arXiv:2305.03689v3 [cs.CV] 3 Nov 2023

$\mathcal{C}ola$：组合式文本到图像检索基准测试

Arijit Ray[1,2] Filip Radenovic[2] Abhimanyu Dubey[2] Bryan A. Plummer[1]Ranjay Krishna[2,3] Kate Saenko[1,2]

{array, bplum, saenko}@bu.edu, {filipradenovic, dubeya}@fb.com, ranjay@计算机科学.washington.edu

[1]波士顿大学， [2]Meta AI（FAIR）， [3]华盛顿大学

摘要

组合推理是人类视觉智能的标志。然而，尽管大型视觉语言模型规模庞大，它们仍难以通过将物体与其属性相结合来表示简单的组合。为衡量这种组合能力的不足，我们设计了 $\mathcal{C}ola$，这是一个用于组合带属性定位的物体的文本到图像检索基准。要解决 $\mathcal{C}ola$，模型必须检索出属性和物体配置正确的图像，并避免选择一张包含相同物体和属性但配置错误的干扰项图像。$\mathcal{C}ola$ 包含约1.2k个组合查询，涉及168个物体和197个属性，涵盖约30K张图像。我们的人类评估发现， $\mathcal{C}ola$ 的准确率为83.33%，与当代组合性基准相近。我们使用$\mathcal{C}ola$ 作为测试平台，探索将预训练视觉语言模型调整为进行组合推理的经验性模型设计。我们在2个开创性的视觉语言模型上，使用以组合性为核心的测试基准—— $\mathcal{C}ola$和CREPE，探索了6种调整策略。我们发现，最优的调整策略是训练一个多模态注意力层，使其同时关注冻结的预训练图像和语言特征。令人惊讶的是，在CLIP上训练多模态层的表现优于调整规模更大的FLAVA模型，后者已经预训练了多模态层。此外，我们的调整策略将CLIP和FLAVA提升到了相当的水平，这表明，使用对比式属性-物体数据训练多模态层是关键，而不是直接使用预训练的多模态层。最后，我们展示了 $\mathcal{C}ola$ 比一个密切相关的当代基准CREPE更难，因为在CREPE上无需多模态层的简单微调策略就已足够，而在 $\mathcal{C}ola$上则不够。不过，我们在最佳调整结果与人类准确率之间仍然看到了显著差距，表明这一领域仍有相当大的进一步研究空间。项目页面：https://cs-people.bu.edu/array/research/cola/

1 引言

组合性是人类智能的一项基本特征，使我们能够将“整体的意义 [视为]其各部分意义的函数” [9]。在语言中，整体是由名词和形容词等词语组成的句子。在视觉中，整体是由物体和属性等视觉元素组成的图像 [28, 20]。例如，“圆形白色桌子”这一表达由名词“桌子”和形容词“圆形”“白色”组合而成，在图 1最左侧的照片中得到了视觉呈现。近期研究一致指出，这种物体与其属性之间的组合性，是现有视觉语言模型难以表征的内容 [55, 35, 25]。相反，它们会将属性分散开来，并将其对应到干扰物体上；例如，它们会错误地将“圆形白色桌子”匹配到

37th 神经信息处理系统大会（NeurIPS 2023）数据集与基准赛道



图 1：我们提出了 $\mathcal{C}ola$，要求模型将带有属性的物体进行组合与定位。要解决 $\mathcal{C}ola$，模型必须将正确的图像与正确的图像描述匹配，而不是匹配一张包含相同物体和属性、但配置错误的干扰图像。我们探索了将现有模型适配到该任务的各种可能机制；结果表明，微调预训练视觉语言表征的简单多模态适配方法效果最佳。

图 1 中左起第二张照片，方法是将属性“圆形”对应到圆盘，而非桌子。涉及两个物体的查询更具挑战性，见图1（右侧）。

本文研究大型视觉语言模型的**C**ompose **O**bjects **L**ocalizedwith**A**ttributes（$\mathcal{C}ola$）能力。不同于通过关系 [55] 和场景图 [35], 研究组合性的相关基线方法，我们聚焦于属性-对象绑定，因为在许多应用中，找到具有正确属性的物体至关重要。例如，当具身人工智能助手接到指令，要求清理“位于大型深色椅子右侧的小木桌”时，它不应开始清理“位于大型棕色木质光滑桌子旁边的棕褐色木椅”。此外，与包含关系或场景图的构图相比，物体-属性绑定从根本上说应该更容易处理。然而，我们发现，现有模型在这种更简单的绑定任务上仍然表现不佳。

为探究这些问题，我们提出 $\mathcal{C}ola$ 这一基准，用于评估对具有多个属性的定位物体进行组合的能力。 $\mathcal{C}ola$ 包含两类构图：单物体查询（图1 左）和多对象查询（图1 右）。在每种情况下，模型都应将查询中的物体与正确的属性相关联，并忽略那些将查询属性关联到干扰物体上的棘手干扰构图。多对象查询更难，因为其中呈现了更多构图。

与当代基准 CREPE [35]中的图像到文本设置不同， $\mathcal{C}ola$采用文本到图像的方式评估模型，即利用文本查询从一组图像中检索出正确的图像。这与以往针对视觉语言模型的基准 [21, 42, 64]保持一致。此外，文本到图像检索比图像到文本检索更难，因为对于给定文本，图像编码器区分图像中细粒度差异的能力不如文本编码器区分细粒度文本差异的能力 [55]。此外，文本到图像更贴合实际应用，例如用户向机器发出文本指令，让机器找到特定物体。

以 $\mathcal{C}ola$ 作为开发测试平台，我们的实验进一步支持了当前关于预训练视觉语言模型在构图方面表现不佳的讨论 [55]。因此，我们在两种具有开创性的视觉语言模型——CLIP [43] 和 FLAVA [52]——上探索了 6 种微调策略，以找出最能促进组合性的方法。我们使用 3 个数据集（GQA [18],、CLEVR [24], 和 PACO [44]）进行微调，并在 2 个测试平台（ $\mathcal{C}ola$ 和 CREPE [35]）上进行评估。尽管探索有效的预训练策略也是一个值得研究的方向，但我们将工作重点限定在适应策略上，因为从头开始训练模型成本高昂，且只有少数研究机构具备执行这一工作的能力。

我们发现，在适配过程中，表现最佳的架构选择是多模态 Transformer 编码器—解码器 [53, 52, 26]，用于进一步编码预训练模型中的视觉和语言表示。多模态适配的表现显著优于微调单模态编码器（仅编码视觉或语言特征）或微调提示。令人惊讶的是，这种适配使 CLIP 和 FLAVA 都能生成性能相当的微调模型，尽管我们的 CLIP 模型参数更少，而且 FLAVA 在预训练时已经使用了多模态 Transformer

层。这表明，使用对比式属性—物体数据训练多模态层至关重要，而不是使用在网络数据上预训练的多模态层。我们的适配方法也显著优于提示微调等适配或微调基础模型的常规方法 [65], 线性探测 [2], , 或微调数量相当的分离编码器层。与近期研究发现模型中存在结构组合性、但其表示中不存在结构组合性相似 [31], , 我们的研究发现，尽管预训练表示可能不具备组合性，但可以通过适配使其具备组合性。然而，人类准确率与我们最佳适配方法之间的显著差距表明，利用我们的基准仍有大量进一步研究的空间。

2 相关工作

组合性与图像检索。组合性是人类智能的关键方面 [10],，尤其是在视觉和语言领域 [7]。视觉-语言组合性已在视觉问答 [1], 、组合式图像检索（例如，具有 Y 风格的 X） [49], 以及生成 [56]中得到探索。组合性是提升对多样化查询的鲁棒性的关键因素之一，这一主题在视觉-语言领域得到了广泛研究 [46, 45, 51, 1, 8]。随着基础模型近期日益普及，各种研究都致力于测试它们的组合推理能力 [35, 55, 25, 62]。与 CREPE [35] 和 ARO [62], 相比，在我们的任务中，模型必须区分难以辨别的图像。文本到难辨图像的检索更具挑战性，因为区分难以辨别的图像（给定图像描述时）比区分难以辨别的图像描述更困难 [55]。Winoground [55]等基准主要评估广泛且复杂的关系组合性（例如，“男人从背后拥抱女人”与“女人从背后拥抱男人”），而我们专门关注查询中的属性-物体绑定。这一研究方向源于实际应用，例如具身智能体在杂乱的工作空间中，尝试检索定制物体（如“带红色橡胶手柄的金属扳手”），而周围存在相似的干扰物体 [39]。该领域的大多数属性-物体图像检索研究，要么关注单个属性 [19, 41] ，要么聚焦于非常狭窄的领域，例如以居中图像和纯色背景呈现的连衣裙 [16], 、动物 [60], 、鞋子[61], 或鸟类 [58]。相比之下，我们关注包含多个物体和属性的场景，其中干扰物体也具有相同的属性。

视觉-语言对齐。 最近，涌现了大量图像-文本对齐模型，以各种方式学习匹配图像和文本的相似度。一些模型为图像和文本使用独立的单模态编码器 [43, 22] ，而另一些模型 [52, 33, 3, 11, 50] 也使用多模态编码器。此外，还采用了诸如困难负样本挖掘 [33], 、概念蒸馏 [40], 以及保持图像-文本映射的动量 [17] 等各种策略来提升性能。我们专注于测试和改进此类模型的属性-对象绑定能力，并选择了最具奠基性的模型 CLIP [43], , 该模型被广泛应用于各种同期的视觉-语言研究/应用中 [54, 27, 47]。与近期的文本定位模型不同，我们的方法不使用任何框标注 [26, 34, 63], ； 我们还发现，这些模型在文本到图像检索方面的表现不佳。

适配基础模型。 由于从头开始训练新的视觉语言模型成本高昂，我们希望构建一个简单的适配器，以改善组合式属性-物体绑定。多项研究探索了通过提示微调 [4] 、线性探测 [65], 以及带残差连接的微调 [2], 来适配基础模型。提示微调 [13]学习词语输入的嵌入层，同时保持模型冻结。受提示微调成功的启发 [32] ，一些研究还探索了在视觉 [32], 和视觉语言 [23, 5] 领域中使用提示，以及将其用于单一属性-物体组合 [65, 48], 。即使在难度更高的设置下，我们的最优微调策略在属性-物体组合任务上的表现也显著优于提示微调和微调。我们的多模态策略与 MAPL [37]相似，不同之处在于，我们的轻量级适配器会关注语言和视觉表征，而 MAPL 仅关注语言。

3 $\mathcal{C}ola$基准

我们的目标是调整视觉语言特征，以改善属性与物体之间的组合绑定。具体而言，我们旨在提升对图像中涉及单个或多个物体及多个属性的查询进行分类的能力。图像和语言由属性、物体等原子概念构成。这些原子概念（“正方形”“盘子”）组成特定的组合（“方盘”），而场景则是由各种此类组合构成的（“白色桌子上的方盘”）。因此，我们创建了一个基准，其中使用这类原子概念的构图来构造查询，

并测试模型能否区分正确包含这些原子概念且构图正确的图像，与虽包含这些原子概念但构图错误的干扰图像。总计，$\mathcal{C}ola$ 基准包含约 1236 个组合查询，涵盖 168 个物体和 197 个属性，涉及来自 4 个数据集的约 3 万张图像。

$\mathcal{C}ol^a$包含以下讨论的两种查询类型：单物体组合和多对象查询。 .

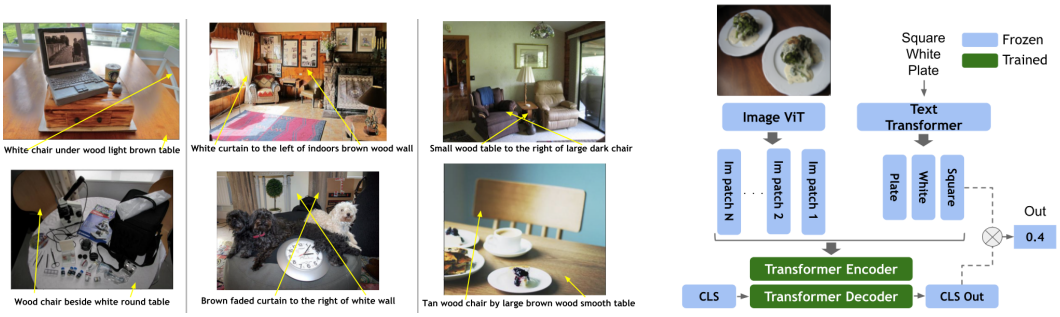
使用单物体查询进行检索。单物体查询包含多个描述同一物体的属性。例如，查询“方形白色盘子”的形式为 $Q = a_1a_2o$，其中$a_i \in A$ 取自有限的可能属性集合，而 $o \in O$ 类似地，是从有限物体集合中选取的一个类别。对于此查询，模型应将图像与正确附加到物体（“盘子”）上的属性（“方形”“白色”）关联起来，并忽略相同属性和物体的不正确搭配（例如方形桌子，而非方形盘子）。因此，这项任务是在难以区分的干扰项中进行图像检索的文本查询。我们首先创建一份查询列表，其中每个查询都包含描述某个物体的多个属性。接着，我们整理一组图像，其中至少包含一个查询词。例如，对于“方形白色盘子”，所有包含“方形”物体、“白色”物体或“盘子”的图像都会列入待检索图像列表。检索任务的目标是为属性与查询物体正确搭配的图像赋予高于其他图像的分数。我们使用三个带有物体和属性标注的数据集构建单物体查询测试集：1）GQA [18]：筛选出至少标注了 1 个属性的物体后，我们得到 320 个单对象查询，涉及 1952 张图像中的 114 个物体和 114 个属性。这些物体和属性均为常见类别，因此此划分适用于实际应用。2）CLEVR[24]：我们在 1.5 万张图像中，使用 8 种颜色、2 种材质和 2 种尺寸组合了 3 种物体形状——立方体、圆柱体和球体。3）PACO [44]：此划分类似于 GQA 中的物体。我们在 7921 张图像中构建了 400 个查询，涉及 51 个物体和 61 个属性。

多对象查询检索。 根据现有文献 [35], ，多对象查询包含多个物体，每个物体都有自己的一组属性。例如，“棕色木桌上的方形白盘”，其形式为 $Q = a_1a_2o_1 + a_3a_4o_2$，其中 $a_i \in A$ 取自有限的可能属性集合，而 $o_j \in O$ 取自有限的物体集合。在这种情况下，我们希望检查模型是否会被错误的物体和属性组合所混淆。因此，我们寻找属性与物体被互换的干扰项图像-查询对。对于查询 $Q = a_1o_1 + a_2o_2$，示例图像的形式为 $I' = a_2o_1 + a_1o_2$。换句话说，我们互换这两个物体的属性。我们精心筛选这些干扰项，以确保 $o_1 \neq o_2$ 和 $a_1 \neq a_2$。采用这一形式化方法定义的检索任务，是对正确图像与正确图像描述进行排序，使其排名高于干扰项图像：学习一种相关性编码 $f(I,\ Q)$，用于图像 I和查询 Q，使得 $f(I,\ Q) > f(I',\ Q)$ & $f(I',\ Q') > f(I,\ Q')$。测试集使用 Visual Genome [29] 数据集的测试集划分构建。

筛选 $\mathcal{C}ola$ 多目标（由众包工人完成）。我们使用 Visual Genome 数据集中的物体、属性和关系标注 [29] 来创建多对象查询。我们按照上述方法筛选掉物体和属性组合被对调的图像-字幕对。随后，我们对筛选后的测试集进行人工标注清理。我们向 I I' 和查询 Q Q' 展示10 位众包工人，并请他们选择与每个查询最相关的图像。我们只保留多数众包工人能够正确将图像与查询匹配的图像-查询对。筛选后，我们剩下210 个数据点（840 个图像-查询对），共有 1680 个可能的图像-查询匹配。在我们的验证集上，人工一致性（准确率）为83。88% ——平均而言，10 位人工评审中有 8.33 位认同第一张图像与第一条图像描述匹配，第二张图像与第二条图像描述匹配。图2a中提供了一些定性示例。

4 探索使用 $\mathcal{C}ola$的微调策略

给定一张图像（I）和一个查询（Q），$\mathcal{C}ola$ 通过衡量模型 $f(I$关联输入查询与正确图像的能力，来评估该模型。$Q)$ 现有的预训练模型在此任务上表现不佳，因为它们无法区分属性-对象组合之间的细粒度差异。因此，我们探索了微调策略，使用包含图像-语言对的数据集，其中的语言描述包含物体和属性。微调数据集的详情见第 5.3 节。我们遵循标准的微调范式，从属性-对象 $\mathcal{C}ola$ 微调数据集中采样图像和文本批次。具体而言，我们在每个批次中将正确图像与正确查询相匹配，并最小化对比学习中通常使用的 NCE损失 [43, 52]。此微调



(a) (b)

图 2：**a)** *Cola* **多物体设置验证集：**一个经过人工清理的困难验证集，用于测试属性-对象绑定。这两张图像包含相似的物体和属性，但其组合配置不同。模型必须将正确的图像与对应的图像描述匹配。**b) 最优适配策略（MM-Adapter）：**在冻结的预训练编码器之上添加一个轻量级多模态 Transformer 编码器。该多模态编码器通过对图像块和文本词元进行交叉注意力处理，构建更强的表示，从而将正确的属性关联到正确的物体。随后，对这一更强的表示进行训练，使其与冻结的文本表示对齐。

步骤旨在改善预训练视觉语言特征中属性与对象的组合绑定。这与在网络图文数据的随机批次上训练多模态层形成对比。

不相交的微调策略由于 CLIP [43] 通常用于各种任务，并且是较为轻量的视觉语言基础模型之一，因此我们主要围绕 CLIP 设计微调策略。不过，我们之后也会将这些微调策略应用于更新且更大的 FLAVA [52] 模型。CLIP [43] 由两个编码器组成：一个用于编码输入图像，另一个以类似方式编码输入文本。两个模态的输出表示分别作为图像和文本的嵌入。为了使这些模型适应特定任务/能力，研究人员通常使用线性探测或提示微调。线性探测利用微调数据集，在冻结的视觉编码器和文本编码器之上训练线性层。提示微调学习查询的词嵌入，使其适应微调数据集域，同时不改变模型的权重 [65]。其他方法会微调两个编码器的后续层。所有这些适配方法都会分别调整两个编码器的参数。

联合多模态策略 我们假设，上述常见的适应策略无法恰当地捕捉实现强属性-对象绑定所需的跨模态交互。然而，CLIP 比近期的多模态模型轻量得多。因此，我们探索了用于适应 CLIP 的轻量级多模态适应策略。我们介绍了在实验中发现的性能最佳的多模态适应策略，该策略也示意于图 2b 中。这种多模态适应借鉴了 Transformer 编码器-解码器 [53, 52, 26]，用于共同关注图像和语言表征。令 $M = [I;\ Q]$ 表示从 CLIP 的视觉编码器提取的拼接图像块表征，以及从查询中提取的词元级表征。我们使用 Transformer 编码器 M 对 $A = Att(M)$ 计算自注意力。最后，我们使用一个分类词元（随机初始化的词元），称为 [CLS],，该词元使用 Transformer 解码器对 [1] 所有自注意力特征A 进行交叉注意力，以生成输出$_{MM}$。这种对自注意力特征进行交叉注意力的方式类似于 FLAVA[52]/MDETR [26]。

大多数基于多模态编码器的预测模型通常会学习一个线性层，以对输出 [CLS] 标记嵌入 [26, 52, 33, 11]进行分类。然而，我们不学习线性预测器，而是计算 [CLS] 与冻结文本编码器中查询标记的表示 q_i 之间的余弦相似度 Q。我们认为，与在更大规模数据上训练的冻结文本编码器对齐，可以起到正则化器的作用，从而提升模型在未见组合上的表现。因此，给定图像-查询对的最终分数为 $f(I,\ Q) = \frac{1}{N_q}\sum_i^{N_q} out_{MM} \odot q_i$ 其中 N_q 是查询中的标记数量。这两种消融方法分别称为 **MM-pred** 和 **MM-adapter**

[1] 查询来自 [CLS], , 键和值来自自注意力 f 特征，A。

（称为“adapter”，因为我们可以将后者视为调整图像特征，使其更好地与文本对齐）在我们的实验中。

我们还尝试了受 FIBER 启发的多模态适配计算的各种变体 [11] 以及 ALBEF [33], , 它们在文本和图像之间使用交叉注意力。我们想强调的是，尽管探索融合视觉和语言特征的新方法是一个有价值的研究方向，但我们更感兴趣的是探索当前融合方法中最能促进组合性的共同主题，以推动未来的研究。为简洁起见，我们报告了最佳方法，并在补充材料的表5中列出了其他策略的准确率。

5 评估设置

我们在第 3 节 $\mathcal{C}ola$和 CREPE [35]数据集上，对上述两类查询进行模型评估。所有模型均使用 PyTorch [38]进行训练，并采用 Hugging Face Transformers 库 [59]。超参数等实现细节详见补充材料第 11 节。

5.1 指标

单物体查询。 对于此类查询，我们报告 $\mathcal{C}ola$ 平均精度均值[2], 即针对难以区分的干扰项计算的平均检索精度均值。我们进一步将平均精度均值区分为 **已见** 和 **未见** 查询。我们将 $\mathcal{C}ola$中的部分属性-物体对从训练集中移除，从而将其划分为已见集和未见集。例如，如果“白色方形盘子”这一组合在微调过程中未出现，则它属于未见组合；但“白色方形碗”或“方形盘子”可能出现过。在测试集中，GQA 有 320 个（150 个未见、170 个已见）查询，对应 1950 张图像；CLEVR 有 96 个（32 个未见、64 个已见）[3] 个查询，对应 22500 张图像；PACO 有 400 个（200 个已见、200 个未见）查询，对应7921 张图像。

多对象查询。 回想一下，我们有两幅图像和两条图像描述，任务是将正确的图像描述与正确的图像进行匹配。若将图像与查询的预测分数记为 $f(I,\ M)$则当 $f(I、\ M) > f(I'、\ M)\&\ f(I'和\ M') > f(I时$，我们认为预测正确 $M')$，其中图像 I 和 I' 分别与图像描述 M 和 M' 配对。根据这一标准，我们计算 $\mathcal{C}ola$多物体准确率。随机基准为 25%，因为将两条图像描述与两幅图像进行匹配共有四种方式。我们还在当代数据集 CREPE 上进行评估 [35],，其中任务是反向的。对于 CREPE，我们计算图像到文本（I2T）准确率，即模型必须从两个选项中为给定图像匹配正确的文本。请注意，在 CREPE 中，两个图像描述选项仅对应一幅图像 [35]。随机基准为50%，因为这是一个二元任务。

5.2 探索的微调策略

请注意，我们发现效果最佳的微调方法是一种多模态适配（**MM-Adapter**），用于调优第 4 节所述的预训练图像和文本特征。我们将其与常见的调优方法进行比较，例如线性探测、调优提示嵌入（提示微调），以及对整个模型（全量微调）或最后两层（后期微调）进行微调。这些适配方法分别应用于基础模型，以便进行比较。更多详情请参阅补充材料（第 11 节）。由于我们的适配方法使用多模态注意力，我们也将其与一个在预训练中使用多模态注意力的开创性模型进行比较。我们选择了 FLAVA [52]，因为它是 CLIP 之后较新的模型之一，规模更大、准确度更高 [52]，而且很容易获取其预训练权重。

5.3 微调数据集

该 $\mathcal{C}ola$训练集的构建方式也与第3 节中描述的相同，使用 GQA [18, 29], CLEVR [24]PACO [44], 和 Visual Genome [29]的训练集划分。对于 GQA，训练集划分包含 1381 个物体和 601 个属性，它们在 7.4 万张图像上构成了 27078 个查询。对于 CLEVR，我们在 7 万张图像上使用由 8 种颜色和 2 种尺寸组成的 3 种形状。最后，对于 PACO，我们有 75 个物体和 55 个属性，它们在 37883 张图像上构成了 18696 个查询。该 $\mathcal{C}ola$多目标

[2]我们还计算了 F1 分数，发现趋势保持不变；更多详情请参见补充材料第10节、第18页。[3]为避免与未见类别失衡，我们仅在32个类别上报告“已见”结果。“全部”平均精度均值基于全部96个类别计算。使用64个类别时，“已见”趋势也保持不变；更多详情请参见附录。

	GQA			CLEVR			PACO		
	All	**未见**	**Seen**	**All**	**未见**	**Seen**	**All**	**未见**	**Seen**
a. CLIP	36.53	39.06	34.24	15.38	15.01	15.32	12.21	8.64	15.79
b. +线性	40.44	42.87	38.24	47.96	29.43	46.75	14.22	6.75	21.68
c. +提示调优	37.40	40.69	34.43	29.61	23.17	28.05	12.76	5.92	19.61
d. +全量微调	38.81	40.85	36.95	52.32	19.00	47.95	14.58	6.49	22.66
e. +后期微调	42.19	44.61	40.01	64.06	27.53	67.48	15.66	8.74	22.58
f. **+MM-Pred（我们）**	**45.99**	**48.6**	**43.64**	**75.80**	**51.98**	**80.72**	**15.49**	**8.00**	**22.94**
g. **+MM-Adapter（我们）**	**46.83**	**48.86**	**44.99**	**88.21**	**89.52**	**77.00**	**18.56**	**11.47**	**25.66**
h. FLAVA	39.65	42.18	37.37	15.41	13.27	15.93	12.53	7.29	17.76
i. +线性	37.07	39.96	34.46	19.30	17.53	18.52	11.65	7.90	15.39
j. +后期微调	39.58	42.26	37.16	77.95	72.72	66.42	12.82	5.79	19.84
k. **+MM-Pred（我们的方法）**	**47.12**	**51.53**	**43.13**	**90.43**	**85.78**	**86.07**	**18.57**	**10.71**	**26.44**
l. **+MM-Adapter（我们的方法）**	**48.54**	**52.55**	**44.91**	**91.10**	**86.64**	**87.39**	**19.36**	**11.16**	**27.55**

表1：在 $\mathcal{C}ola$ 单个对象组合设置下的平均精度均值结果。我们的多模态适配（MM-Adapter）优于常见的微调方法。此外，通过多模态注意力适配图像表示（MM-Adapter），比单纯将其用作预测头（MM-Pred）具有更好的泛化能力。在 CLIP 上使用 MM-Adapter 优于对更大的 FLAVA 预训练多模态注意力层进行微调（+后期微调）。MM-Adapter 还能进一步提升 FLAVA 的表现。

复合项训练集划分包含 71,174 张图像上的 551,980 个多目标复合项。只有测试集划分使用人工标注进行了清理。对于基于 GQA [18] 和 Visual Genome [29], 构建的数据集，我们利用标注来探究不同类型数据查询的影响。我们使用 Visual Genome 中的区域描述（在表格中记为**RegionCap** ），以测试语言多样性是否优于模板化图像描述（$\mathcal{C}ola$单个对象和多目标）。我们还与来自 $\mathcal{C}ola$ 多目标图像对的困难负样本进行比较。最后，我们设置了一个综合场景，在其中使用所有数据。

6 结果

请注意，我们在两种设置下进行评估： $\mathcal{C}ola$——单物体组合设置和第3节定义的多目标复合设置。下面讨论定量结果，图3展示了一些定性结果。

6.1 $\mathcal{C}ola$单对象检索

多模态适配比其他调优方法更有效：表1中，与提示微调（行 c）、对整个 CLIP 进行微调（行 d）或对后几层进行微调（行 e）相比，调优一个参数量相同或更少的多模态注意力层，能获得更高的平均精度均值（行 f 和 g）。线性探测（行 b）虽然成本更低，但表现明显较差。这并不令人意外，因为对图像区域和文本词元进行多模态注意力处理，为模型提供了更大的灵活性，使其能够学会将正确的属性绑定到正确的物体区域。调优整个模型也不如调优后部的单模态层（行 d 与行 e）。这可能是因为微调整个模型需要更大的批次尺寸和显著更多的数据。图3展示了调优单模态层与我们提出的多模态适配方法之间的比较（因为调优单模态层的性能最为接近）。补充材料中的图 9–15 展示了其他各类适配方法的定性示例。

仅使用预训练的多模态注意力层还不够——在属性-对象组合上训练这些层才是关键：在表1中，我们看到，在 CLIP ViT B-32（151 M 参数）上使用 MM-Adapter（第 g 行）的效果优于微调 FLAVA B-16 [52]（241M 参数）模型的最后两个多模态层（第 j 行），也优于微调线性探测器（第 i 行）。令人惊讶的是，微调 FLAVA 的最后两个多模态层（第 j 行）的效果不如替换这些层，并使用 MM-Pred 和 MM-Adapter 层进行训练（第 k 和 l 行）。这表明，在适配期间（而非预训练期间）训练多模态层才是关键。

MM-Adapter 优于将多模态注意力用作预测头： 请回想一下，我们方法的消融实验之一是将多模态编码器的输出与冻结的

	多目标准确率 ↑	
	T2I *Cola*	**I2T** **CREPE** [35]
- 随机基准	25.00	50.00
- 人工	83.88	-
o. CLIP	21.42	77.43
a. +线性	30.47	87.35
b. +提示调优	27.14	80.81
c. +全量微调	34.76	82.39
d. +后期微调	36.19	87.14
e. + MM-Pred（我们的）	41.42	77.84
f. +MM-Adapter（我们的）	40.95	87.02
g. FLAVA	24.76	65.10
h. +线性	22.38	55.10
i. +后期微调	22.38	58.11
j. + MM-Pred（我们的方法）	39.04	81.37
k. +MM-Adapter（我们的方法）	40.47	74.81

(a)

数据类型	单对象 *Cola*GQA		
	All	**未见**	**Seen**
a. RegionCap	0.4711	0.4965	0.4481
b. SingleObj	0.4683	0.4886	0.4499
c. +多目标	0.4641	0.4795	0.4501
d. +困难负例	0.4688	0.4843	0.4548
e. Combined	**0.4788**	**0.4983**	**0.4612**

	多目标	
	Cola	**CREPE**
a. RegionCap	0.3114	0.8833
b. 单对象	0.2745	**0.9023**
c. +多目标	**0.3975**	0.8702
d. +困难负例	0.3483	0.8775
e. 组合	0.3893	0.8798

(b)

表2：a. 在我们的 *Cola* 任务和 CREPE 的多目标复合设置下的结果。在 CREPE 上，较简单的方法就足够了，但在 *Cola* 上则不然，这表明 *Cola* 更难。红橙黄表示准确率依次递减。CLIP 上的 MM-Adapter 和 MM-Pred 在两项任务上的平均表现都很好。b. 展示对比批次训练中所用数据类型影响的表格。在数据中加入多目标图注有助于提升 *Cola*表现，同时保持 CREPE 的性能。

文本嵌入，使其成为多模态图像特征“适配器”（MM-Adapter）。这与将多模态模块用作带线性层的预测头（MM-Pred）形成对比。如表1所示，MM-Adapter 的表现优于 MM-Pred（第 g 行与第 f 行相比），尤其是在未见类别以及 CLEVR 等详尽标注的数据集上 [24]和 PACO [44]。我们认为，与冻结的文本表示对齐起到了正则化器的作用，因为该表示是在更多数据上预训练的。

6.2 多目标检索

对于 CREPE，较简单的方法就足够了，但对于 *Cola*则不然，这表明 *Cola* 是一项更困难的任务：如表2a所示，线性探测（第 a 行）或简单微调（第 c 行和第 d 行）对于 CREPE 已经足够 [35]。然而，我们的 MM-Adapter 在 *Cola* 上进一步提升了表现（第 f 行与第 a、b、c、d 行相比），同时保持了在 CREPE 上的表现。这也表明文本到图像匹配比图像到文本匹配更困难，这一点在 Winoground 中也有所体现 [55]。

基线 CLIP 和 FLAVA 的表现低于随机水平：如果我们使用开箱即用的 CLIP [43] 和 FLAVA [52] 在我们的 *Cola* 数据集上进行评估，可以从表 2a 中看到，它们的表现低于随机基准（第 o 行和第 g 行）。这一结果与 Winoground [55]中的研究发现一致。

从头开始训练后期多模态层的效果，CLIP+*MM-Adapter* 整体表现更佳：如表 2a的 e、f、j、k 行所示，我们通过在 CLIP 和 FLAVA 的基础上从头开始训练多模态层，提升了性能。有趣的是，如表 2a所示，对 FLAVA 预训练的多模态层进行线性探测或微调会损害性能（g 行与 h、i 行相比）。这可能是因为微调会对在大规模数据上训练的参数造成不利扰动。最后，CLIP+MM-Adapter（f 行）的表现与在 FLAVA 上微调多模态层（f 行与 j、k 行相比）相当。

6.3 微调数据的影响

自由形式的描述与模板化 *Cola* 查询之间的差异很小。因此，*Cola*模板化查询很有用：在表 2b （第 a 行与第 c、d、e 行相比），我们看到模板化查询的表现

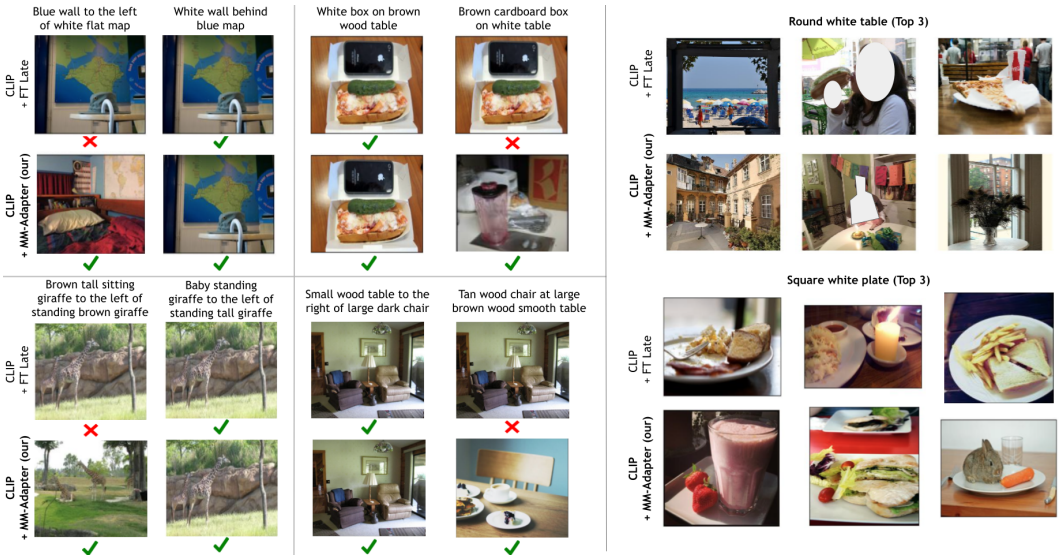


图3：多目标匹配（左）和检索具有多个属性的单个对象（右）的定性结果。



图4：模型难以处理多个物体-属性组合的案例定性结果（左）。单对象组合检索效果提升最大和最小的案例如右侧所示。

与自由形式的区域描述一样好。这表明 $Cola$ 查询对于属性-对象绑定仍然有用，尽管不是自由形式的。

在图像描述中包含多个物体及多个属性会有所帮助： 当我们将单个对象图像描述与多目标图像描述结合时，可以在表 2b 中看到，性能得到提升，尤其是在多目标 $Cola$ 设置下，同时保持了单对象设置下的性能（第 b 行与第 c 行对比）。多目标和单对象图像描述由数量相同的图像构成。将所有类型的数据与困难负样本结合（第 e 行）似乎不会对性能产生太大影响。

6.4 $Cola$Benchmark 的质量

我们的 $Cola$ mAP 更难，提升也更困难：我们还根据 mAP 的标准定义计算了性能，其中所有图像都包含在待检索列表中

。相比之下，$\mathcal{C}ola$mAP 仅纳入第3节所定义的困难干扰项。可以看到，采用标准 mAP 时，趋势依然相同。不过，我们的$\mathcal{C}ola$mAP 更难提升。标准 MAP（补充材料，表 4）在 GQA 上提升了 2 倍以上，在 CLEVR 上提升了 10 倍。相比之下，采用难度更高的$\mathcal{C}ola$MAP，我们在 GQA 上仅提升了 1.09 倍，在 CLEVR 上提升了 2 倍。

颜色歧义、空间关系和尺寸是模型在$\mathcal{C}ola$基准测试中表现不佳的一些常见主题：我们分析了模型在我们的基准测试中难以处理的构图查询类型。从图4所示的定性示例中可以看出，对于涉及多个物体的情况，包含模糊颜色（由光照引起）和空间关系的构图较难处理。这可能是因为训练数据中对空间关系的标注往往不一致——例如，“在...左侧”有时可能是以观察者视角或图像视角为参照。更多示例见补充材料中的图 8（数据歧义）、图 20 和图 21（预测不准确）。

在非显著/被遮挡物体上取得显著提升。在较大物体上的提升微乎其微：对于单个对象的属性-物体组合检索，正如图中所示，我们的适配方法带来的最大提升出现在非显著或被遮挡的物体上（例如小标志）。4 （右）。训练集中常见的查询（例如“晴朗的蓝天”——天空最常见的颜色是蓝色）从预训练模型中获得的提升微乎其微。

7 讨论与局限性

本研究通过微调视觉语言模型，检验了用于组合式属性-对象检索的设计选择。因此，与原始预训练模型相比，[43], 我们可能会失去一些其他通用能力，例如问答、生成图像描述等。例如，我们最好的适配模型在 CIFAR 上的零样本准确率为 83%10 [30], ，而预训练 CLIP 的准确率为 87%。不过，仅使用5% 的 CIFAR 训练数据（不到 1 个 epoch），我们就达到了92% 的准确率，同时在我们的 $\mathcal{C}ola$ 任务上保持了性能。我们还想强调，我们的目标不是训练一个新的基础模型，而是探索能够实现组合推理的设计选择。虽然我们关注属性-对象组合性，但其他类型的组合结构仍有很大的探索空间，例如关系、场景图和计数 [15, 12]。未来研究的一条重要方向是，在更大规模的图像集上收集更详细的标注（关于查询中物体和构图的类型），以帮助找出模型表现不佳的主题。还需要进一步测试模型在涉及敏感属性和物体时的表现——它是否会因数据中的种族或政治偏见而倾向于将错误的属性与物体关联起来 [6]。此外，有必要结合新近提出的视觉语言模型重新评估我们的结果。最后，我们想指出，我们的微调策略并非专为构图设计，这表明它也有望应用于其他下游任务，尤其是类似的策略和模块已用于其他计算机视觉任务 [53, 52, 26]。

8 结论

我们提出了一项新任务 $\mathcal{C}ola$，用于测试视觉语言模型的组合式属性-对象绑定能力。这对于各种实际应用都很重要，例如辅助代理需要理解杂乱工作环境中物体之间的细粒度差异。我们探索了适配大型视觉语言模型的架构选择，以促进此类推理。我们证明，轻量级多模态适配器可以提升预训练视觉语言模型的这项能力，并为后续研究提供强有力的基线。我们希望 $\mathcal{C}ola$ 能成为一个强有力的基准，而我们的适配方案能成为提升组合式视觉语言智能的强有力基线。

致谢。我们感谢 Dhruv Mahajan 和 Jang Hyun (Vincent) Cho 在项目初期给予的宝贵指导。我们也感谢匿名评审提出的深思熟虑的评论和建议。本材料部分基于 DARPA 根据协议编号 HR00112020054 提供的资助，该协议授予波士顿大学的 Kate Saenko 和 Bryan Plummer。本材料中表达的任何观点、发现、结论或建议均属于作者本人，并不一定反映资助机构的观点。

参考文献

[1] Aishwarya Agrawal、Aniruddha Kembhavi、Dhruv Batra 和 Devi Parikh。C-VQA：视觉问答（VQA）v 1 0 2017 1704 08243 [ ] 3 的构成性划分。数据集，4月。。[2] arXiv:。计算机科学。Guillaume Alain 和 Yoshua Bengio。使用线性分类器探针理解中间层，2018年11月。arXiv:1610.01644 [计算机科学，统计学]。3[3] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katie Millican、Malcolm Reynolds、Roman Ring、Eliza Rutherford、Serkan Cabi、Tengda Han、Zhitao Gong、Sina Samangooei、Marianne Monteiro、Jacob Menick、Sebastian Borgeaud、Andrew Brock、Aida Nematzadeh、Sahand Sharifzadeh、Mikolaj Binkowski、Ricardo Barreira、Oriol Vinyals、Andrew Zisserman 和 Karen Simonyan。Flamingo：用于少样本学习的视觉语言模型，2022年11月。arXiv:2204.14198 [计算机科学]。3[4] Amit Alfassy、Assaf Arbelle、Oshri Halimi、Sivan Harary、Roei Herzig、Eli Schwartz、Rameswar Panda、Michele Dolfi、Christoph Auer、Kate Saenko、PeterW J. Staar、Rogerio Feris 和 Leonid Karlinsky。FETA：面向专家任务应用的基础模型专项化，2022年12月。arXiv:2209.03648 [计算机科学]。3[5] Hyojin Bahng、Ali Jahanian、Swami Sankaranarayanan 和 Phillip Isola。探索用于适配大规模模型的视觉提示，2022 2203 17274 [ ] 3 年6月。arXiv:。计算机科学。Joy Buolamwini 和 Timnit Gebru。性别阴影：商业性别分类中的交叉准确率差异。发表于公平性、问责制与透明度会议，第77–91页。PMLR，2018 10。Noam Chomsky。《句法理论的若干问题》。麻省理工学院出版社，第50版，1965年。3[8] Gordon Christie、Ankit Laddha、Aishwarya Agrawal、Stanislaw Antol、Yash Goyal、Kevin Kochers- berger 和 Dhruv Batra。共同消解语言与视觉歧义：带图像描述场景中的联合分割与介词附着消解，2016 1604 02125 [ ] 3 年9月。arXiv:。计算机科学。[9] MJ Cresswell。逻辑与语言。。[10] M. J. Cresswell。逻辑与语言。《Synthese》，40(2):375–387，1973年。出版商：Springer。3[11] Zi-Yi Dou、Aishwarya Kamath、Zhe Gan、Pengchuan Zhang、Jianfeng Wang、Linjie Li、Zicheng Liu、Ce Liu、Yann LeCun、Nanyun Peng、Jianfeng Gao 和 Lijuan Wang。通过骨干网络融合进行由粗到细的视觉语言预训练，2022年11月。arXiv:2206.07643 [计算机科学]。3、5、6、18[12] Mona Gandhi、Mustafa Omer Gul、Eva Prakash、Madeleine Grunde-McLaughlin、Ranjay Krishna 和 Maneesh Agrawala。衡量视频问答中的构成一致性。发表于IEEE/CVF计算机视觉与模式识别会议论文集，第5046–5055页，2022年。10[13] Peng Gao、Shijie Geng、Renrui Zhang、Teli Ma、Rongyao Fang、Yongfeng Zhang、Hongsheng Li 和 Yu Qiao。CLIP-Adapter：通过特征适配器打造更好的视觉语言模型，2021 2110 04544 年10月。arXiv:。计算机科学。[ ] 3[14] Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna Wallach、Hal Daumé III au 2, 2021 14 和 Kate Crawford。数据集数据表，。[15] Madeleine Grunde-McLaughlin、Ranjay Krishna 和 Maneesh Agrawala。Agqa：用于构成性时空推理的基准。发表于IEEE/CVF计算机视觉与模式识别会议论文集，第 11287 11297, 2021 10 –页。Sheng Guo、Weilin Huang、Xiao Zhang、Prasanna Srikhanta、Yin Cui、Yuan Li、Hartwig Adam、Matthew R. Scott 和 Serge Belongie。iMaterialist时尚属性数据集。发表于2019年IEEE/CVF国际计算机视觉会议研讨会（ICCVW），第3113–3116页，韩国首尔，2019年10月。IEEE。[17] Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。用于无监督视觉表征学习的动量对比，2020 1911 05722 [ ] 3 年3月。arXiv:。计算机科学。Drew A. Hudson 和 Christopher D. Manning。GQA：用于真实世界视觉推理与构成性问答的新数据集，2019 1902 09506 [ ] 2, 4, 6, 7, 15, 17, 18, 20 年5月。arXiv:。计算机科学。Phillip Isola、Joseph J. Lim 和 Edward H. Adelson。在图像集合中发现状态与变换。发表于2015年IEEE计算机视觉与模式识别会议（CVPR），第 1383 1391, 2015 3 –页，美国马萨诸塞州波士顿，6月。IEEE。Jingwei Ji、Ranjay Krishna、Li Fei-Fei 和 Juan Carlos Niebles。动作基因组：将动作表示为时空场景图的构成。发表于IEEE/CVF计算机视觉与模式识别会议论文集，第10236–10247页，2020年。1[21] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。通过带噪文本监督扩展视觉与视觉语言表征学习。发表于国际机器学习会议，第 4904 4916 2021 2 –页。PMLR，。Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc V. Le、Yunhsuan Sung、Zhen Li 和 Tom Duerig。通过带噪文本监督扩展视觉与视觉语言表征学习，2021年6月。arXiv:2102.05918 [计算机科学]。3[23] Menglin Jia、Luming Tang、Bor-Chun Chen、Claire Cardie、Serge Belongie、Bharath Hariharan 和 Ser-Nam Lim。视觉提示调优，2022 2203 12119 [ ] 3 年7月。arXiv:。计算机科学。Justin Johnson、Bharath Hariharan、Laurens van der Maaten、Li Fei-Fei、C. Lawrence Zitnick 和 Ross Girshick。CLEVR：用于构成性语言与初级视觉推理的诊断数据集，2016年12月。arXiv:1612.06890 [计算机科学]。2、4、6、8、15、17、18

[25] Aishwarya Kamath、Sara Price、Jonas Pfeiffer、Yann LeCun 和 Nicolas Carion。TRICD：通过上下文短语检测测试稳健的图像理解。 2023 1, 3 。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，2022年10月。 2021 2104 12763 。arXiv：。计算机科学。[27] [ ] 2, 3, 5, 10, 18, 19 Gwanghyun Kim、Taesung Kwon 和 Jong Chul Ye。DiffusionCLIP：用于稳健图像操控的文本引导扩散模型，2022年8月。arXiv:2110.02711 [计算机科学]。3[28] 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。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):32–73，2017年5月。1[29] 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 Fei-Fei Li。Visual Genome：利用众包密集图像标注连接语言与视觉，2月。 2016 1602 07332 。arXiv：。计算机科学。[30] [ ] 4, 6, 7, 15 Alex Krizhevsky、Geoffrey Hinton 等。学习来自微小图像的多层特征。 2009 10 。Michael A Lepori、Thomas Serre 和 Ellie Pavlick。拆解它：神经网络中结构组合性的证据。arXiv预印本，arXiv: *2301 10884*, 2023 3 。。Brian Lester、Rami Al-Rfou 和 Noah Constant。规模对参数高效提示调优的力量，2022年9月。 2021 2104 08691 [ ] 3 。arXiv：。计算机科学。Junnan Li、Ramprasaath R. Selvaraju、Akhilesh Deepak Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Hoi。在融合之前对齐：通过动量蒸馏进行视觉与语言表征学习，2021年10月。arXiv:2107.07651 [计算机科学]。3、5、6、18[34] Liunian Harold Li、Pengchuan Zhang、Haotian Zhang、Jianwei Yang、Chunyuan Li、Yiwu Zhong、Lijuan Wang、Lu Yuan、Lei Zhang、Jenq-Neng Hwang、Kai-Wei Chang 和 Jianfeng Gao。基于定位的语言-图像预训练，2022年6月。 2022 2112 03857 [ ] 3 。arXiv：。计算机科学。Zixian Ma、Jerry Hong、Mustafa Omer Gul、Mona Gandhi、Irena Gao 和 Ranjay Krishna。CREPE：视觉语言基础模型能否进行组合推理？，2023年1月。arXiv:2212.07796 [计算机科学]。1、2、3、[36] 4, 6, 8 Oscar Mañas、Pau Rodriguez、Saba Ahmadi、Aida Nematzadeh、Yash Goyal 和 Aishwarya Agrawal。MAPL：用于视觉语言少样本提示的单模态预训练模型参数高效适配，2022年10月。 2022 2210 07179 [ ] 3 。arXiv：。计算机科学。Nihal V. Nayak、Peilin Yu 和 Stephen H. Bach。学习组合软提示以实现组合零样本学习，2022年9月。arXiv:2204.03574 [计算机科学]。3[38] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Köpf、Edward Yang、Zach DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。PyTorch：一种命令式风格的高性能深度学习库，2022年12月。 2019 1912 01703 [ ] 6 。arXiv：。计算机科学、统计学。Xavier Puig、Kevin Ra、Marko Boben、Jiaman Li、Tingwu Wang、Sanja Fidler 和 Antonio Torralba。VirtualHome：通过程序模拟家庭活动。载于2018年IEEE/CVF计算机视觉与模式识别会议，8494–8502页，美国犹他州盐湖城，2018年6月。IEEE。3[40] Filip Radenovic、Abhimanyu Dubey、Abhishek Kadian、Todor Mihaylov、Simon Vandenhende、Yash Patel、Yi Wen、Vignesh Ramanathan 和 Dhruv Mahajan。视觉语言预训练中的过滤、蒸馏与困难负样本，2023年1月。 2023 2301 02280 [ ] 3 。arXiv：。计算机科学。Filip Radenovic、Animesh Sinha、Albert Gordo、Tamara Berg 和 Dhruv Mahajan。大规模属性-对象组合，2021年5月。arXiv:2105.11373 [计算机科学]。3[42] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，8748–8763页。PMLR，2021年。2[43] 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。通过自然语言监督学习可迁移视觉模型，2021年2月。arXiv:2103.00020 [计算机科学]。2、3、4、5、8、10、19、20[44] Vignesh Ramanathan、Anmol Kalia、Vladan Petrovic、Yi Wen、Baixue Zheng、Baishan Guo、Rui Wang、Aaron Marquez、Rama Kovvuri、Abhishek Kadian、Amir Mousavi、Yiwen Song、Abhimanyu Dubey 和 Dhruv Mahajan。PACO：常见物体的部件与属性，2023年1月。arXiv:2301.01795 [计算机科学]。2、4、[45] 6, 8, 14, 15, 17, 18 Arijit Ray、Gordon Christie、Mohit Bansal、Dhruv Batra 和 Devi Parikh。VQA中的问题相关性：识别非视觉问题与错误前提问题，2022年9月。 2016 1606 06622 [ ] 3 。arXiv：。计算机科学。Arijit Ray、Karan Sikka、Ajay Divakaran、Stefan Lee 和 Giedrius Burachas。外面晴朗还是阴暗？！通过蕴含式问题生成提高VQA中的答案一致性，2022年9月。 2019 1909 04696 。arXiv：。计算机科学。[47] [ ] 3 Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。利用潜在扩散模型进行高分辨率图像合成，2022年4月。arXiv:2112.10752 [计算机科学]。3

[48] Kuniaki Saito、Kihyuk Sohn、Xiang Zhang、Chun-Liang Li、Chen-Yu Lee、Kate Saenko 和 Tomas Pfister。前缀条件化统一语言与标签监督，2022年6月。arXiv:2206.01125 [计算机科学]。3[49] Kuniaki Saito、Kihyuk Sohn、Xiang Zhang、Chun-Liang Li、Chen-Yu Lee、Kate Saenko 和 Tomas Pfister。Pic2Word：将图像映射为词语以实现零样本组合图像检索，2023年2月。arXiv: 2302 03084 [ ] 3。计算机科学。Madeline Chantry Schiappa、Michael [50] Cogswell、Ajay Divakaran 和 Yogesh Singh Rawat。探究大型视觉语言模型的概念理解，2023 3。Meet Shah、Xinlei Chen、Marcus Rohrbach 和 Devi [51] Parikh。循环一致性：实现稳健的视觉问答。载于2019年IEEE/CVF计算机视觉与模式识别会议（CVPR），6642–6651页，美国加利福尼亚州长滩，2019年6月。IEEE。3[52] Amanpreet Singh、Ronghang Hu、Vedanuj Goswami、Guillaume Couairon、Wojciech Galuba、Marcus Rohrbach 和 Douwe Kiela。FLAVA：基础语言与视觉对齐模型，2022年3月。arXiv: 2112 04482 [ ] 2, 3, 4, 5, 6, 7, 8, 10, 18, 19。计算机科学。Hao Tan 和 Mohit Bansal。LXMERT：从 [53] Transformer中学习跨模态编码器表征，2022年12月。2019 1908 07490 [ ] 2, 5, 10。arXiv：。计算机科学。Reuben Tan、Arijit Ray、Andrea Burns、Bryan A. Plummer、Justin Salamon、Oriol Nieto、Bryan Russell [54] 和 Kate Saenko。通过三模态一致性进行语言引导的音视源分离，2023年。3[55] Tristan Thrush、Ryan Jiang、Max Bartolo、Amanpreet Singh、Adina Williams、Douwe Kiela 和 Candace Ross。Winoground：探究视觉与语言模型的视觉语言组合性，2022年4月。2022 2204 03162 [ ] 1, 2, 3, 8。arXiv：。计算机科学。Ben Usman、Dina Bashkirova 和 Kate Saenko。[56] 通过受限信息流进行解耦无监督图像转换，2022年11月。2021 2111 13279 [ ] 3。arXiv：。计算机科学。Ashish Vaswani、Noam [57] Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need，2022年12月。2017 1706 03762 [ ] 18。arXiv：。计算机科学。 [58] Catherine Wah、Steve Branson、Peter Welinder、Pietro Perona 和 Serge Belongie。加州理工学院-加州大学圣迭戈分校鸟类200-2011数据集，2011年7月。专题：2010-001；页数：8；编号：2010-001；地点：加利福尼亚州帕萨迪纳；出版方：加州理工学院。3[59] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Remi 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 Rush。Transformer：最先进的自然语言处理。载于2020年自然语言处理实证方法会议：系统演示，38 45, 2020 6 –线上，2022年10月。计算语言学协会。 [60] Yongqin Xian、Christoph H. Lampert、Bernt Schiele 和 Zeynep Akata。零样本学习——对优、劣与丑的全面评估，2022年9月。2020 1707 00600 [ ] 3。arXiv：。计算机科学。Aron Yu 和 [61] Kristen Grauman。通过局部学习进行细粒度视觉比较。载于2014年IEEE计算机视觉与模式识别会议，192–199页，美国俄亥俄州哥伦布，2022年6月。IEEE。Mert Yuksekgonul、Federico Bianchi、Pratyusha Kalluri、Dan Jurafsky 和 James Zou。视觉语言模型何时以及为何表现得像词袋，以及如何应对？，2022年10月。2022 2210 01936。arXiv：。计算机科学。[63] [ ] 3 Haotian Zhang、Pengchuan Zhang、Xiaowei Hu、Yen-Chun Chen、Liunian Harold Li、Xiyang Dai、Lijuan Wang、Lu Yuan、Jenq-Neng Hwang 和 Jianfeng Gao。GLIPv2：统一定位与视觉语言理解，2022年10月。arXiv:2206.05836 [计算机科学]。3[64] 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（CVPR）论文集，5579–5588页，2021年6月。[65] 2 Kaiyang Zhou、Jingkang Yang、Chen Change Loy 和 Ziwei Liu。学习提示：面向视觉语言模型。《国际计算机视觉杂志》，130(9):2337–2348，2022年9月。arXiv:2109.01134 [计算机科学]。3、5

9 额外数据详情

Ourdata 可在公开数据中获取，网址为 https://github.com/arijitray1993/COLA。请注意，我们的数据集建立在四个公开数据集之上。我们将进一步介绍数据集的整理方式。

*Cola*单个对象数据

GQA GQA 的图像中包含物体和属性的标注。我们利用这些标注构造诸如“白色方盘子”之类的查询。我们忽略边界框。在测试时，我们筛选出测试集划分中每个物体标注至少包含 2 个属性的图像。最终，测试集中有 1952 张图像和 320 条查询。为了创建一个查询均未见过的挑战性数据集，我们从测试集的 320 条查询中选取训练集中出现频率最低的 150 个属性-物体元组，并将这些

图像和查询从训练集中完全移除。这样，我们最终得到 150 条未见查询，同时对训练集规模的影响最小。我们在这个测试集上报告所有结果，其中包含 150 条未见查询和 170 条已见查询。我们使用 GQA 训练集划分进行训练（移除了测试集中未见查询及其对应的图像）。因此，训练集约有 6.7 万张图像和 2.7 万条查询。配对样例数量为 45 万个图文对。

CLEVR 在 CLEVR 上，我们在 22,500 张图像中的 96 个类别上进行测试。我们采用其组合划分。按照其论文和数据集网站中的描述，我们在条件 A 上进行训练，并在条件 B 上进行测试。在这两个划分中，立方体和圆柱体具有未见的颜色和大小组合。然而，对于球体，所有颜色和形状都是已见的。由于平均精度均值对类别数量较为敏感，我们使已见类别和未见类别的数量保持一致。因此，在报告已见与未见的结果时，我们不计入球体。但是，对于全部平均精度均值，我们会将球体计入其中。因此，“未见”有 32 个类别，“已见”有 32 个类别，而“全部”有 96 个类别。在训练中，我们使用 70K 张图像上的 168 个可能查询（立方体和圆柱体的颜色替换为测试集中的颜色）。

PACO PACO [44] 数据集在 9443 张测试集图像中的 75 个物体类别上标注了 55 个属性。由于物体和属性的所有组合会产生数量难以处理的可能构图，我们从测试集中采样了 400 个实际出现的多属性-对象组合。这 400 个类别通过采样排名前 200 的已见属性-物体查询和排名前 200 的未见属性-物体查询得到。如果某个物体与属性的组合在训练数据中从未是该物体对应属性集合的子集，则该属性-物体查询被定义为未见。这样，我们得到覆盖 7921 张图像的 400 个类别，并在其上报告结果。我们有 37K 张训练图像、18K 个查询以及 55K 个成对图像-文本样例。

$Cola$ **多对象数据**多对象数据是根据以下划分创建的：训练集和测试集划分$Cola$ 单对象 GQA 数据划分。只有测试集划分使用人工标注进行清理。图 7 展示了人工清理的测试集中的一些定性示例。我们也发现，即使经过人工清理，仍有一些案例存在歧义，如图 8 所示。这些情况通常涉及对大小（大与小）和颜色（不同光照条件下的蓝色与白色）的感知差异。不过，尽管存在少量噪声，10 名独立标注人员在我们的验证集上的准确率为 84%。相比之下，我们表现最好的模型准确率为 45%。因此，我们认为仍有很大的改进空间。

数据集数据表答案

我们认为，数据集数据表论文中的大多数问题 [14] 已经在正文中得到了回答。在此，我们补充一些答案。此外，我们还在图6中提供了数据中物体类型和属性的直方图。

数据集资助机构本项目由授予 KS 和 BP 的 DARPA Semafor 资助。论文中报告的发现和结果并不代表美国政府或国防部的观点。

数据集是否包含所有可能的实例，还是从更大的集合中抽取的实例样本（不一定是随机样本）？ 该数据集是从更大的数据集中精选的样本，专门用于测试模型的属性-对象组合性。

每个实例由哪些数据组成？ 原始图像、文本说明，以及对图像中物体和属性的基于文本的场景图。请注意，部分物体和属性标注可能缺失。

单个实例是否缺少任何信息？ 是的，由于要详尽标注所有可能的物体和属性非常困难，因此部分标注可能缺失。

该数据集是自包含的，还是会链接到或以其他方式依赖外部资源（例如网站、推文或其他数据集）？ 由于该数据集基于公开可用的数据集构建，部分标注（如场景图）链接到了外部数据集。

数据集是否包含可能被视为机密的数据？ 据我们所知并不包含，因为我们使用的是已发布数据集中的公开数据。

该数据集是否包含直接查看时可能令人反感、受辱、受威胁或以其他方式引发焦虑的数据？ 据我们所知，没有。

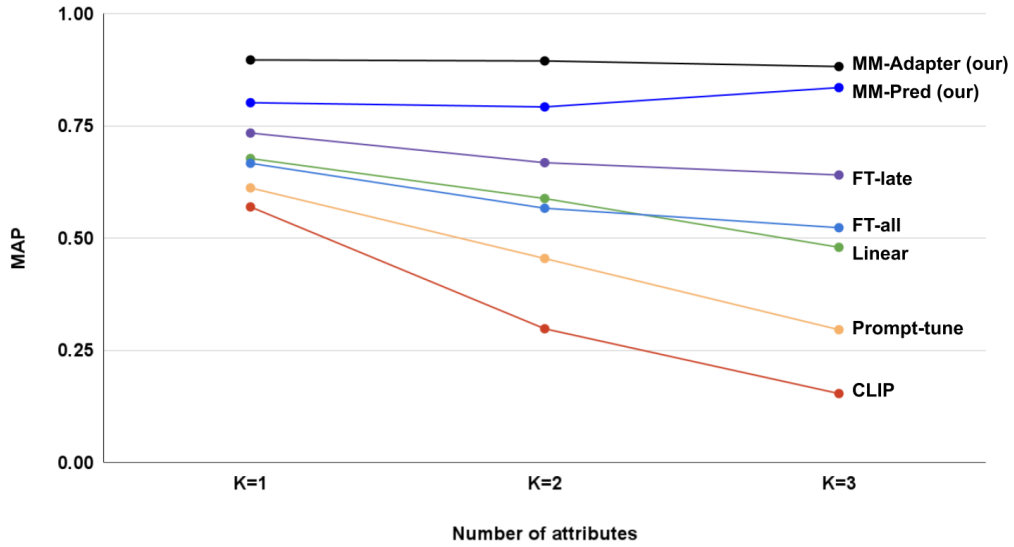


图 5：CLEVR 数据集上查询中属性数量对应的平均精度均值。请注意，即使属性数量逐渐增加，MM-Adapter 仍表现出色。

该数据集是否标识任何子群体（例如按年龄、性别划分）？ 数据中不包含任何个人身份信息。我们也不会针对种族、年龄、性取向或宗教等敏感属性开展分析。图像中可能标注了性别和发色等属性，但我们不会对其进行专门分析，因为这并非该基准测试或论文的研究重点。

能否直接或间接（即结合其他数据）从该数据集中识别出个人（即一位或多位自然人）？ 数据中不包含个人身份信息。现实生活中的某个人可能出现在数据图像中，这纯属巧合。

该数据集是否包含任何可能被视为敏感的数据？ 据我们所知，没有。

数据是在什么时间段内收集的？ Visual Genome [29]于 2016 年收集。GQA[18] 于 2018 年收集。PACO [44] 于 2022—23 年收集。CLEVR [24] 于 2016 年收集。

谁参与了数据收集过程（例如，学生、众包工作者、承包商）？他们获得了怎样的报酬？ 学生负责运行数据收集流程，众包工作者负责标注数据。我们不知道他们为我们基于其构建的数据集获得了多少报酬。不过，对于我们经人工清理的多目标 $\mathcal{C}ola$ 测试集，我们向众包工作者支付的平均报酬为每小时 15 美元；如果他们正确标注示例，还会获得奖金。

是否开展过任何伦理审查流程？ 是的，由于该数据不涉及任何个人信息或敏感信息，我们获得了机构审查委员会（IRB）的豁免。

是否对该数据集及其使用对数据主体的潜在影响（例如，数据保护影响分析）进行了分析？ 没有，但这可能是未来一项有趣的工作。

用于预处理、清理或标注数据的软件是否可用？ 是的，我们将发布用于整理数据的代码。

该数据集是否会在版权许可下分发？知识产权限制和出口管制限制有哪些？ 不存在此类限制。许可证与构成我们基准的数据集所使用的许可证相同。

数据集会更新吗？ 我们可能会定期添加更多示例或标注来更新数据。

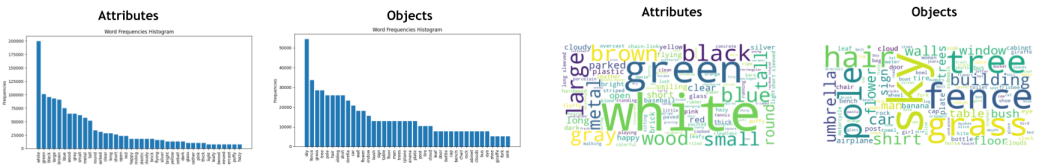


图 6：我们的数据集中各类属性和物体的概览。它们对应于日常生活中的实际物体。

10 更多指标与分析

请记住，我们的目标是调整视觉语言特征，以改善属性与物体之间的组合绑定。具体而言，我们旨在提升对图像中包含单个或多个物体及多个属性的查询进行分类的能力。因此，我们进行了一些分析，考察属性数量增加对性能的影响。请记住，我们还报告了基于 $\mathcal{C}ola$ 平均精度均值 的数值，该指标评估模型从属性与物体绑定错误的困难干扰项中，将属性正确绑定到目标物体的图像排在前面的能力。我们还展示了采用其他平均精度均值方案以及常用标准平均精度均值时的结果。我们展示了主论文中所选的平均精度均值方案为何更难，尽管采用所有方案时的趋势都保持一致。最后，我们还展示了 MM-Adapter 和 MM-Pred 方法采用其他多模态融合方案时的性能，结果表明，这种适配策略同样适用于多种其他方案。

按属性数量评估性能 我们在单个对象设置下改变查询中的属性数量，并检查属性数量增加时的性能。结果如图 5 所示。我们发现，基线 CLIP 和微调在单属性查询上的性能高于多属性查询。我们的 MM-Adapter 在单属性和多属性情况下均保持了更优的性能。

其他评估指标

QueryAll 平均准确率均值 请注意，在主论文中，我们针对$\mathcal{C}ola$ 单对象设置计算困难干扰项中的平均精度均值。困难干扰项是包含查询中属性和物体词语任意一个的图像。模型需要将属性正确附加到所需的物体上的图像（而不是仅仅存在于某处的图像）排在前面，才能获得更高的平均精度均值。在此，我们设计了另一种类似的困难平均精度均值。这里，我们将候选池中的图像列表限制为包含全部查询属性和物体的图像。因此，对于查询“青色金属圆柱体”，我们在同时包含“圆柱体”AND“青色”物体AND“金属”物体的图像中进行排序。在主论文中，平均精度均值使用的是 OR 而不是 AND 操作。结果如表3 所示，我们观察到，使用该指标时，所有趋势仍与主论文中的结果相同。不过，我们发现，该指标只能应用于 CLEVR，因为其标注是完整的。在 GQA 等真实数据集中，这类已标注困难干扰项的数量有限；因此，我们使用 OR 操作，以保留较多可供排序的图像。应用于 GQA 等数据集时，趋势相同，但由于干扰项图像极少，数值会虚高。

真实目标平均排名（✓）与干扰项（×） 基于我们为上文 QueryAll 平均准确率均值构造的困难干扰项，我们还报告了正确将属性与物体匹配的图像的平均排名，以及错误将属性与物体匹配的图像的平均排名。结果也列于表 3 中。我们观察到，所有趋势都保持不变，因此正文中只报告其中一种。

标准平均精度均值 与我们的困难 $\mathcal{C}ola$平均精度均值不同，我们还计算了验证集内全部图像的平均精度均值，不论干扰项是困难还是简单。再次可以看到，所有趋势都保持不变，如表 4所示。不过，我们注意到，全部图像上的平均精度均值数值低得多。这有两个原因：a）需要进行排名的图像数量更多；b）GQA 等数据集存在缺失的标注，因此许多图像由于缺少标注而被标记为负检索结果。当我们将图像限制为至少包含一个查询词时，这种噪声会有所减少。不过请注意，相较于正文中报告的更困难的 $\mathcal{C}ola$ 平均精度均值，整体平均精度均值更容易得到提升。这表明模型可以迅速

	QueryAll 平均准确率均值 ↑			平均排名	
	All	Seen	未见	✓ ↓	× ↑
CLIP+线性	49.56	48.21	31.78	21.78	54.17
提示调优	31.07	29.34	29.43	31.37	52.72
全部微调	54.58	52.05	21.27	21.06	54.32
后期微调	66.16	70.15	30.99	14.03	55.35
MM-Pred（我们的）	85.51	77.85	81.18	9.69	55.99
MM-Adapter（我们的）	**90.35**	**81.95**	**90.53**	**8.85**	**56.12**

表 3：在 CLEVR [24] 数据集上计算的硬指标的另外两种选择。

		总体平均精度均值		
		All	**未见**	**Seen**
GQA [18]	CLIP	0.65	0.35	0.91
	+ 提示调优	8.72	9.63	7.91
	+ 线性探测	12.81	13.29	12.52
	+ 全量微调	11.47	10.87	11.96
	+ 后期微调	13.39	13.70	13.10
	+MM-Pred（我们的方法）	16.76	17.45	16.15
	+MM-Adapter（我们的方法）	17.40	16.79	17.95
	FLAVA	7.33	6.43	8.15
	+ 后期微调	9.71	9.49	9.90
	+MM-Pred（我们的）	17.68	19.24	16.29
	+MM-Adapter（我们的）	20.03	20.70	19.42
CLEVR [24]	CLIP	6.42	6.36	6.29
	+ 提示调优	29.42	23.02	27.79
	+ 线性探测	47.83	29.33	46.54
	+ 全量微调	51.99	18.40	47.63
	+ 后期微调	63.93	27.20	67.00
	+MM-Pred（我们的方法）	83.40	76.82	76.10
	+MM-Adapter（我们的方法）	88.15	89.40	76.90
	FLAVA + 线性	18.76	16.77	17.82
	+ 后期微调	77.59	71.91	66.25
	+MM-Pred（我们的）	90.41	85.74	86.05
	+MM-Adapter（我们的方法）	91.08	86.60	87.39
PACO [44]	CLIP	0.71	0.11	1.31
	+ 提示调优	6.19	2.78	9.61
	+ 线性探测	8.22	3.83	12.61
	+ 全量微调	7.19	3.00	11.38
	+ 后期微调	9.29	5.37	13.21
	+MM-Pred（我们的）	9.63	4.00	15.26
	+MM-Adapter（我们的）	10.00	6.50	15.22
	FLAVA + 线性	3.45	1.73	5.17
	+ 后期微调	6.31	2.00	10.60
	+MM-Pred（我们的）	10.77	4.77	16.76
	+MM-Adapter（我们的）	12.02	6.36	17.67

表 4：测试集内标注的物体上具有多个属性的全部图像的标准平均精度均值（而不仅仅是像我们的 $\mathcal{C}ola$MAP 那样的困难干扰项）。请注意，我们的表现可以大幅提升（例如，在 GQA 划分上从 0.65 提升至 17.40，达到 10 倍），但在我们的 $\mathcal{C}ola$ MAP 上提升幅度要小得多，因为它仅针对困难干扰项。

			*Cola*单目标平均精度		
		参数量	All	未见	Seen
GQA [18]	单模态	13M	42.19	44.61	40.01
	FLAVA	9.9M	47.43	48.95	46.05
	ALBEF	1090万	45.2	48.23	42.6
	MDETR	7.8M	46.83	48.86	44.99
	FIBER	15M	47.08	50.01	44.43
	FIBER-MM	12M	46.05	48.91	43.47
CLEVR [24]	单模态	13M	64.05	27.53	67.48
	FLAVA	9.9M	88.21	89.52	77
	ALBEF	10.9M	85.56	85.47	72.97
	MDETR	7.8M	89.35	89.4	80.2
	FIBER	15M	82.9	76.97	73.5
	FIBER-MM	12M	86.6	88.56	72.98
PACO [44]	单模态	13M	15.66	8.74	22.58
	FLAVA	9.9M	18.56	11.47	25.66
	ALBEF	10.9M	18.22	10.57	25.8
	MDETR	7.8M	19	11.13	26.87
	FIBER	15M	12.34	5.34	19.35
	FIBER-MM	12M	11.83	4.49	19.17

表5：受文献中研究者采用的多模态融合方法启发，我们尝试了不同的多模态融合方案。请注意，这些并非论文中所提出模型的数值，而是在冻结的 CLIP 特征之上采用相应多模态融合方式所得到的准确率。与在单模态注意力层上微调相近或更多数量的参数相比，大多数多模态变体表现更好。正文中的数值采用 MDETR 风格的多模态融合。

在区分粗粒度差异方面有所提升，但区分细粒度差异则更困难（根据 *Cola*平均精度均值 评估）。

F1 分数 我们还在 COLA 单对象的 GQA 划分上抽取了一部分样本进行评估，并使用 F1 分数进行评估。结果发现，比较 F1 分数时，所有趋势都保持不变。例如，CLIP 基线的 F1 分数为 0.28，而 FT-all 为 0.31，FT-Late 为 0.33，我们的 MM-Adapt 为 0.39，MM-Pred 为 0.40。我们的结论不变：调整多模态注意力层优于调整拆分模态注意力层（FT-Late）、微调整个模型或进行线性探测。

10.1 多模态融合的其他选择

在我们提出的 MM-Adapter 和 MM-Pred 方法中，我们使用了多模态融合。实现多模态融合有多种方式。其中一些重要的选择受到 FLAVA [52], ALBEF [33],MDETR [26], 以及 FIBER [11]的启发。下面介绍我们尝试多模态融合的一些方式：

– 受**FLAVA**启发的——对与图像块和文本词元拼接的 [CLS] 词元应用自注意力——这里，我们取图像块特征和文本词元特征，并在拼接后的图像、文本和 [CLS] 词元上使用自注意力 Transformer[57] 。

– 受**MDETR**启发：对图像块和文本词元进行自注意力计算，然后使用一个 [CLS] 词元交叉关注图像和文本词元-在 MDETR [26] 论文中，他们对图像和文本特征进行自注意力计算，然后使用多个任务词元交叉关注经过自注意力处理的图像-文本特征，以执行各种任务。由于这里只有一个任务，即检索，因此我们使用一个 [CLS]。我们还尝试使用多个（100 个）[CLS] 词元，以观察它们是否会学习不同的内容。我们发现，所有 [CLS] 词元学习到的内容都相同，性能差距极小。这是我们在论文中针对 MM-Adapter 和 MM-Pred 方法报告的多模态融合选择。

– 受**ALBEF**启发的文本对图像交叉注意力–这里，首先在图像块和文本词元特征上分别使用单模态自注意力层。然后，文本词元特征与图像块特征以及 [CLS] 词元进行交叉注意力计算。接着， [CLS] 输出用于 MM-Pred（使用全连接层进行预测）或 MM-Adapter（与冻结的文本特征计算余弦相似度）。

– 受 FIBER 启发的方案——文本与图像进行双向交叉注意力。首先，我们分别对图像块特征和文本词元特征应用独立的单模态自注意力层。然后，文本词元特征与一个［CLS］词元一起，对图像块特征执行交叉注意力。与此同时，图像块特征也与另一个［CLS］词元一起，对文本词元特征执行交叉注意力。最后，我们计算这两个［CLS］词元之间的余弦相似度。

– **FIBER-MM**- 在上述 FIBER 和 ALBEF 风格的融合中，我们在交叉注意力之前，分别对图像块和文本词元特征使用独立的单模态自注意力层。在这里，我们设计了一种改进方法：首先，像 FLAVA 一样，对图像块和文本词元使用多模态自注意力。然后，再按照上述描述，像 FIBER 一样执行交叉注意力。

在 $\mathcal{C}ola$上述多模态选项下单对象情形在 GQA、CLEVR 和 PACO 上的准确率如表 5所示。我们观察到的趋势与论文中报告的多模态注意力选择趋势相似。所有多模态融合方法的效果都优于单模态融合。此外，虽然有些选项优于其他选项，但请注意，对于所有设计选项，将多模态层用作特征适配器（MM-Adapter）都比将其用作预测头（MM-Pred）效果更好。

11 实现细节

现在，我们介绍主论文中使用的模型和适应策略的更多实现细节。我们还提供有关所使用数据集的更多细节。

11.1 模型架构详情

请注意，我们使用了 CLIP [43] 图像和文本编码器来提取图像和文本区域特征。以下是我们尝试的各种适配方式的更多细节：

– **线性**：我们在预训练表示上训练线性探测器。对于 CLIP [43]，我们分别在图像和文本池化特征上训练一个线性层。每个线性层都会将 512 维的图像和文本表示转换为另一个 512 维嵌入。最后，我们计算这两个转换后嵌入之间的余弦相似度。– **提示调优**：我们调整训练查询中所用文本词语的嵌入层，同时冻结其他所有部分。– **全量微调**：我们对整个模型进行微调。对于 CLIP，这涉及调整 1.51 亿个参数。– **后期微调**：我们从 CLIP 的图像和文本编码器中提取倒数第二层特征。其中有 49 个图像块特征和 K 个文本词元特征（K 取决于输入查询的长度，但上限为 77）。我们在 49 个图像块嵌入和 K 个文本词元上训练一个独立的 Transformer 编码器层。该 Transformer 编码器包含 2 个 Transformer 编码器自注意力层，每层有 4 个注意力头。我们尝试了 1 层、2 层和 3 层的变体，并报告了表现最佳的结果。我们选择这一设计，是为了使其参数数量和方法尽可能接近我们的多模态适配方法，从而构成一个强基线。– **MM-Pred**：在这里，我们将多模态注意力用作预测头，这与常见的多模态模型 [52, 26], 类似，但在冻结的 CLIP [43] 基础图像和文本编码器上进行训练。同样，该多模态 Transformer 编码器包含 2 层，每层有 4 个注意力头。我们使用全连接层对多模态注意力的［CLS］词元输出进行预测，将 512 维嵌入映射为 1 维分数。– **MM-Adapter**：这与我们的 **MM-Adapter** 方法不同；在该方法中，我们使用多模态注意力来适配图像表示，并计算其与文本特征之间的余弦相似度。

对于图文匹配损失，我们会为一个批次中的每个图文对计算一个分数。对于每个分数，我们计算二元 sigmoid 交叉熵，并取批次内的平均值。我们使用 sigmoid 交叉熵，因为对于每张图像，可能有多个正确的文本查询，反之亦然。对于基于 CLIP 的模型，我们使用 0.00001 的学习率和 0.0001 的权重衰减进行训练。对于基于 FLAVA 的适配，我们发现需要更高的学习率才能更快收敛，因此使用 0.001 的学习率和 0.0001 的权重衰减。



图 7：来自 $\mathcal{C}ola$多目标设置的一些示例。

12 定性结果

单对象案例 图 9、10、11、12 展示了基于常见适配方法和我们的 MM-Adapter 方法进行的前 5 个检索示例，针对 $\mathcal{C}ola$GQA 上的单对象设置 [18]数据集。图像中的每一行代表一种不同的适配方法（基于主论文表1中展示的方法）。请注意，我们在附着于不显著且较小物体上的多个属性方面有所改进。图 13、14、15 展示了一些案例，其中现成的 CLIP 或更简单的适配技术（如微调或线性探测）带来的改进有限。我们观察到，这些有限的改进大多出现在涉及图像中大面积区域（如天空和水域）的查询上。现有的 CLIP [43] 模型对这类显著的大型物体的表现相当不错，尤其是当它们与常见属性搭配时，例如物体“叶子”的“绿色”。

多目标案例 图16、17、18和19 展示了 $\mathcal{C}ola$ 多目标设置下的一些结果。与单对象设置中的观察结果类似，即使图像中的物体并不显著，我们也提升了属性-对象绑定能力。此外，如图20和 21所示，在关系组合性方面，我们的方法仍无法理解属性相对强度的细微差别，以及物体被严重遮挡的情况。

我们使用的所有图像均来自公开数据集，且据我们所知，这些图像与现实生活中可识别的人物没有任何对应关系。

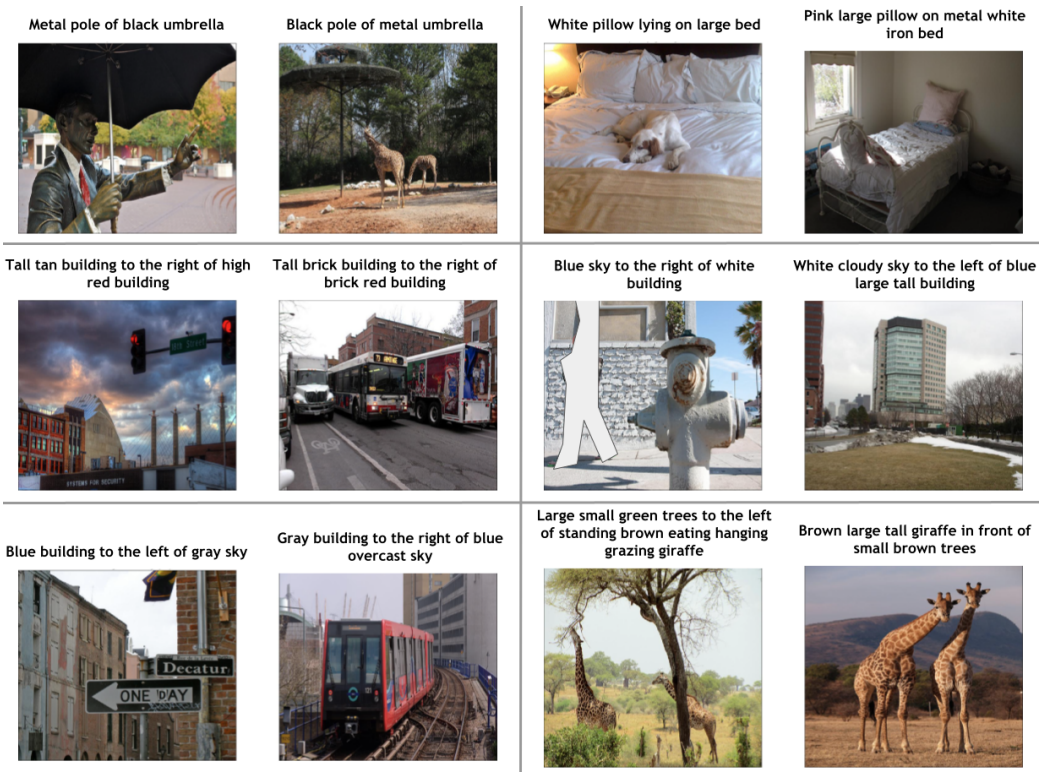


图 8：$Cola$多物体设置中的一些示例，即使经过人工清理仍存在一定歧义。请注意，这些歧义大多与颜色（在不同光照下看起来可能不同）和尺寸（具有主观性）有关。

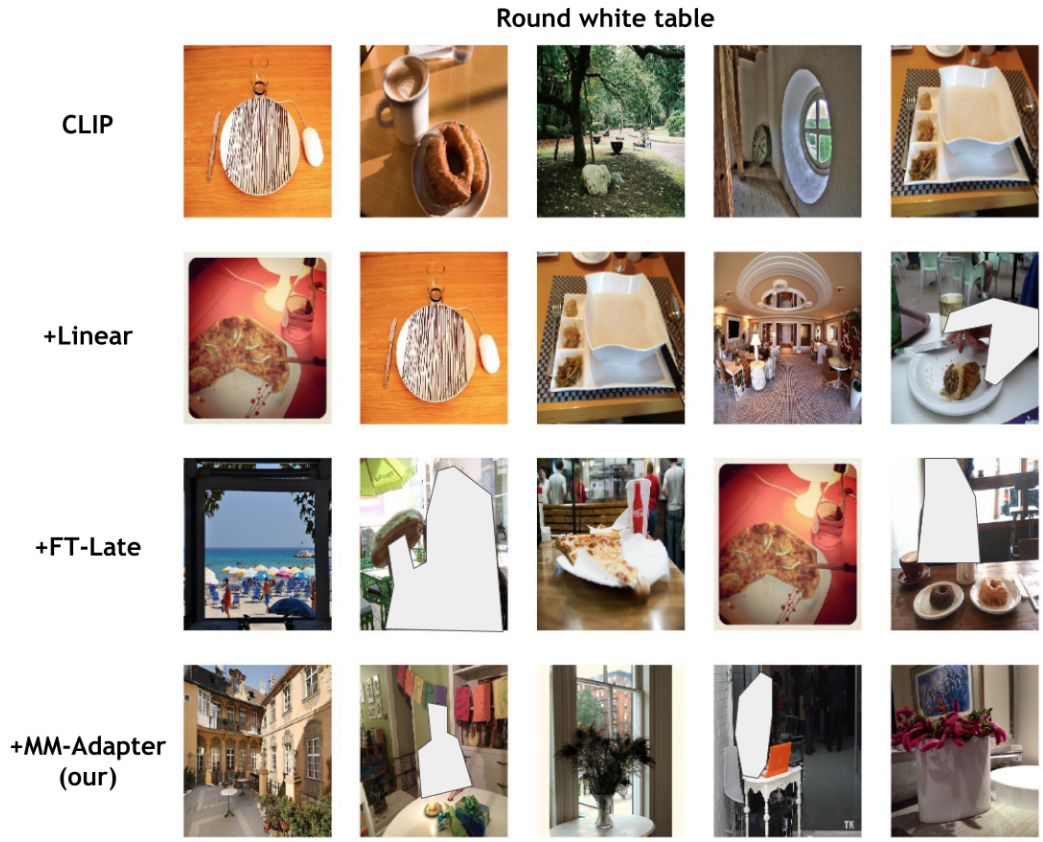


图 9：附着于物体的多个属性的定性结果。请注意，我们在杂乱场景中附着于小型非显著物体的多个属性方面取得了显著改进。测试图像中的圆形白色桌子通常尺寸较小，因此原始模型难以找到它们。请注意，原始 CLIP 只能找到显著的黑色金属椅子（第一行），相比之下，我们还能找到较小的非显著物体（最后一行）。



图 10：物体具有多个属性时的定性结果。请注意，我们在杂乱场景中对附着于小型非显著物体的多个属性进行了改进。

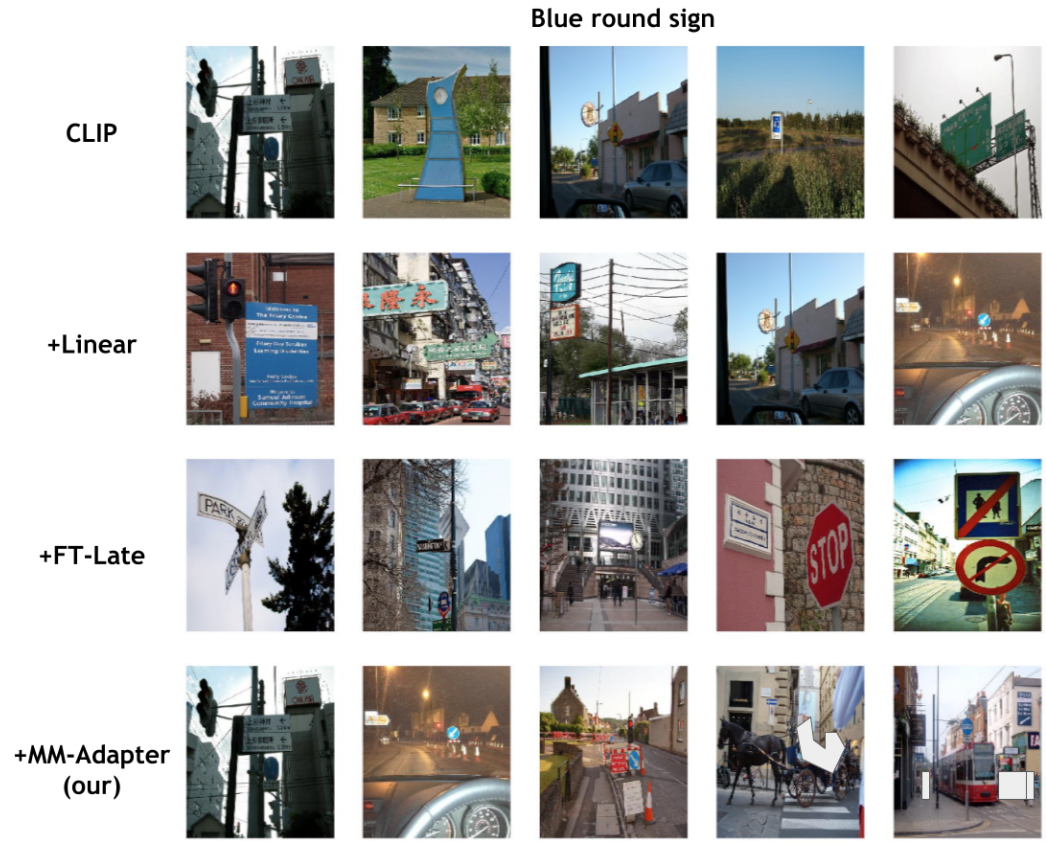


图11：附加于物体的多个属性的定性结果。请注意，我们在杂乱场景中附加于小型非显著物体的许多属性上取得了改进。

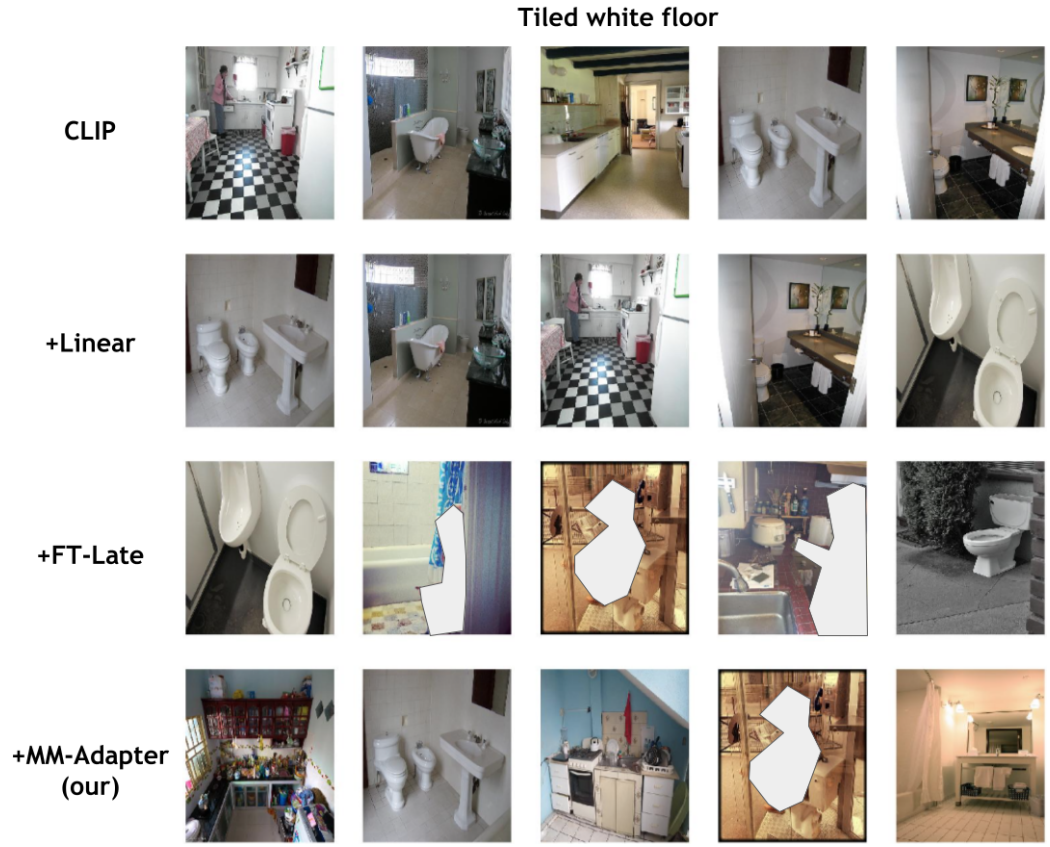


图 12：物体所带多个属性的定性结果。请注意，我们在杂乱场景中的小型非显著物体所带的许多属性上都有所提升。

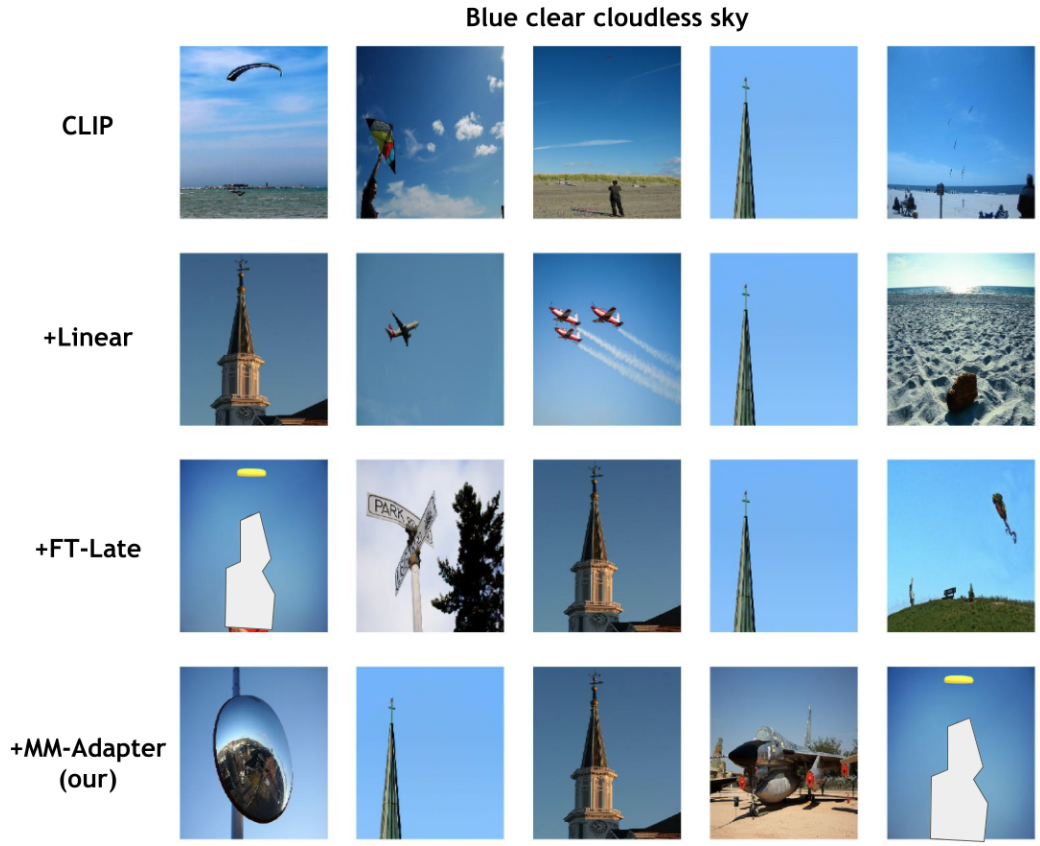


图13：具有属性的查询若覆盖范围广且包含蓝天等常见属性，那么采用现成模型或简单的适应策略所带来的提升微乎其微，因为现有模型对这类查询的表现已经很好。

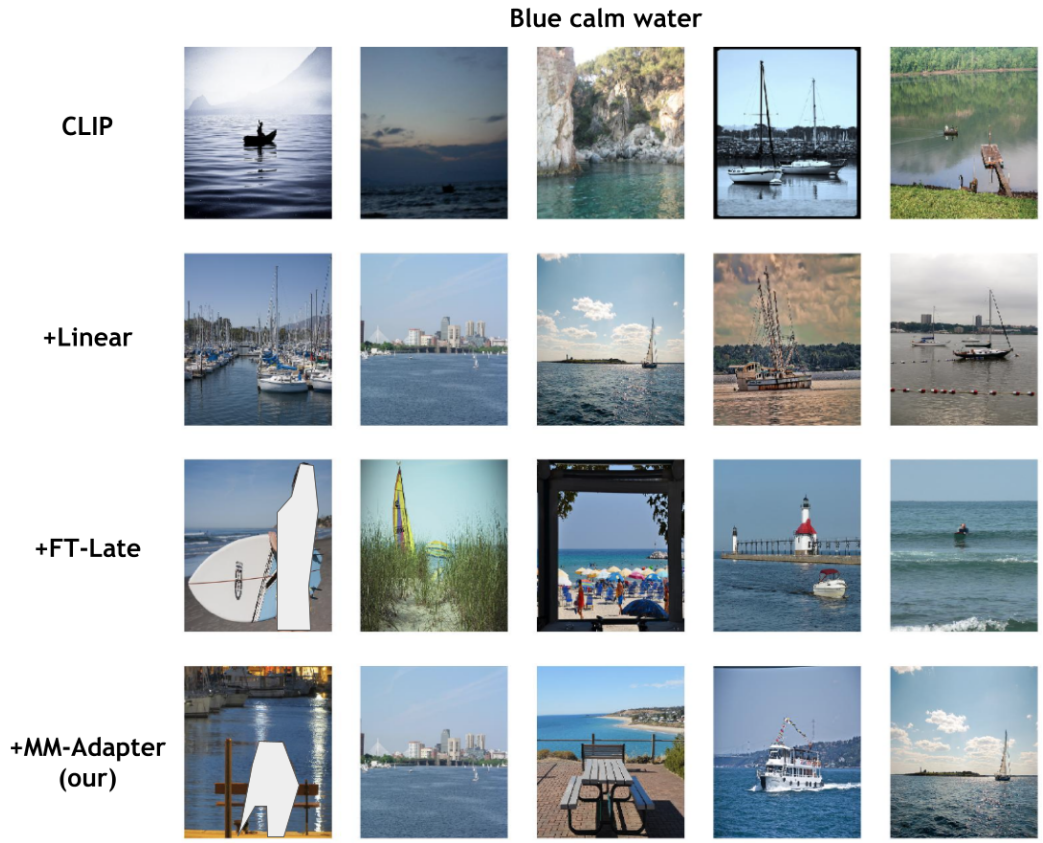


图14：覆盖范围较广的带属性查询（例如水体）从现成模型或简单适应策略中获得的提升很小，因为现有模型在这类查询上已经表现良好。

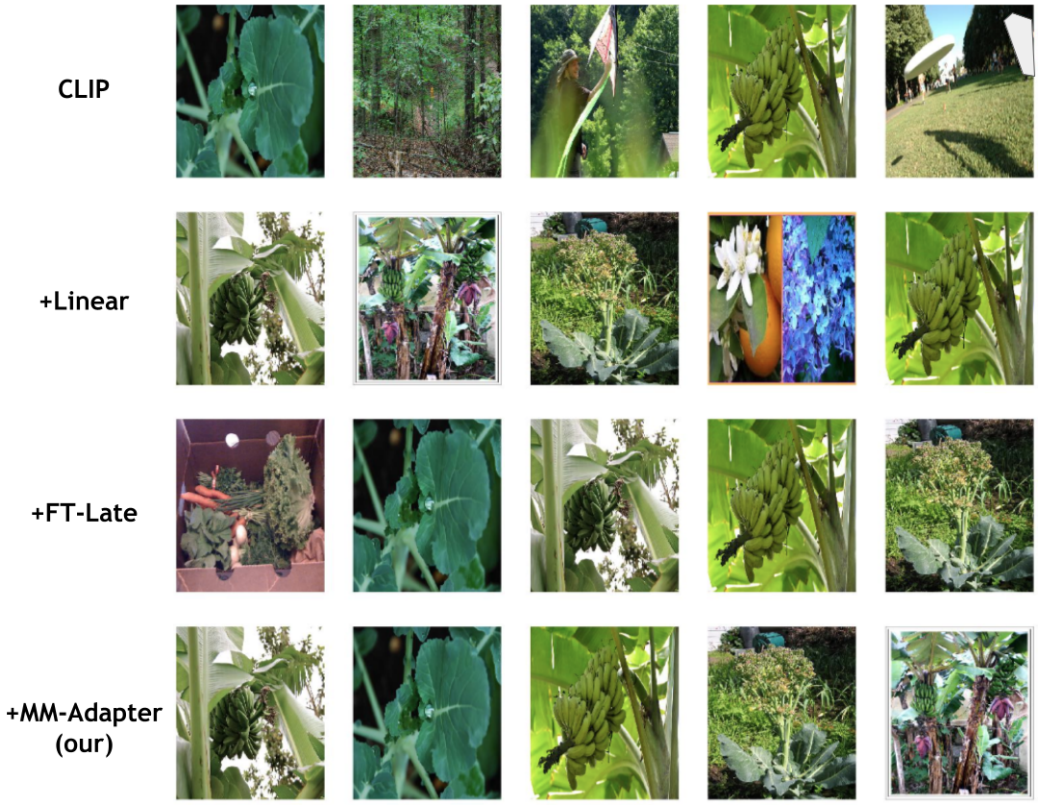


图15：具有覆盖较大区域且常见属性的查询（例如大型绿色叶片）从现成模型或简单适应策略中获得的改进最小，因为现有模型在处理此类查询时已经表现良好。



图16：多目标案例的定性结果。我们再次看到，涉及小型非显著物体（例如小标志）的构图取得了显著改进。

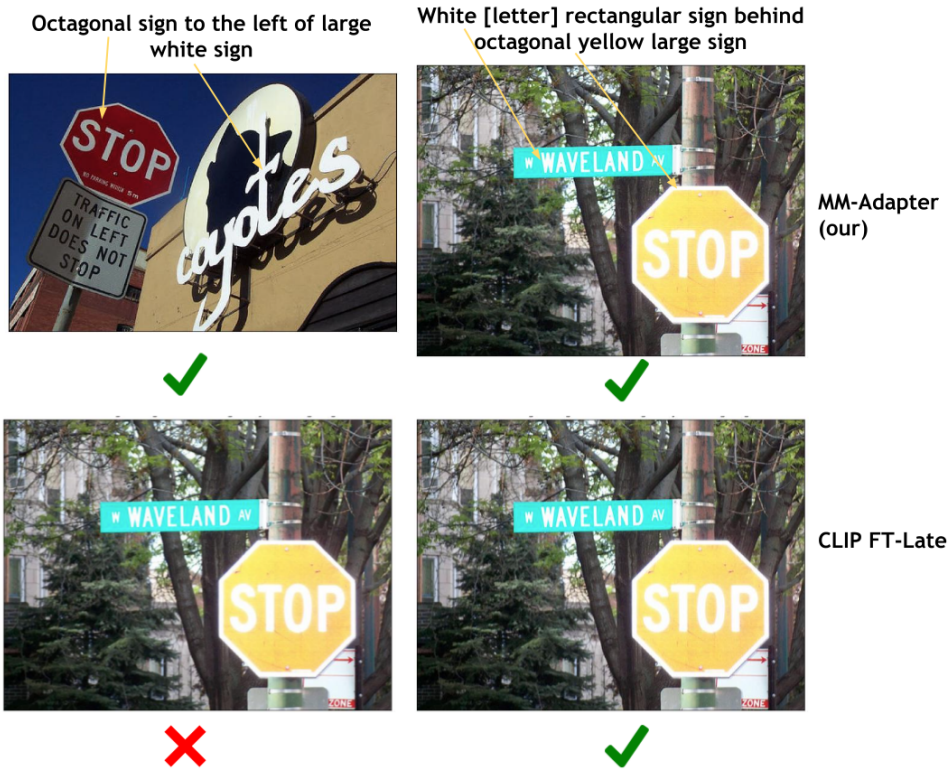


图 17：多目标案例的定性结果。

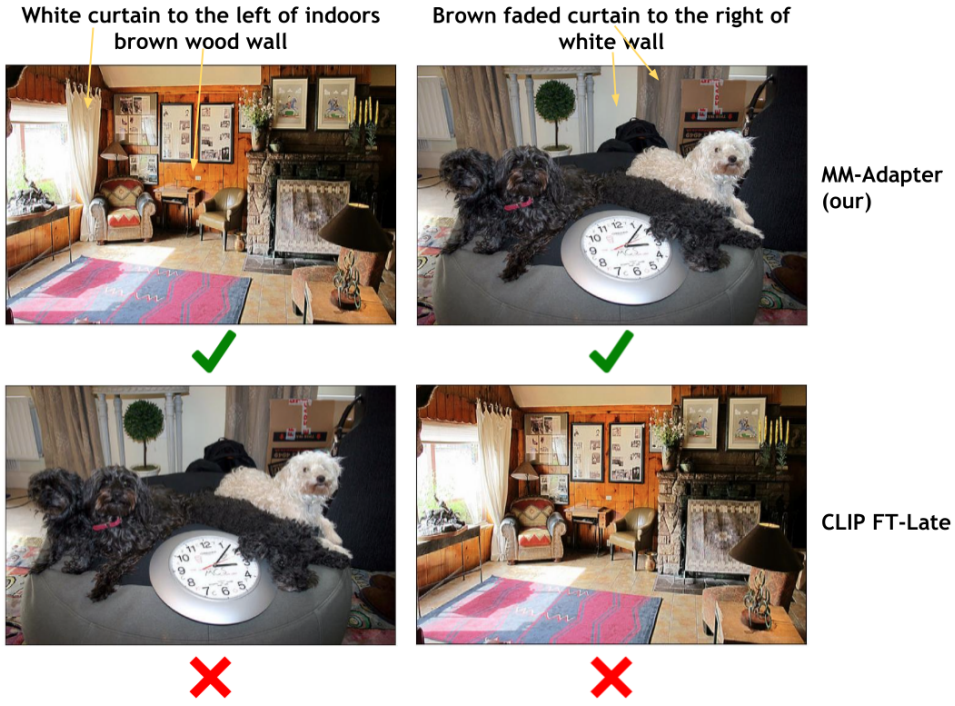


图 18：多目标案例的定性结果。我们发现，在图像中杂乱元素和干扰物体较多的构图上，效果有了显著提升——场景中有许多白色和棕色的物体。

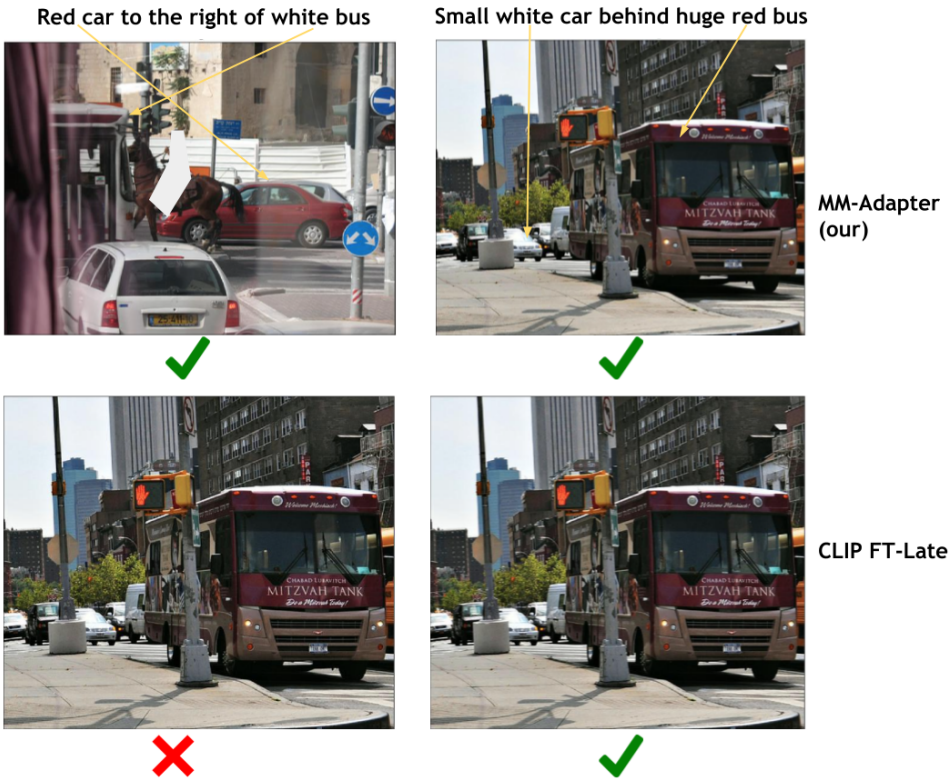


图19：多目标案例的定性结果。

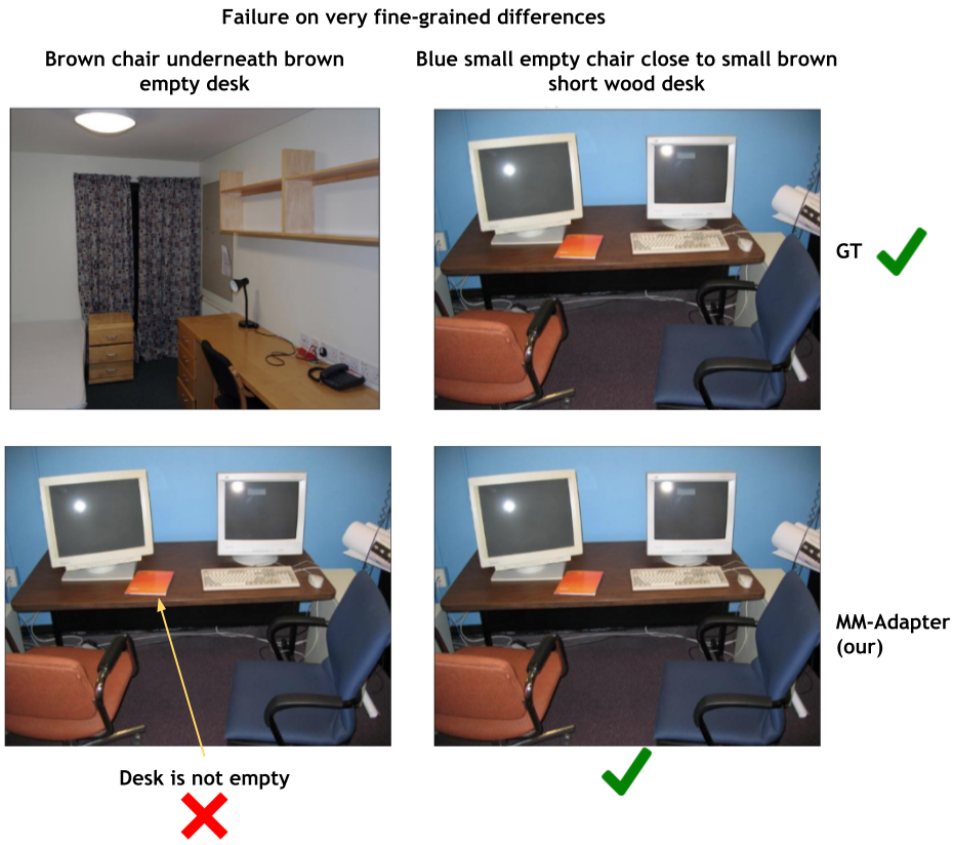


图 20：我们的方法在处理非常细微的相对差异时表现略差。在上面的示例中，一把棕色椅子位于一张棕色书桌下方，但书桌并非空无一物。事实上，与该图像描述对应的正确图像中的书桌严格来说也不是空的，但它比干扰项中的书桌更空，而我们的模型未能理解这种相对差异。

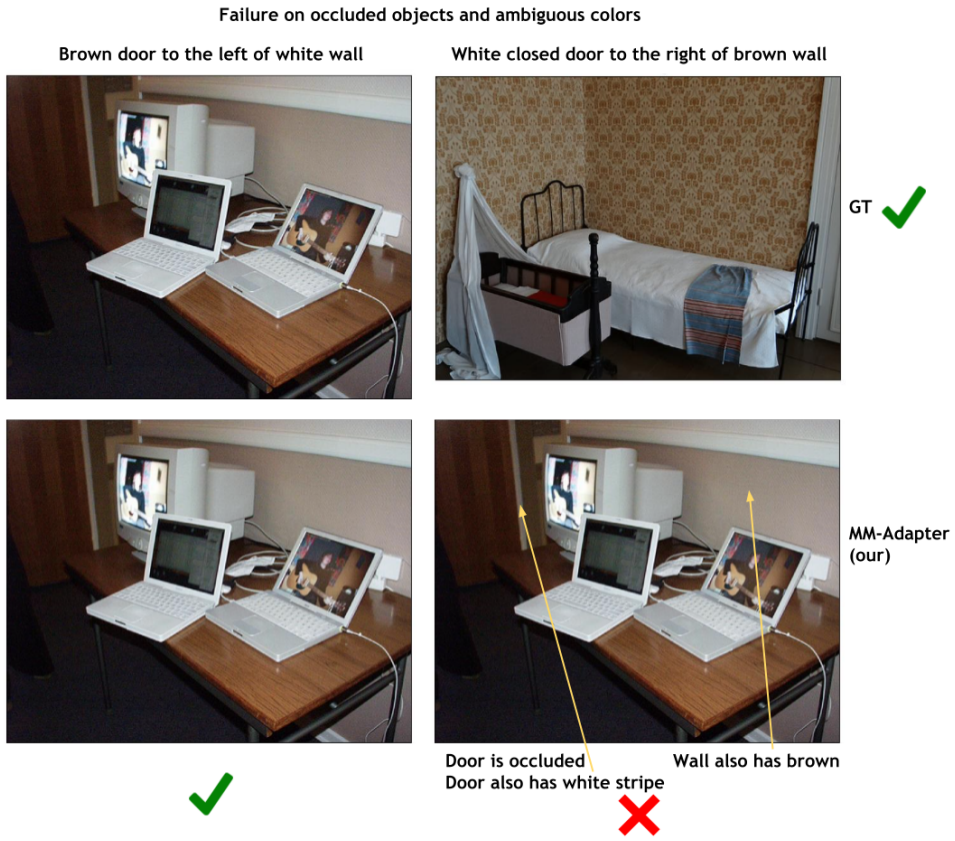


图 21：当物体被遮挡，或物体也具有干扰项的某些属性时，我们的方法表现也不佳。在上面的示例中，门没有清晰地出现在视野中。此外，棕色的门上也有一条白色条纹，这进一步混淆了模型。