**Evaluating Object Hallucination in Large Vision-Language Models**

大型视觉语言模型中的物体幻觉评估

**摘要**

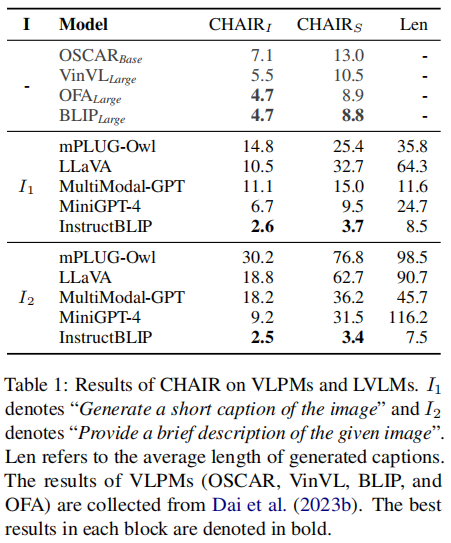
受大型语言模型（LLM）优越的语言能力的启发，大型视觉语言模型（LVLM）最近被提出，通过集成强大的LLM以提高复杂多模态任务的性能。尽管LVLM在技术上取得了很好的进展，本文发现它们仍然存在物体幻觉，也就是说，它们倾向于产生与目标图像不一致的内容。故本文对物体幻觉进行了系统分析。 本文对几个经典LVLM上进行了evaluation实验，结果表明它们都存在严重的物体幻觉问题。然后本文通过进一步探索prompt（visual instruction)和幻觉的关联得出结论：经常出现在prompt中的物体，以及，常与图像物体共同出现的物体 更容易被LVLM错误产出。分析之外，本文设计了一种基于轮询的查询方法POPE，以更好地评估物体幻觉。实验结果表明，本文的POPE能够更稳定、更灵活地评估物体幻觉。

**引言**

大型视觉语言模型（LVLM），即通过用LLM替换原来的textual encoder来增强整体模型性能。经过vision-language 预训练和visual 指令微调后，LVLM在各种视觉语言任务(VQA, Image Captioning)上表现十分优异，展现了LVLM极佳的研究前景。

尽管LVLM取得了一定成功，但大量研究都指出了LVLM中的LLM存在幻觉问题。LLM倾向于产生意外文本的幻觉。而传统的VLM也会产生物体幻觉。在实际应用中，幻觉会降低模型性能，极大地损害用户体验。因此，很自然会提出这样一个问题：LVLM中仍然存在幻觉吗？故本文系统地评估了现有的LVLM中的物体幻觉问题。

本文首先采用CHAIR（一种基于图像相关性的幻觉评估方式）评估了几个经典LVLM的幻觉程度（基于MSCOCO数据集），结果如表1所示。实验表明大多数LVLM物体幻觉问题严重，它们甚至比原来的小型VLM更容易产生幻觉。



此外，本文发现目前的物体幻觉评估指标可能不适用于LVLM，故本文进一步提出了Polling-based Object Probing Evaluation（POPE，基于轮询的物体检测评估）。其基本思想是将对幻觉的评估转换为二元分类任务，通过用简单的是或否的简短问题询问探测物体（例如，图像中是否有汽车？）。实验结果表明，这种方法更加稳定和灵活。此外，通过使用不同的物体采样策略，本文验证了现有的LVLM容易产生幻觉的物体类型：在视觉指令数据集中频繁出现或同时出现的物体。

本文的主要贡献如下：

(1)对几个具有代表性的LVLM进行了对物体幻觉的实证研究，发现它们受到物体幻觉的高度影响。

(2)讨论了这一问题背后的潜在原因，例如，LVLM倾向于生成在指令语料库中频繁出现或同时出现的物体。

(3)提出了一种被称为POPE的物体幻觉评估方法，该方法更稳定，可以更容易地扩展到无标注的数据集。

**背景**

**大型视觉语言模型**

一般来说，LVLM由一个视觉编码器、一个语言编码器（即LLM）和一个跨模态对齐网络组成。LVLM的训练一般由三个主要步骤组成。

• 首先，使用大规模单模态数据（分别为图像和文本数据）对视觉编码器和语言编码器进行预训练。

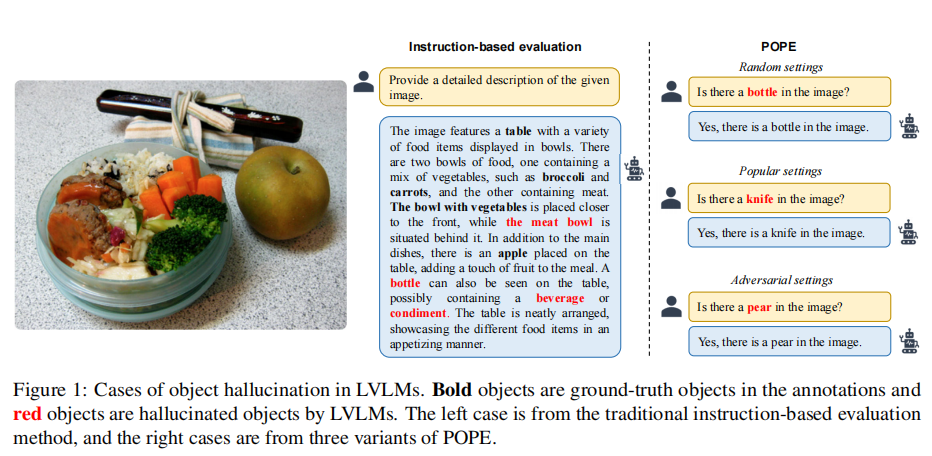
• 其次，这两个编码器通过图像-文本对齐预训练进行对齐，这使LLM能够为给定的图像生成一个有意义的说明文字。

• 最后，使用图像-文本指令对对齐模型进行微调，从而生成关于特定图像的自然语言问题的满意答案。

在第二步和第三步中，可以有选择地微调不同的组件，而不是执行全参数的微调。一旦视觉编码器和LLM很好地对齐，导出的LVLM就可以显示出优越的视觉理解能力。

**物体幻觉**

尽管LVLM在解决语言任务方面很强大，但也像VLM一样存在物体幻觉的问题。物体幻觉是指模型生成的description和image不一致(图片中的红球被说成了绿球，或者图片中压根没有球，模型却说有红球)。一般来说，物体幻觉可以在不同的语义层次上进行定义。最直接的方法是在物体级别上定义，而更细粒度的定义则与物体的属性或特征有关。本文重点关注description中的粗粒度的物体幻觉，而细粒度的物体幻觉(如物体的数量、属性和位置)，暂先留至未来研究。在图1中展示了一个产生物体幻觉的例子，其中LLM回复中的标红的“装肉的碗”、“瓶子”、“饮料”、“调味品”都是由底层的LLM错误产生的。

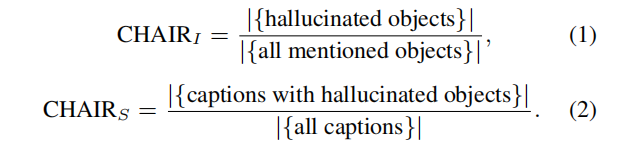


**LVLM中的物体幻觉**

在本节中，研究人员使用CHAIR指标来评估流行的LVLM中的物体幻觉问题。

**评估设置**

CHAIR（Caption Hallucination Assessment with Image Relevance）是一种在Image Captioning任务中评估物体幻觉的常用指标。给定图像中的标注物体，CHAIR计算「出现在模型输出description但不真实出现在image」物体的比例。现有的工作通常采用它的两种变体，即CHAIRI和CHAIRS，它们分别在物体实例水平和句子水平上评估幻觉程度。它们可以表述为：



研究人员选择了五个最近发布的LVLM，用以下prompt让LVLM对MSCOCO中每个image生成description：

· I1：Generate a short caption of the image.

· I2：Provide a brief desciption of the given image.

然后根据生成的description计算CHAIR。

**实验结果**

**幻觉的严重程度**

表1所示的结果表明，大多数指令调优的LVLM都存在物体幻觉问题，甚至比小模型更严重。作为比较，InstructBLIP比其他LVLM的幻觉更少。一个可能的原因是，它的视觉指令是从各种公开的数据集收集而来的，这些数据集相对较短。相比之下，其他的LVLM大多使用由单模态LLM生成的视觉指令。这种生成的视觉指令虽然时间更长，信息更丰富，但可能涉及与图像不一致的意外描述性信息（LLM固有的幻觉），这可能会误导LVLM。

**CHAIR的缺点**

如表1所示，评价结果也可能受到指令的设计和描述的长度等其他因素影响。具体来说，虽然采用的两个指令具有相似的语义意义，但指令2提示的LVLM可以比指令1提示的CHAIR指标值增加一倍，并且一些LVLMs的性能顺序也会发生变化。表明了当使用不同的指令时，CHAIR指标的不稳定性。此外，由于CHAIR需要检查生成的描述中的物体是否产生幻觉，因此需要复杂的人工解析规则来进行精确匹配，这并没有适应LVLM的特殊生成风格，可能导致误分类错误。

**指令数据对物体幻觉的影响**

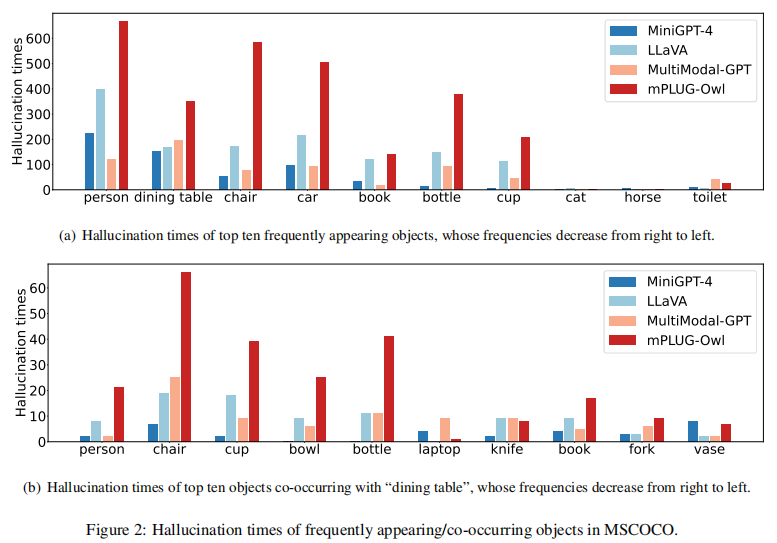
**假设**

（1）LVLM容易对视觉指令数据集中频繁出现的物体产生幻觉。

（2）LVLM容易对经常和图像中的标注物体同时出现的物体产生幻觉。

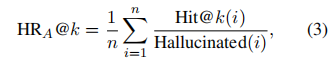
**定性分析**

对于第一个假设，在MSCOCO中经常出现的前10个物体和出现幻觉的次数之间绘制了一个条形图；对于第二个假设，选择了前10个经常与“餐桌”同时出现的物体，并绘制了一个条形图，以显示它们在真正包含“餐桌”的图像中出现幻觉的次数。在图2中显示了MiniGPT-4、LLaVA、MultiModal-GPT和mPLUG-Owl的结果。显然，随着物体出现频率的减少，所有四种LVLM的幻觉次数均显著减少。说明在视觉指令数据集中经常出现和同时出现的物体确实更有可能被LVLM产生幻觉。



**定量分析**

为了进一步巩固上述发现，我们采用top-k命中率（HR@k）来测量物体出现频率和出现幻觉次数之间的一致性，定义为：





其中，HRA和HRC分别表示幻觉次数与出现频率和共同出现频率之间的相关性。n是图像的总数，Hallucinated(i)表示第i个例子中幻觉物体的数量，Hit@k(i)表示在Hallucinated(i)中且在MSCOCO前k个频繁出现的物体的数量，而Hit@k（i，o）表示在Hallucinated(i)中的前k个频繁与探测物体o同时出现物体的数量。因此，HR@k可以反映所有幻觉物体中经常出现或同时出现的顶部物体的比例。

[本文在表2中展示了前30个物体的HRA和HRC（餐桌），并在附录C中留下了HRC@（椅子）和HRC@（汽车）。所有LVLM的HRA@10和HRC@10（餐桌）分别接近0.5和0.6。这表明，平均而言每张图像中大约有一半的幻觉物体属于前10个经常出现的COCO物体，而超过一半是经常与图像中已经出现的物体共存的前10个物体。当我们将观察范围扩大到前30个物体时，这一比例就会继续增加。这些发现进一步证实了LVLM对视觉指导数据中常见物体的幻觉。](mailto:本文在表2中展示了前30个物体的HRA和HRC（餐桌），并在附录C中留下了HRC@（椅子）和HRC@（汽车）。所有LVLM的HRA@10和HRC@10（餐桌）分别接近0.5和0.6。这表明，平均而言每张图像中大约有一半的幻觉物体属于前10个经常出现的COCO物体，而超过一半是经常与图像中已经出现的物体共存的前10个物体。当我们将观察范围扩大到前30个物体时，这一比例就会继续增加。这些发现进一步证实了LVLM对视觉指导数据中常见对象的幻觉。)

**POPE**

**POPE概述**

POPE将对物体幻觉的评估定义为一个二元分类任务，提示LVLM输出“是”或“否”，例如，“图像中有椅子吗？”。这样，通过对LVLM容易产生幻觉的物体进行采样，我们就可以构建一组困难的问题来投票调查LVLM。由于这些问题的标准答案只是“是”或“否”，我们可以不需要复杂的解析规则就很容易地识别它们，避免了指令设计和标题长度的影响，从而保证了稳定性、公平性和灵活性。

**定义**

给定一个图像描述数据集，POPE专注于构建一组三元组，每个三元组包含一个图像、多个问题和它们的答案（“是”或“否”）。对三元组的明确定义可以描述为：



其中x表示图像，q（oi）是基于模板“图像中是否存在<物体>？”的问题，oi是要询问的第i个物体，ai是问题的答案（“是”或“否”），l表示每张图像的轮询问题的数量。oi可以从注释或类似SEEM等自动分割工具的结果中获得。对于标签平衡，设置标注和不存在的物体之间的比率为1：1。在构建了评估三元组之后，可以直接用它们来投票给LVLM，并收集预测的答案。

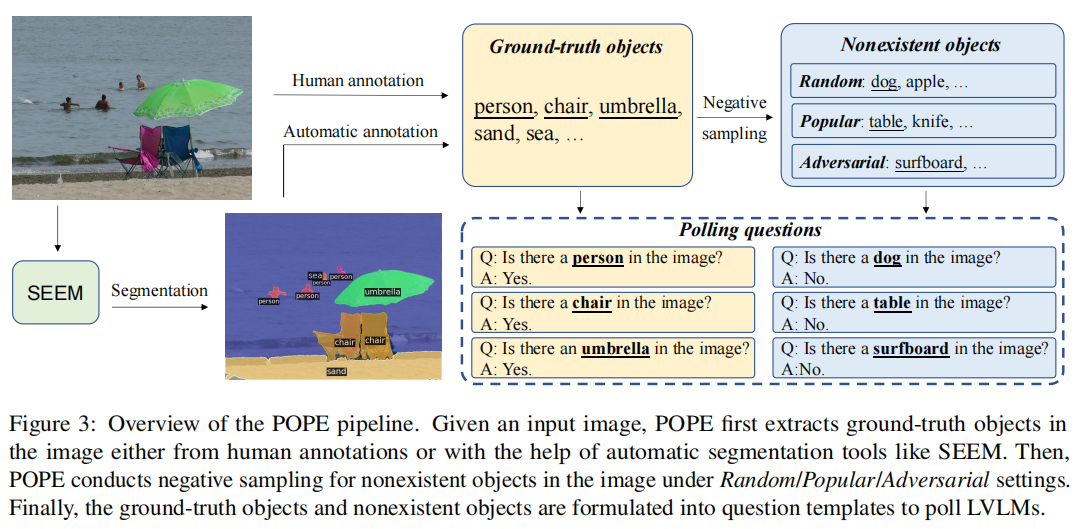
**综合解决方案**

整个POPE如图3所示。在获取了图像中的物体后，就可以开始构建轮询问题了。答案是“是”的问题可以直接使用标注的物体来构建，而答案是“否”的问题可以通过从负面物体中采样来构建。因此，通过设计不同的采样策略，我们可以验证LVLM是否容易产生特定物体，如第4节中讨论的经常出现或同时出现的物体。因此，设计了以下三种采样策略：

随机抽样：随机抽样图像中不存在的物体。

流行抽样：选择整个数据集中不存在的前top-k最常见的物体，其中k=l/2

对抗采样：首先根据所有物体与标注物体的共出现频率对其进行排序，然后选择图像中不存在的前k个频繁物体。



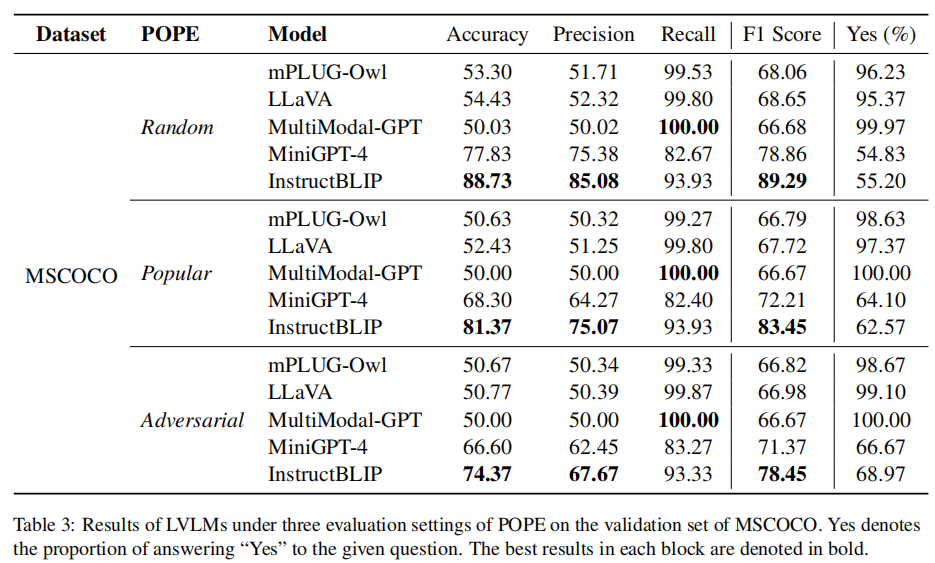
**指标**

采用准确性、精密度、召回率和F1评分作为评价指标。准确性反映了正确回答的问题的比例。精确度和召回率分别反映了正确回答答案为“是”或“否”的问题的比率。F1分数结合了精度和召回率的结果，选择它作为评估的主要指标。此外，还报告了LVLM回答“是”的比率，作为分析模型行为的参考。

**在MSCOCO上的评估**

使用POPE在MSCOCO验证集上评估LVLM。在注释中随机选择500张具有超过3个标注物体的图像，并为每个图像构造6个问题（即l = 6）。

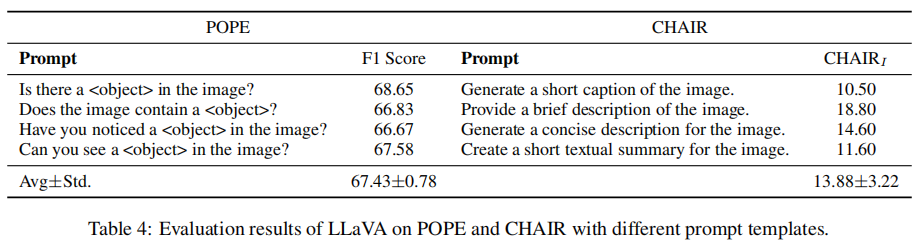
结果如表3所示，可以得到与表1相似的结论，即InstructBLIP表现最好，而LLaVA、MultiModal-GPT和mPLUG-Owl的幻觉问题更严重，其F1评分低于70。这表明POPE可以很好地估计LVLM中幻觉问题的程度。此外，发现LLaVA、MultiModal-GPT和mPLUG-Owl非常容易回答“是”（接近99%）。表明这三个LVLM过于自信，导致对答案为“否”的问题的准确性较低。此外，从随机设置到流行设置和对抗性设置，LVLM的性能持续下降。这与在第4节中的发现一致，因为LVLM容易对经常出现和同时出现的物体产生幻觉。



**POPE的优点**

**稳定性**

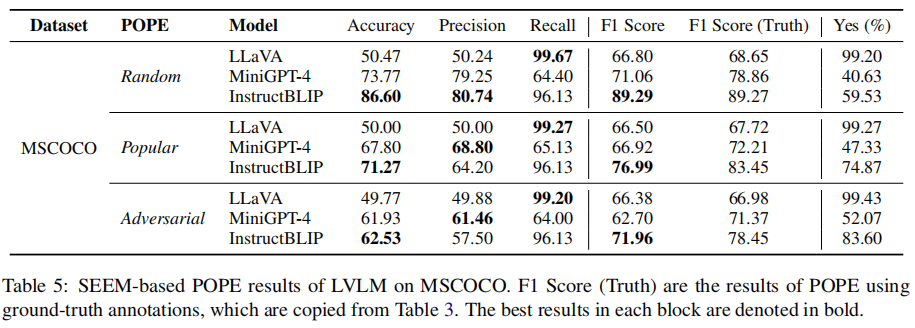
不管提示模板的变化如何，POPE要求LVLM回答简单的封闭式问题，与基于指令的方法相比，这不太可能引入歧义，有助于稳定性。为了验证，使用POPE和CHAIRI评估LLaVA，每个都有四种不同的提示。评价结果见表4。可以观察到，F1评分标准偏差显著低于CHAIRI标准偏差，这证实了POPE在面对不同提示时表现出更高的稳定性。

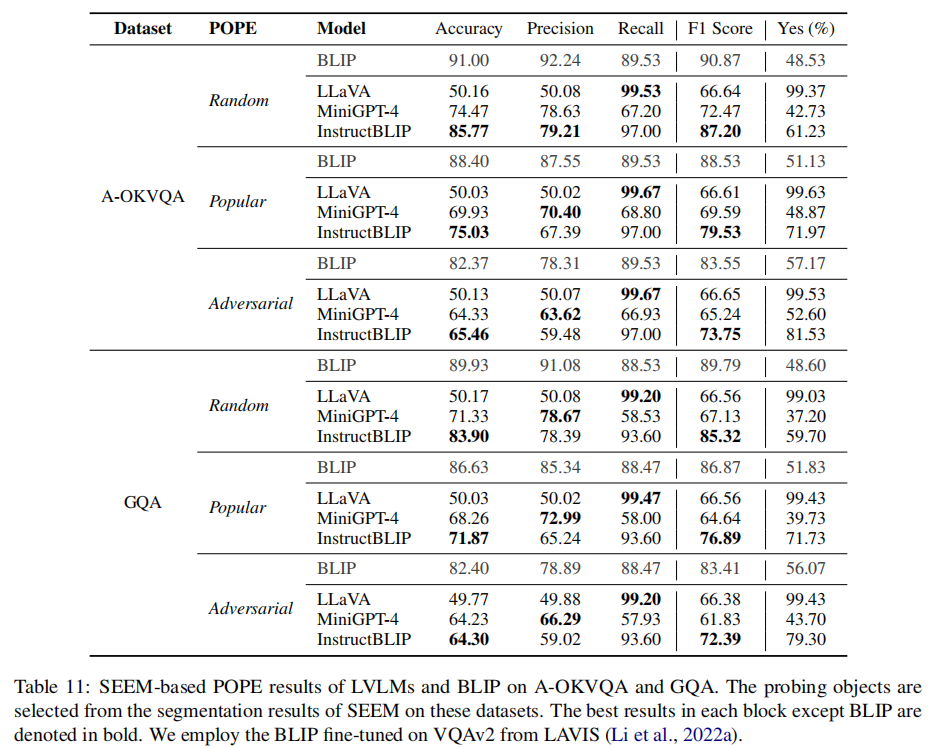


**可扩展性**

在自动分割工具的帮助下，POPE可以很容易地扩展到没有注释的数据集。为了验证这一点，采用SEEM对三个数据集（MSCOCO、A-OKVQA和GQA）的图像进行注释，并基于分割结果构建POPE。对InstructBLIP、MiniGPT-4和LLaVA进行了评估，并将结果报告在表5和表11中。在表5中，所有LVLM的性能大多与表3中基于注释的POPE相同，即随机>流行>对抗和InstructBLIP>MiniGPT-4>LLaVA。这种一致性表明了基于SEEM的POPE的可靠性。

然而，这两种设置之间存在性能差距。这种现象可以归因于由SEEM生成的分割结果的粒度更细，这使得POPE更具挑战性。总之，当与自动分割工具结合时，POPE可以很容易地扩展到未注释的数据集，并对它们进行有效的评估。





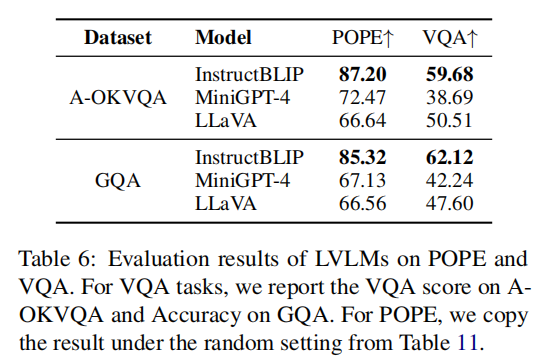
**一致性**

POPE的一个潜在问题是，LVLM的“是/否回应”是否真正反映了对物体的感知。为了验证这一点，测量了POPE响应和由LVLM生成的描述之间的一致性。具体来说，检查收到“否”响应的物体是否很少出现在描述中，以及描述中经常提到的物体是否通常会收到“是”的答案。从InstructBLIP和MiniGPT-4收集数据，考虑到它们相对平衡的是/否分布。研究结果显示，在InstructBLIP和MiniGPT-4给出的1303和1445个“否”回答的物体中，只有0个和5个物体在描述中出现。此外，在这些模型的描述中提到的664个和1034个物体中，分别有664个和961个得到了“是”的判决。这些结果强调了物体在描述中的存在与在POPE问题中是/否的回答之间的稳健相关性，验证了POPE评估的可靠性。

**幻觉对视觉任务的影响**

虽然现有的LVLM确实存在严重的物体幻觉问题，但这些幻觉是否对其他视觉任务有强烈影响仍是一个悬而未决的问题。因此，研究人员比较了它们在VQA和图像字幕任务上POPE表现。对于VQA任务，评估了在A-OKVQA和GQA数据集上的基于SEEM的POPE和VQA得分。由于LVLM倾向于以一种开放的方式生成答案，因此利用ChatGPT来帮助解析生成的结果，以更好地评估VQA的性能。对于图像字幕任务，使用传统指标评估了POPE中的500张图像的描述。

实验结果见表6。InstructBLIP在所有设置下表现最好，突出了大型视觉指令语料库上指令调优的重要性。注意，由于InstructBLIP已经接受过A-OKVQA培训，因此应该谨慎考虑结果。此外，尽管MiniGPT-4获得了更高的F1分数，但它在VQA任务上的表现相对较差。一个可能的原因是，MiniGPT-4的指令数据集仅来自图像标题数据，而LLaVA使用了涉及复杂视觉问题的158K视觉指令数据。结果表明，幻觉的程度可能并不总是与VQA的表现相一致，这两个评价方面都很重要，应该在现实应用中加以考虑。



**结论**

在这项工作中，研究人员对几个LVLM进行了评估实验，并检查了它们的物体幻觉的问题。通过调查物体幻觉的原因，揭示了视觉指令的物体分布会影响LVLM的物体幻觉。此外，还发现现有的幻觉评估方法可能会受到输入指令和LVLM生成的文本的影响，从而导致评估结果较不可靠。为了解决这个问题，提出了一种基于轮询的查询方法，称为POPE，为LVLM的物体幻觉提供了一种改进的评估方法。实验结果表明，本文提出的POPE可以更好地评价LVLM的物体幻觉问题。

**局限性**

首先，本文只关注LVLM中的物体幻觉问题，而没有考虑其他可以反映LVLM能力的方面。这意味着当前的评估任务不能衡量LVLM的整体性能。如果某个模型在本文的评估设置中获得了更高的分数，它并不一定表明比分数较低的模型有更强的整体能力。

其次，由于计算资源的限制，只能在每个数据集的部分验证集上评估所有模型。报告的结果可能会受到相应的数据分布的影响，尽管已经仔细地设置了实验。

第三，本文提出的POPE采用基于匹配的方法来确定LVLM是回答“是”还是“否”，而在经验上，LVLM偶尔可能无法提供明确包含这些词的答案，这可能导致不准确的评估结果。

第四，当与自动分割工具结合时，会根据工具设置的标签对物体进行标注，这可能与收集到的人工标注不一致，导致评价结果存在差异。

最后，这项工作只比较了少量的LVLM，而不包括一些最近发布的或闭源代码的LVLM。