号。掩码语言建模就是一种成功的监督方式。然而，大多数视觉语言预训练方法只是采用与 BERT 相同的掩码语言建模设计（Devlin 等人，2019）。我们认为，视觉语言预训练中的语言建模监督仍缺乏系统性的探索。

本文提出了一种用于视觉语言预训练（VLP）的精细化视觉-语言建模（RVLM）方案。首先，我们提出一种同音词句子改写算法（HSR），以生成细粒度监督。如图 1 所示，我们将图注中的一个动词 / 名词 / 形容词 / 量词词元替换为其在 WordNet（Miller，1995）中的同音词。改写后的句子用作负样本，其语义与原始图注仅有一个词元之差。随后，我们设计了三个自监督目标来利用这一监督信号。在单模态编码器上，精细化图文对比（RITC）任务最大化图像与真实词元之间的相似度，并最小化图像与被替换词元之间的相似度。在多模态编码器上，替换语言建模（RLM）任务预测哪个词元被替换，精细化图文匹配（RITM）任务则要求图像-图注的相似度高于图像-改写文本的相似度。

我们在多种下游视觉-语言任务上验证了所提出的细粒度视觉语言预训练框架的有效性。RVLM 显著优于弱监督模型，且能够达到与细粒度监督模型相当或更好的性能。

我们的主要贡献可概括为以下三点：
1. 我们从语言学角度提出了一种细粒度视觉语言预训练方案，通过同音词句子重写算法，在没有物体标注的情况下，从图像-描述对中获取词元级监督。2. 我们设计了一系列精细化视觉-语言建模任务，即 RITC、RITM 和 RLM，以利用细粒度监督。3. 在多个下游视觉语言任务上开展的大量实验，验证了所提方案的有效性。此外，我们还将发布代码及相关参数，以帮助其他研究人员。

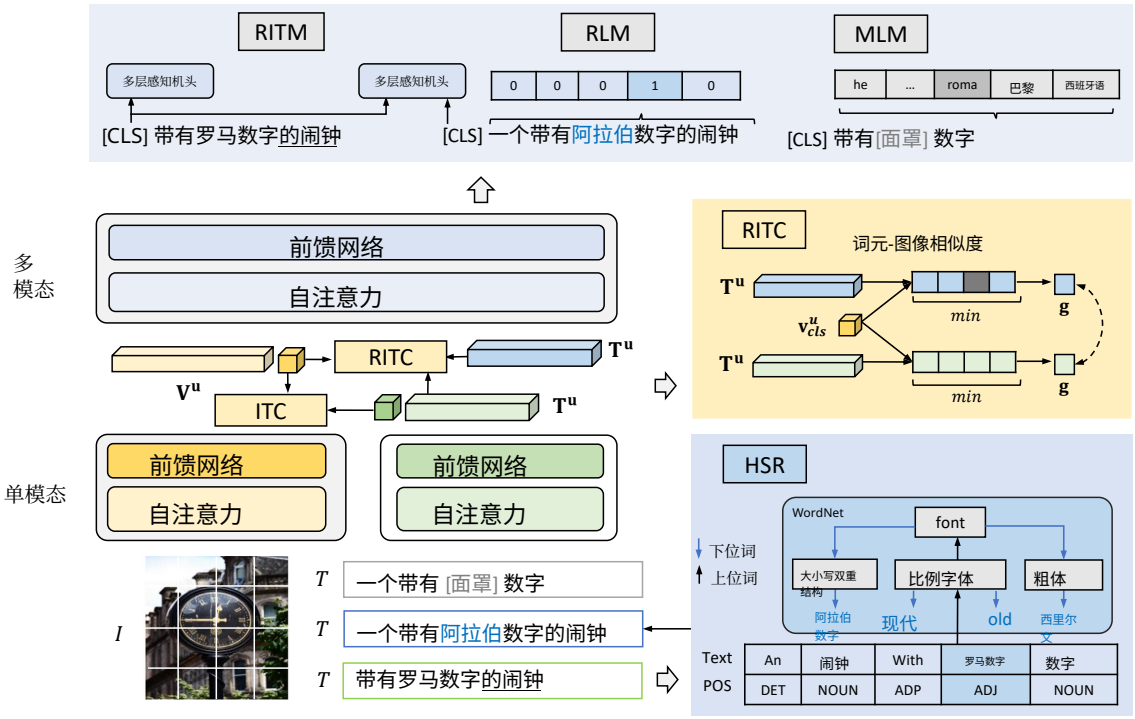
2 相关工作

2.1 细粒度视觉-语言预训练

早期视觉语言预训练模型使用的目标级标注来自自底向上和自顶向下特征（Anderson 等人，2018），这些特征由在 Visual Genome（Krishna 等人，2017）上预训练的 Faster-RCNN（Ren 等人，2015）生成。（Tan 和 Bansal，2019；Lu 等人，2019）分别使用两个单模态网络处理文本和图像，随后通过跨模态 Transformer 结合这两种来源。基于对象标注，研究者设计了掩码对象建模目标，以学习物体与文本之间的细粒度对应关系。后续方法（Li 等人，2019，2020a；Chen 等人，2020；Su 等人，2019；Li 等人，2020b）将感兴趣区域视为图像标记，并通过 BERT（Devlin 等人，2019）网络统一图像和文本编码器。这些模型改进了预训练目标，从而更好地学习跨模态对应关系，并取得更出色的性能。VinVL（Zhang 等人，2021）探究了对象检测器的重要性，并借助改进的对象检测器（Anderson 等人，2016），在下游视觉和语言任务中取得更高性能。最近，（Li 等人，2022b）将对象标签监督扩展到来自场景图解析器（Anderson 等人，2016）的多粒度对象概念。同时，（Zeng、Zhang 和 Li，2021）将对象框引入预训练，并提出短语定位目标。这些多粒度方法通过更丰富的对象标注显著提升了性能。事实证明，对象检测器和场景图解析器比视觉Transformer（Dosovitskiy 等人（2020））更加耗费计算资源（Li 等人，2021a）。另一方面，这些模型依赖特定场景中成本高昂的边界框标注，这不可避免地限制了视觉语言预训练在更广泛领域中的应用。RITM 探索了不依赖语言模态对象标注的细粒度监督。此外，由于无需对象特征提取阶段，RITM 比这些细粒度视觉-语言预训练方法更快。

2.2 弱监督细粒度视觉-语言预训练

弱监督细粒度视觉语言预训练方法不使用物体标注。(Gao 等人，2020) 首次探索了用于视觉语言预训练的基于图像块的图像词元，并应用于电子商务检索。他们将时尚图像切分为图像块，并使用 ResNeXt-101 (Xie 等人，2017) 提取特征，作为视觉输入。ViLT (Kim、Son 和 Kim，2021) 提出将图像块词元用作 ViT (Dosovitskiy 等人（2020）)，该模型完全无卷积。由于没有对象标注，模型使用图文匹配和掩码语言建模目标进行训练。与使用相近语料库的细粒度视觉语言预训练方法相比，该模型推理速度更快，但性能较低。ALBEF (Li 等人，2021a) 通过先对齐后融合框架弥补了性能差距。该框架针对单模态编码器输出优化图文对比目标，并使用动量蒸馏网络指导训练过程。在此框架基础上，(Li 等人，2022a；Yu 等人，2022) 提出将自回归语言生成作为语言建模任务，并使用图像描述生成器扩大训练数据规模。这些方法借助扩充数据取得了出色的性能。然而，据报道 (Li 等人，2022b)，在使用同一数量级的训练语料时，这些方法仍不及细粒度方法。这些方法只是采用与单模态预训练 BERT (Devlin 等人，2019) 相同的语言建模目标，并未充分探索视觉与语言的语言建模任务。RITM 为视觉语言预训练提出了一种精细化语言建模方案，可从语言模态提供细粒度监督。



ˆ 图 2：RVLM 的整体架构。输入描述 T 通过 HSR 算法改写为负文本 T。通过精细化图文对比（RITC）目标学习单模态表示。通过精细化图文匹配（RITM）和替换语言建模（RLM）目标学习多模态表示。

3 方法论

所提出的 RVLM 从图像-描述输入 I 和 T中学习细粒度跨模态对齐。如图 2 所示，RVLM 包含三个 Transformer 模块：用于单模态阶段的图像编码器和文本编码器，以及用于多模态阶段的跨模态编码器。

形式上，我们首先定义一些符号。我们使用粗体大写字母（例如，$\mathbf{X}$）表示矩阵，使用粗体小写字母（例如，$\mathbf{x}$）表示向量。非粗体字母（例如，x）用于表示标量，希腊字母（例如，β）用于表示参数。

3.1 模型输入

输入图像 I 被切分为 8x8 像素的图像块，然后编码为嵌入序列：$\mathbf{V}=\{\mathbf{v}_{cls}, \mathbf{v}_1, ..., \mathbf{v}_N\}$，遵循 ViT（Dosovitskiy 等人（2020））的做法，其中 $\mathbf{v}_{cls}$ 是 [CLS] 标记的嵌入。句子 T 被分词为词元索引序列$\mathbf{T}=\{\mathbf{w}_{cls}, \mathbf{w}_1, ..., \mathbf{w}_L\}$，其中 L 是 WordPiece（Johnson等，2017年）分词器输出序列的长度。

在模型预训练阶段，我们通过同音异义句子改写，将输入文本 T改写为 ˆT，以进行细粒度的 ¯监督。我们准备了掩码文本 T 用于掩码语言建模。

同音异义句子改写 同音异义句子改写（HSR）旨在为图像提供细粒度的负句子。基于词元是语言基本语义单位这一原则，我们可以改变句子中的一个词元。HSR 监督不需要物体标注。

为了获得一个难负句子，替换词的语义应与真实词相关但有所不同。因此，我们建议从 WordNet（Miller, 1995）中选择所选词元的同音异义词。由于图像说明具有描述性，我们仅选择具体的名词、形容词、动词和量词作为改写候选词。

如图 2 中的 HSR 模块所示，给定一个句子 T，我们首先提取每个词的词性 S，然后从名词/形容词/动词/量词中随机选择一个替换词 T_i，其中 i 是词索引。接着，我们从 WordNet 中查找上位词，并遍历其下位词以获取同义词。如果该词没有同义词，我们会进一步查找二级上位词及其二级下位词。随后，将改写后的文本 ˆ ˆT 分词为 $\mathbf{T}$，用于单模态和多模态学习。

3.2 精炼的单模态学习

ˆ单模态阶段以$\mathbf{V}$、$\mathbf{T}$ 和$\mathbf{T}$ 为输入，学习它们的细粒度单模态表示。文本编码器采用 6 层 Transformer（Vaswani 等人（2017）），图像编码器则采用 12 层视觉 Transformer ViT-B/16（Dosovitskiy 等人，2020）。输入被编码为单模态表示$\mathbf{V}^u=$ˆ$\{\mathbf{v}^u_{cls}, \mathbf{v}^u_1, ..., \mathbf{v}^u_N\}$、$\mathbf{T}^u=\{\mathbf{w}^u_{cls}, \mathbf{w}^u_1, ..., \mathbf{w}^u_L\}$以及$\mathbf{T}^u=$

{$\hat{\mathbf{w}}^u_{cls}$, $\hat{\mathbf{w}}^u_1$, ..., $\hat{\mathbf{w}}^u_L$}。随后，通过图像文本对比和 RITC 对该表示进行优化。

图文对比学习 图像文本对比（ITC）主要关注图像与句子之间的粗粒度对齐。对于经过 L2 归一化的 [CLS]表示 $h_v(\mathbf{v}^u_{cls})$ 和 $h_w(\mathbf{w}^u_{cls})$，我们使用 InfoNCE（Yang 等人，2019）损失，其中 h_v 和 h_w 分别是文本和图像表示的线性投影层。与（Li 等人，2021a）类似，我们使用动量网络辅助训练。将动量网络的表示记为 $h'_v(\mathbf{v}'^u_{cls})$ 和 $h'_w(\mathbf{w}'^u_{cls})$，文本与图像之间的相似度记为 $s = h_v \; {}^u_{cls} \; \underset{\intercal}{h_v} \; {}'^u_{cls}$ (T, I') (w) (v)，则文本到图像的对比损失可表示为式 1：

$$\mathcal{L}^t_c = \tag{1}$$

其中，τ 是可学习的温度参数，N 是批次大小，M 是动量队列长度。对称地，我们有式 2。

$$\mathcal{L}^t_c = \tag{2}$$

ITC 损失通过对分数计算交叉熵损失得出：

$$\mathcal{L}_{itc} = \mathcal{L}^t_c + \mathcal{L}^v_c \tag{3}$$

精细化图文对比RITC 损失旨在计算图像与词元之间的细粒度相似度。我们计算图像与词元$\mathbf{g} = \{h_v(\mathbf{v}^u_{cls}) \; h_w(\mathbf{w}^u_i)\}^L_{i=1}$ 以及重写文本 $\hat{\mathbf{g}}_{\intercal} = \{h_v(\mathbf{v}^u_{cls}) \; h_w(\hat{\mathbf{w}}^u_i)\}^L_{i=1}$之间的相似度。线性投影层与 ITC 共用相同的权重。随后，我们使用$\mathbf{g}$ 和 $\hat{\mathbf{g}}$ 的下界来衡量文本与图像的相似度，从而通过式 4 中的间隔损失优化细粒度对齐：

$$\mathcal{L}_{ritc} = \max(0, (\beta + \min(\mathbf{g}) - \min(\hat{\mathbf{g}}))) \tag{4}$$

其中，β 是间隔超参数。RITC 可视为一种弱监督学习，用于对齐图像与词元。

3.3 精细化多模态学习

在多模态阶段，我们通过 RITM 损失促进跨模态语义学习，以实现粗粒度对齐；并通过 MLM 和 RLM 损失实现细粒度对齐。多模态编码器是一个 6 层的跨模态Transformer，可将单模态表示融合到语言模态中。它为说明文字、掩码文本和重写文本生成多模态表示$\mathbf{T}^m$,$^-$ $\hat{\mathbf{T}}^m$ 和$\mathbf{T}^m$ 。

精细化图文匹配 RITM 损失通过学习文本-图像对的全局表示，预测文本与图像是否匹配。我们采用两种负样本对进行匹配。第一种是批次内的困难负样本，根据单模态阶段的相似度分数选取。第二种是图像与改写后的文本。损失函数是在 [CLS] 表示上进行的二元分类损失，分为两部分，计算方式如公式 5 所示。

$$\begin{aligned}\mathcal{L}_{ritm} = \mathbb{E}_{(I,T,\hat{T})\sim D} & H(\mathbf{y}^{itm}, \mathbf{p}^{ritm}(\mathbf{T}^m_{cls})) \\ & + H(\hat{\mathbf{y}}^{itm}, \hat{\mathbf{p}}^{ritm}(\hat{\mathbf{T}}^m_{cls}))\end{aligned} \tag{5}$$

	COCO VG	CC	SBU	OI	Flickr30k	总计
图像	113k 100k	2.8M	860k	504k	30k	4.4M
text	567k 769k	2.8M	860k	507k	150k	5.7M

表 1：预训练数据集的统计信息。

其中，$\mathbf{y}^{itm}$ 和 $\hat{\mathbf{y}}^{itm}$ 是表示对齐标签的二维独热向量，H 表示交叉熵函数，$\mathbf{p}^{ritm}$ 和 $\hat{\mathbf{p}}^{ritm}$ 是带有 softmax 的单层线性投影，用于将 [CLS] 表示投影为二元概率。

掩码语言建模 结合视觉线索的掩码语言建模是视觉语言预训练中的传统且重要的方法。它利用多模态上下文预测被掩码词。除了随机掩盖 15% 输入词元的传统设置外，我们还会掩盖 HSR 指示的替换词元。随后，80% 的被掩码词元会被替换为 [MASK]，10% 会被替换为随机词元。掩码语言建模项的公式如式 6 所示

$$\mathcal{L}_{mlm} = \mathbb{E}_{(I,\bar{T})\sim D} H(\mathbf{y}^{msk}, \mathbf{p}^{msk}(\bar{\mathbf{T}}^m)) \tag{6}$$

其中，$\mathbf{y}^{msk}$ 是一种独热词汇分布，真实词元的概率为 1；$\mathbf{p}^{mlm}$ 是模型的预测头网络。

掩码语言建模目标通过分类学习多模态上下文与词元之间的细粒度对应关系。这种方法有两个缺点：首先，可能有多个合适的词元可填入掩码位置，但该目标忽略了其他正确的词元。其次，它并未直接学习多模态上下文与词元之间的对齐关系。

替换语言建模 替换语言建模任务根据多模态内容预测哪个词元被替换。它弥补了上述掩码语言建模的两个局限。首先，二元词元对齐是确定的。其次，该目标直接学习图像与词元之间的细粒度对齐。细粒度对齐学习采用弱监督方式。损失的计算公式如式 7 所示

$$\mathcal{L}_{rlm} = \mathbb{E}_{(I,\hat{T})\sim D} H(\mathbf{y}^{rlm}, \mathbf{p}^{rlm}(\hat{\mathbf{T}}^m)) \tag{7}$$

其中，$\mathbf{y}^{rlm}$ 是长度为 L 的 one-hot 分布，被替换词元对应的概率为 1；$\mathbf{p}^{rlm}$ 是模型的预测头网络。

RVLM 的完整预训练目标由上述各项组合而成：

$$\mathcal{L} = \mathcal{L}_{itc} + \mathcal{L}_{ritc} + \mathcal{L}_{ritm} + \mathcal{L}_{mlm} + \mathcal{L}_{rmlm} \tag{8}$$

模型在预训练数据集上进行端到端优化，以最小化 $\mathcal{L}$。

3.4 预训练数据集

我们在大规模视觉语言语料库上对模型进行预训练，其中包括 MSCOCO（Lin 等，2014）、Visual Genome（Krishna 等人，2017）、Flickr30k（Plummer

方法	Flickr30k 测试集（1k 张图像）							COCO 测试集（5k 张图像）							# 预训练图像
	文本检索			图像检索				文本检索			图像检索				
	R@1	R@5	R@10	R@1	R@5	R@10	AR	R@1	R@5	R@10	R@1	R@5	R@10	AR	
弱监督细粒度视觉语言预训练															
ALIGN（2021年）	95.3	99.8	100	84.9	97.4	98.6	96.0	77.0	93.5	96.9	59.9	83.3	89.8	83.4	1.8B
FILIP（2021年）	**96.6**	**100**	**100**	**87.1**	**97.7**	**99.1**	**96.8**	78.9	94.4	97.4	61.2	84.3	90.5	84.5	340M
ViLT（2021年）	83.5	96.7	98.6	64.4	88.7	93.8	87.6	61.5	86.3	92.7	42.7	72.9	83.1	73.2	14M
ALBEF（2021a）	94.3	99.4	99.8	82.8	96.7	98.4	95.2	73.1	91.4	96.0	56.8	81.5	89.2	81.3	4M
ALBEF-14M（2021a）	95.9	99.8	100	85.6	97.5	98.9	96.2	77.6	94.3	97.2	60.7	84.3	90.5	84.1	14M
细粒度监督视觉语言预训练															
Uniter（2020年）	85.9	97.1	98.8	72.5	92.4	96.1	90.5	65.7	88.6	93.8	52.9	79.9	88.0	78.1	4M
Oscar（2020b）	-	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5	80.0	4M
UNIMO（2021b）	89.7	98.4	99.1	74.7	93.4	99.1	92.4	-	-	-	-	-	-	-	4M
ROSITA（2021年）	88.9	98.1	99.3	74.1	92.4	96.1	91.4	71.3	91.6	95.6	54.4	80.9	88.6	80.4	4M
VinVL（2021年）	93.6	99.1	99.9	82.0	95.7	97.7	94.6	74.6	92.6	96.3	58.1	83.2	90.1	82.5	5.7M
MVPTR（2022b）	95.2	99.7	100	84.0	96.8	98.5	95.7	77.3	93.6	96.9	60.1	84.0	90.7	83.8	4.7M
RVLM	94.7	99.8	99.8	82.4	96.1	98.0	95.1	77.3	94.2	97.2	60.9	84.6	91.2	84.2	4.4M
RVLM-Large	95.6	99.8	**100**	85.7	97.6	98.8	96.2	**79.5**	**95.1**	**97.9**	**63.1**	**85.6**	**91.9**	**85.5**	14M

表 2：Flickr30k 和 COCO 数据集上的图文检索结果，以 R@k 指标表示。

等人，2015）、Conceptual Captions（Sharma 等人，2018）[1],SBU（Ordonez、Kulkarni 和 Berg，2011）以及 OpenImages（Kuznetsova 等人，2020）。表 1 展示了预训练数据集中的图像和文本统计信息。

与 Li 等人（2022b）一样，我们排除了 Flickr30k 的验证集/测试集划分。此外，我们按照 Li 等人（2021a）的方法，使用规模更大的网络数据 Conceptual Concept 12M（Changpinyo 等人，2021）扩展 RVLM，该数据集共包含 1400 万张图像。据报道，大型 SBU 和 Conceptual Captions 数据集存在噪声（Hendricks 等人，2021），其中的图像说明文字可能与图像并不完全对应。因此，噪声词语可能会模糊 HSR 的词元级监督。为解决这一问题，我们将训练过程分为两个阶段：前 30 个训练轮次使用全部预训练语料库进行训练，随后使用两个由人工标注的 MSCOCO 和 Visual Genome 数据集训练 10 个训练轮次。

3.5 实现细节

图像编码器使用 MAE-base（He 等人，2022年）的前 12 层权重进行初始化，该模型在 ImageNet-1k 上经过无标签预训练。文本编码器和跨模态编码器使用 BERT-base（Devlin 等人，2019）的权重进行初始化。采用 AdamW（Loshchilov 和 Hutter，2018年）优化器，学习率为 $1e^{-4}$，权重衰减为 0.02；前 20 个训练轮次进行预热，随后在 10 个训练轮次内线性衰减至 0。批次大小设为 32，动量队列大小为 65536。文本词元的最大序列长度设为 30。模型在配备 Ubuntu 系统和 8 个 V100 GPU 的 NVIDIA DGX 上以半精度进行训练。

4 实验

为评估所提出的方法，我们在三类下游 V+L 任务上进行了大量实验。我们将在下文介绍各项任务和实验结果。

[1]~0.5M 张图像因 URL 无效而被排除。

4.1 图文检索

我们评估 RITM 在图像到文本检索和文本到图像检索两项任务上的表现。基准数据集为 Flickr30K 和 COCO。检索性能以前 k 个样本的召回率（R@k）衡量。对于文本到图像检索及其反向任务，我们报告三个 k 值：R@1、R@5 和 R@10。微调时，我们使用与预训练相同的数据集进行优化。推理时，我们采用 Li 等人（2021a）提出的召回与重排序策略，以提高检索效率。首先，按照公式 3 的单模态相似度分数，为所有图文对召回前 k 个候选项，然后计算其图文匹配分数并进行排序。此外，我们还采用 CXC （ Parekh 等人（ 2020 ））指标，该指标旨在为 COCO 图像提供更精确的排序。

表 2 展示了 RVLM 与当前先进的视觉语言预训练方法在 COCO 5k 和 Flickr30k 1k 测试集上的图文检索性能对比。总体而言，目标标注监督方法的表现优于弱监督方法。然而，RVLM 的表现与 MVPTR 等目标标注监督方法相当。与弱监督方法相比，RVLM 的表现更优，尤其是在 COCO 数据集上。在 400 万预训练设置下，RVLM 的 R@1 分数比 ALBEF 高.2% 和 4.1%，这表明 RVLM 能够学习更好的细粒度对齐关系，从而区分相似图像。与目标标注监督的 MVPTR 模型相比，RVLM 在 Flickr30k 上的平均表现仅比 MVPTR 低 0.6%，但在 COCO 5k 测试集上则高出 0.5%。

由于某些查询对应多个语义对齐的图像或文本（Parekh 等人（2020）），我们进一步在 CXC 指标上进行实验，以验证我们的改进

方法	COCO 测试集 (CXC)					
	文本检索			图像检索		
	R@1	R@5	R@10	R@1	R@5	R@10
ALBEF	74.48	92.34	96.20	59.43	83.67	89.87
ALBEF-14M	78.48	94.66	97.64	63.63	85.80	91.45
MVPTR	78.16	94.98	97.39	63.52	85.03	90.97
RVLM	78.22	95.10	97.70	63.74	86.46	92.22
RVLM-14M	**80.76**	**95.32**	**98.16**	**65.64**	**87.18**	**92.62**

表 3：采用 CXC（Parekh 等人，2020年）指标的 COCO 数据集图文检索结果。

在 COCO 数据集上。如表 3 所示，RVLM 仍持续优于对比方法，并在 R@1 分数上取得最大提升，这表明 RVLM 擅长学习细粒度对齐关系。

4.2 多模态分类

我们在两个广泛使用的多模态分类任务上评估 RVLM，即视觉问答（VQA）和视觉蕴含（VE）。

VQA 要求模型根据图像和问题作为上下文来选择答案。我们采用广泛使用的 VQA v2（Goyal 等人，2017年）基准数据集作为评估数据集。评估指标为答案选择的准确率。遵循（Li 等人，2021a；Cho 等人，2021年），我们将 VQA 视为答案生成问题。RVLM 将图像和问题作为输入，并使用 6 层 Transformer 解码器生成 3，192 个候选答案。

VE 是一项细粒度视觉推理任务，旨在预测图像与文本之间的关系。关系类别包括蕴含、中性或矛盾。评估指标为关系分类的准确率。我们遵循视觉语言预训练的常见做法（Li 等人，2021a；Chen 等人，2020年），将 VE 构建为三分类问题，并在多模态编码器的 [CLS] 标记表示$\mathbf{T}_{cls}^{m}$上使用多层感知机（MLP）层预测类别概率。

表 4 展示了 RVLM 与最先进的 VLP 模型在 VQA 和视觉蕴含任务上的比较。在这两项任务中，基于目标标注的模型都略优于现有的弱监督模型。在 VQA v2 基准测试中，RVLM 优于其他弱监督方法，其准确率也接近最先进的目标标注监督模型，证明了 RVLM 在学习细粒度语义方面的有效性。在视觉蕴含任务中，RVLM 优于这两类方法，证明了该模型学习跨模态关系的卓越能力。

4.3 弱监督视觉定位

视觉定位要求模型从输入图像中定位特定文本描述所对应的区域。我们的 RVLM 在 RefCOCO+ （Yu 等，2016）数据集上进行微调。由于 RVLM 不依赖任何目标框输入，因此无法使用该数据集的边界框标注作为

方法	VQA		SNLI-VE	
	测试开发集	测试标准集	val	test
ViLT‡	70.94	-	-	-
ALBEF‡	74.54	74.70	80.14	80.30
Uniter	72.70	72.91	78.59	78.28
OSCAR	73.16	73.44	-	-
VILLA	73.59	73.67	79.47	79.03
UNIMO	73.79	74.02	80.00	79.10
VinVL	75.95	76.12	-	-
MVPTR	**76.16**	**76.36**	80.30	80.17
RVLM‡	75.85	75.93	**80.52**	**80.41**

表 4：VQAv2 和 SNLI-VE 数据集上的多模态分类结果（带有 ‡ 的模型为弱监督方法）。

方法	RefCOCO+		
	dev	测试A	测试B
ARN（Liu 等，2019）	32.78	34.35	32.13
CCL（Zhang 等，2020）	34.29	36.91	33.56
ALBEF-14M	58.46	65.89	46.25
RVLM-14M	**58.75**	**67.32**	**46.74**

表 5：Re-fCOCO+ 数据集上的弱监督视觉定位结果。

监督信号。因此，我们采用 Li 等人，2021a 中的弱监督设置，并基于图文匹配分数 $\mathcal{L}_{itc}$ 对预训练模型 1400万 进行微调。在推理过程中，我们首先使用 Grad-CAM（Selvaraju 等，2017）获取每个像素的热力图，并利用这些热力图对 Yu 等，2018 提供的检测到的候选区域进行排序。表 5 展示了 RefCOCO+ 上弱监督视觉定位的对比结果。RVLM 在两个测试集上的表现均优于现有的弱监督方法，证明了其学习细粒度对齐的能力。

4.4 进一步分析

基于 Grad-CAM 的细粒度对齐可视化为检查 RVLM 学到的词元级对齐，我们使用 Grad-CAM 可视化每个词在图像上的交叉注意力图。图 3 展示了几种可视化结果。第一行中，名词（例如闹钟和柱子）以及动词（安装）都与其对应区域对齐。特别地，我们展示了同一内容中替代词“罗马数字”和“阿拉伯数字”的注意力图。RVLM 正确对齐了“罗马数字”，而“阿拉伯数字”没有与罗马数字对齐。第二行包含量词“一”和“二”，它们与长颈鹿上正确数量的区域对齐。第三行展示了图像与多个物体之间的细粒度对齐。可以看出，词元与物体实现了精确对齐。具体而言，两个动词中，“喝”正确关联到杯子和嘴，“拿”则关联到双手，

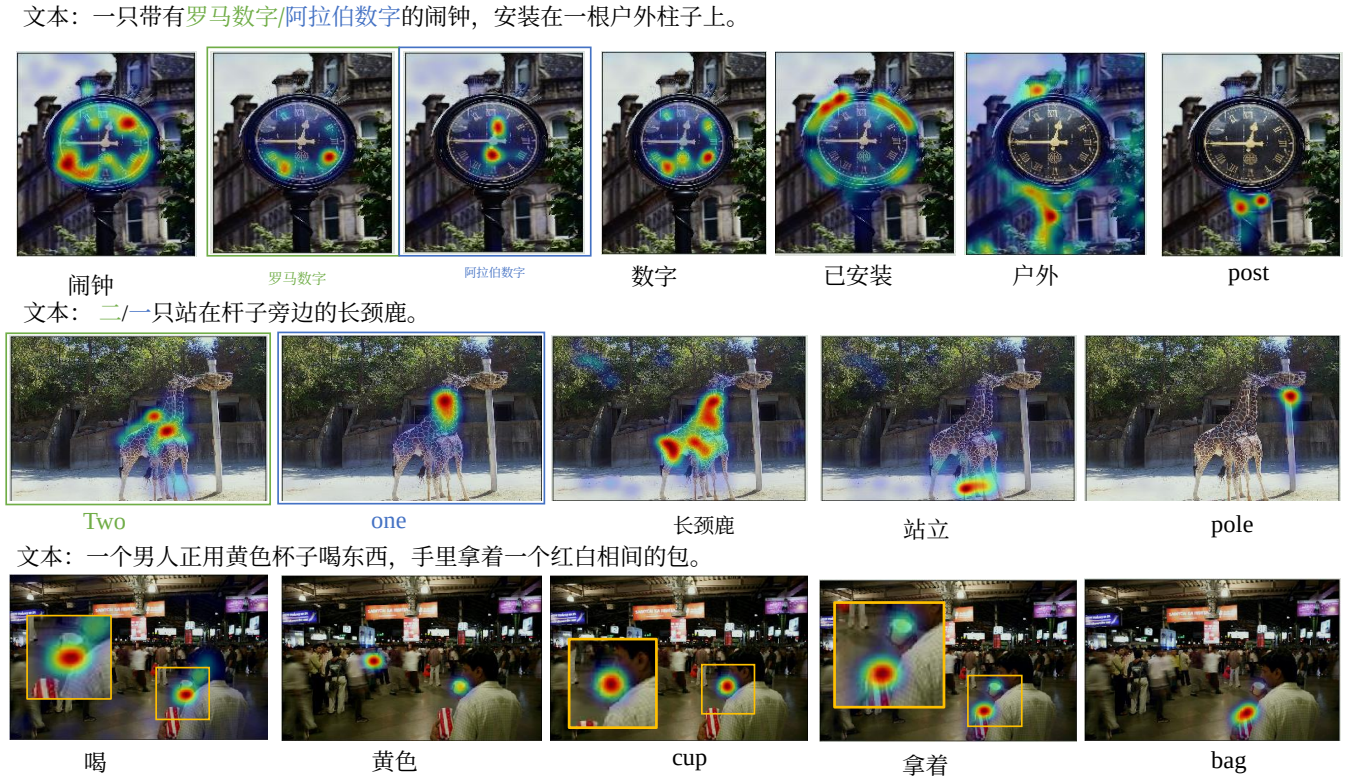


图3: 区域与词语之间交叉注意力图上的 Grad-CAM（放大查看效果更佳n）。被绿色包围的图像 色或蓝色方框圈出的图像来自包含相应词语的句子。

方法	SNLI-VE		COCO（5k）	
	dev	test	I2T	T2I
RVLM	80.52	80.41	77.3	60.9
不使用 RITM	80.32	80.23	77.0	59.7
不使用 RITC	80.14	80.11	75.9	59.3
不使用 RLM	79.09	79.03	76.7	58.9
不使用 HSR	78.89	78.24	75.1	58.8

表6：针对 VE 和图文检索任务的消融研究。使用 COCO 5k 测试集上的 R@1 分数。

这证明 RVLM 能够区分这些动词。两个同音异义词“杯子”和“包”也被正确定位，验证了 RVLM 的细粒度对齐能力。

消融研究 我们在 SNLI-VE 和 COCO 5k 数据集上进行了详细的消融研究，以分析各个提议组件的有效性。具体而言，我们在预训练的第二阶段移除了 RITM、RLM 和 RITC 损失。我们还通过使用随机选取的词语进行改写来验证 HSR策略。所有变体均使用相同的超参数和训练轮次进行训练。表6展示了两个基准数据集的结果。可以观察到，移除任一组件都会导致两项任务的性能下降。使用随机选取的词语进行改写会造成显著的性能下降，这证明了同音异义词语的有效性。

5 结论

本文提出了一种新颖的精细化视觉语言建模方法（RVLM），用于弱监督的细粒度视觉语言预训练。我们不使用对象标注，而是提出了一种同音异义句子改写（HSR）算法，以生成词元级负句。为从改写后的句子中获取细粒度监督，我们提出了精细化图文对比损失（RITC），以在单模态阶段学习细粒度对齐；并在多模态阶段采用精细化图文匹配（RITM）和替换语言建模（RLM）任务。大量实验和分析表明，RVLM 在多个下游基准上优于当前最先进的模型。

可以考虑若干未来研究方向和改进。我们的主要目标是展示弱监督精炼视觉语言模型框架的潜力。HSR仅提供词元级监督，因此可以考虑采用短语级改写策略来实现短语级监督。此外，开发适用于图像模态的改写技术也有望实现细粒度监督。

参考文献

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016年。SPICE：语义命题图像描述评估。收录于欧洲计算机视觉会议，382–398。施普林格。

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018年。用于图像描述和视觉问答的自下而上与自上而下注意力。收录于*IEEE*计算机视觉与模式识别会议论文集，6077–6086。

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021年。Conceptual 12M：推动网络规模图文预训练，以识别长尾视觉概念。收录于*IEEE/CVF*计算机视觉与模式识别会议论文集，3558–3568。Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020年。Uniter：通用图文表征学习。收录于欧洲计算机视觉会议，104–120。Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021年。通过文本生成统一视觉与语言任务。收录于国际机器学习会议，1931–1942。PMLR。

Cui, Y.; Yu, Z.; Wang, C.; Zhao, Z.; Zhang, J.; Wang, M.; and Yu, J. 2021年。ROSITA：通过跨模态与模态内知识整合增强视觉与语言语义对齐。收录于第*29*届*ACM*国际多媒体会议论文集，797–806。Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019年。BERT：用于语言理解的深度双向变换器预训练。收录于*2019*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（长短论文），4171–4186。Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; 等。2020年。一张图像胜过16×16个词：用于大规模图像识别的Transformer。收录于国际学习表征会议。Gao, D.; Jin, L.; Chen, B.; Qiu, M.; Li, P.; Wei, Y.; Hu, Y.; and Wang, H. 2020年。FashionBERT：利用自适应损失进行文本与图像匹配，实现跨模态检索。收录于第*43*届国际*ACM SIGIR*信息检索研究与发展会议论文集，2251–2260。Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017年。让VQA中的“V”发挥作用：提升图像理解在视觉问答中的作用。收录于*IEEE*计算机视觉与模式识别会议论文集，6904–6913。He, K.; Chen, X.; Xie, S.; Li, Y.; Doll´ar, P.; and Girshick, R. 2022年。掩码自编码器是可扩展的视觉学习器。收录于*IEEE/CVF*计算机视觉与模式识别会议论文集，16000–16009。Hendricks, L. A.; Mellor, J.; Schneider, R.; Alayrac, J.-B.; and Nematzadeh, A. 2021年。解耦数据、注意力和损失在多模态变换器中的作用。汇刊计算语言学协会》，9：570–585。Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021年。通过噪声文本监督扩大视觉与视觉语言表征学习的规模。*arXiv*预印本 *arXiv:2102.05918*。Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Vi´egas, F.; Wattenberg, M.; Corrado, G.; 等。2017年。谷歌的多语言神经机器翻译系统：实现零样本翻译。《计算语言学协会汇刊》，5：339–351。Kim, W.; Son, B.; and Kim, I. 2021年。ViLT：无需卷积或区域监督的视觉与语言 Transformer。发表于国际机器学习会议，5583–5594。PMLR。Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; 等。2017年。Visual Genome：利用众包密集图像标注连接语言与视觉。《国际计算机视觉杂志》，123(1)：32–73。Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; 等。2020年。Open Images 数据集 V4。发表于《国际计算机视觉杂志》，128(7)：1956–1981。Li, G.; Duan, N.; Fang, Y.; Gong, M.; and Jiang, D. 2020a。UniCoder-VL：通过跨模态预训练实现视觉与语言通用编码器。发表于*AAAI*人工智能会议论文集，第34卷，11336–11344。Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a。BLIP：通过自举式语言-图像预训练实现统一的视觉语言理解与生成。*arXiv*预印本 *arXiv:2201.12086*。Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021a。先对齐后融合：通过动量蒸馏进行视觉与语言表征学习。第34卷。Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019。VisualBERT：用于视觉与语言的简单而高效的基线。*arXiv*预印本 *arXiv:1908.03557*。Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021b。UNIMO：通过跨模态对比学习实现统一模态理解与生成。发表于计算语言学协会第*59*届年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），2592–2607。Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; 等。2020b。OSCAR：面向视觉语言任务的物体语义对齐预训练。发表于欧洲计算机视觉会议，121–137。Li, Z.; Fan, Z.; Tou, H.; and Wei, Z. 2022b。MVP：通过多层语义对齐进行多阶段视觉语言预训练。*arXiv*预印本 *arXiv:2201.12596*。

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Doll´ar, P.; and Zitnick, C. L. 2014。Microsoft COCO：情境中的常见物体。发表于欧洲计算机视觉会议，740–755。施普林格。Liu, X.; Li, L.; Wang, S.; Zha, Z.-J.; Meng, D.; and Huang, Q. 2019。用于弱监督指代表达定位的自适应重建网络。发表于电气电子工程师学会/计算机视觉基金会国际计算机视觉会议论文集，2611–2620。Loshchilov, I.; and Hutter, F. 2018。解耦权重衰减正则化。发表于国际学习表征会议。Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019。ViLBERT：面向视觉与语言任务的预训练任务无关视觉语言表征。*arXiv*预印本 *arXiv:1908.02265*。Miller, G. A. 1995。WordNet：英语词汇数据库。*ACM*通讯，38(11)：39–41。Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011。Im2Text：使用100万张带说明文字的照片描述图像。发表于神经信息处理系统（*NIPS*）。Parekh, Z.; Baldridge, J.; Cer, D.; Waters, A.; and Yang, Y. 2020。交叉字幕：针对 MS-COCO 的扩展型模态内与跨模态语义相似度判断。*arXiv*预印本 *arXiv:2004.15020*。Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015。Flickr30k Entities：收集区域到短语的对应关系，以构建更丰富的图像到句子模型。发表于电气电子工程师学会国际计算机视觉会议论文集，2641–2649。Qi, D.; Su, L.; Song, J.; Cui, E.; Bharti, T.; and Sacheti, A. 2020。ImageBERT：利用大规模弱监督图像-文本数据进行跨模态预训练。*arXiv*，1–12。Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015。Faster R-CNN：利用区域提议网络实现实时目标检测。发表于神经信息处理系统进展，91–99。Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017。Grad-CAM：通过基于梯度的定位为深度网络提供视觉解释。发表于电气电子工程师学会国际计算机视觉会议论文集，618–626。Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018。Conceptual Captions：用于自动图像描述的清洗后、上位词化图像替代文本数据集。发表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。Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019。VL-BERT：通用视觉语言表征的预训练。发表于国际学习表征会议。Tan, H.; and Bansal, M. 2019。LXMERT：从 Transformer 中学习跨模态编码器表征。发表于*2019*年自然语言处理实证方法会议暨第*9*届国际自然语言处理联合会议论文集，5100–5111。Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017年. 注意力机制就是你所需要的一切。神经信息处理系统进展，30. Xie,S.; Girshick, R.; Doll´ar,P.; Tu, Z.; and He,K. 2017年. 深度神经网络的聚合残差变换。载于*IEEE*计算机视觉与模式识别会议论文集，1492–1500. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019年. Xlnet：用于语言理解的通用自回归预训练。神经信息处理系统进展，32. Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; and Xu, C. 2021年. 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。Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhos- seini, M.; and Wu, Y. 2022年. Coca：对比式图像描述模型是图像-文本基础模型。*arXiv*预印本 *arXiv:2205.01917*。Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018年. Mattnet：用于指代表达理解的模块化注意力网络。载于*IEEE*计算机视觉与模式识别会议论文集，1307–1315. Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016年. 指代表达中的上下文建模。载于欧洲计算机视觉会议，69–85. 施普林格。Zeng, Y.; Zhang, X.; and Li, H. 2021年. 多粒度视觉语言预训练：使文本与视觉概念对齐。*arXiv*预印本 *arXiv:2111.08276*。Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021年. 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，5579–5588. Zhang, Z.; Zhao, Z.; Lin, Z.; He, X.; et al. 2020年. 面向弱监督视觉语言定位的反事实对比学习。神经信息处理系统进展，33: 18123–18134.