【BBVPE: 通过黑盒视觉提示工程缓解大视觉语言模型中的物体幻觉问题】

——【Black-Box Visual Prompt Engineering for Mitigating Object Hallucination in Large Vision Language Models】

1 相关资源

pdf: https://aclanthology.org/2025.naacl-short.45.pdf
ppt:
短视频:
数据集:
源码:
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: NACCL 2025
【给出具体会议名称和年限，不要仅仅写 ACM，IEEE】
论文类别: Large Language Model
【论文的类别，比如移动计算、轨迹处理、深度学习等】
论文关键字: LLMs, Hallucination Mitigation
推荐程度：3 （其他说明可标注）
（5 非常棒，建议认真研读、小组讨论和复现；4 好，建议细读，考虑复现；3 可以，部分内容值得注意；2 一般，简单浏览即可；1 没有意义，不建议阅读）

3 工作团队

作者: Sangmin Woo, Kang Zhou, Yun Zhou, Shuai Wang, Sheng Guan, Haibo Ding, Lin Lee Cheong
单位:

1. Amazon AWS AI
2. KAIST

团队情况描述:

4 论文介绍

(1) 研究目的
【研究背景是什么？本文工作有什么用?】

大型视觉-语言模型（LVLM）已能在图像描述、视觉问答等任务中生成流畅且上下文相关的文本，但在医疗、自动驾驶、机器人等高可靠性场景中，模型经常“脑补”出图像中并不存在的物体（即对象幻觉），导致错误决策。
本文提出了一个名为黑箱视觉提示工程（BBVPE）的框架，旨在通过视觉提示来减少对象幻觉，而无需访问模型的内部结构。

(2) 研究现状
【当前的最好研究做到什么程度了？存在的问题是什么？这里采信论文的说法，可以给出自己的评点】
当前的研究在减少 LVLMs 的对象幻觉方面已经取得了一定进展，主要集中在数据改进、训练方法优化和解码策略调整等方面。例如，通过收集更高质量的数据集、利用外部数据集进行监督学习、采用强化学习或偏好优化等方法来调整模型输出，以及通过改进解码策略来减少幻觉。然而，这些方法大多需要访问模型的内部结构，如注意力权重、预测概率等，这使得它们在处理专有模型（如 OpenAI 的 GPT 系列）时受到限制。此外，现有方法大多假设模型是白盒的，即可以访问模型的内部信息，这在实际应用中并不总是可行的。本文指出，尽管这些方法在一定程度上减少了对象幻觉，但它们的适用范围有限，且无法直接应用于黑箱模型

(3) 本文解决的问题
【一句话概括本文解决的核心问题】
本文的核心问题是提出一种无需访问模型内部信息的方法，以系统地识别和应用最优的视觉提示（VPs），从而减少 LVLMs 中的对象幻觉。

(4) 创新与优势
【本文的创新之处是什么？新场景？新发现？新视角？新方法？请明确指出】
【本文工作的贡献或优点是什么?】

1. 发现 LVLM 存在针对图像的 Oracle VP，可大幅减少物体幻觉
2. 提出 BBVPE 框架系统化识别最优 VP
3. 在 POPE 和 CHAIR 基准测试中，本方法显著降低开源和商业 LVLM 的物体幻觉率

(5) 解决思路
【本文是怎样解决问题的？包括方法、技术、模型等，以自己理解的方式表述清楚】
作者提出 BBVPE 框架，系统化识别并应用最优 VP 来减少 LVLM 物体幻觉。该方法将 LVLM 视为” 黑盒”，仅依赖输入-输出对而不修改模型本身。框架包含三个核心组件：(1) 预定义 VP 池，(2) 评估提示效果的评分函数，(3) 根据输入-输出行为动态选择最佳提示的路由模型。本方法无需访问模型内部，适用于开源和商业 LVLM。
为识别图像 I 中的相关物体，首先使用物体定位模型 $\mathcal{L}$。该模型检测并输出物体坐标集 $O = \{o_1, o_2, \ldots, o_m\}$。
作者定义了一个候选视觉提示 (VP) 池 $P = \{p_1, p_2, \ldots, p_n\}$，包含圆形、箭头等视觉标记。每个 $VP_{p_i} \in P$ 通过高亮定位对象 O 来修改图像 I，生成 I_{p_i}。将图像-文本对 (I_{p_i}, T)（其中 T 为文本提示）输入 LVLM$\mathcal{M}$ 以生成响应。
量化对象幻觉，为评估模型抗对象幻觉的鲁棒性，我们定义评分函数 S 来测量关于对象存在的响应准确度：

$$S = \frac{|\text{correct responses}|}{|\text{total presence questions}|} \tag{1}$$

数据集构建，对于给定图像 I，选择最优 VPp^* 以最大化 S：

$$p^* = \arg\max_{p_i \in P} S(\mathcal{M}(I_{p_i}, T)) \tag{2}$$

为确保唯一性，排除多个 VP 同获最高分的情况。最终生成训练集 D_{train}，将图像映射至唯一最优提示（包括不应用任何 VP 的选项）：

$$D_{\text{train}} = \{(I_j, p_j^*) \mid \text{unique } p_j^*\} \tag{3}$$

训练路由器模型，路由器模型 $\mathcal{R}_\theta$ 基于 D_{train} 训练，用于预测给定图像 I 的最优 VPp^*。它为每个 VP 分配得分 $\hat{s}_{p_i}$：

$$\hat{s}_{p_i} = \mathcal{R}_\theta(I, p_i) \tag{4}$$

通过 softmax 将得分转换为概率：

$$\hat{P}(p_i \mid I) = \frac{\exp(\hat{s}_{p_i})}{\sum_{p_j \in P} \exp(\hat{s}_{p_j})} \tag{5}$$

使用预测概率分布 $\hat{P}(p_i \mid I)$ 与独热编码真实最优 VPp^* 之间的交叉熵损失训练路由器模型：

$$\mathcal{L} = -\sum_{p_i \in P} \mathbb{1}_{p_i = p^*} \log \hat{P}(p_i \mid I) \tag{6}$$

LVLM 推理阶段，训练好的路由器模型 $\mathcal{R}_\theta$ 预测最优 $VP\hat{p}$：

$$\hat{p} = \arg\max_{p_i \in P} \hat{s}_{p_i} \tag{7}$$

将 $\hat{p}$ 应用于 I 中的定位对象 O，生成 $I_{\hat{p}}$。该结果与文本提示 T 共同输入 LVLM$\mathcal{M}$，获得减少对象幻觉的响应。

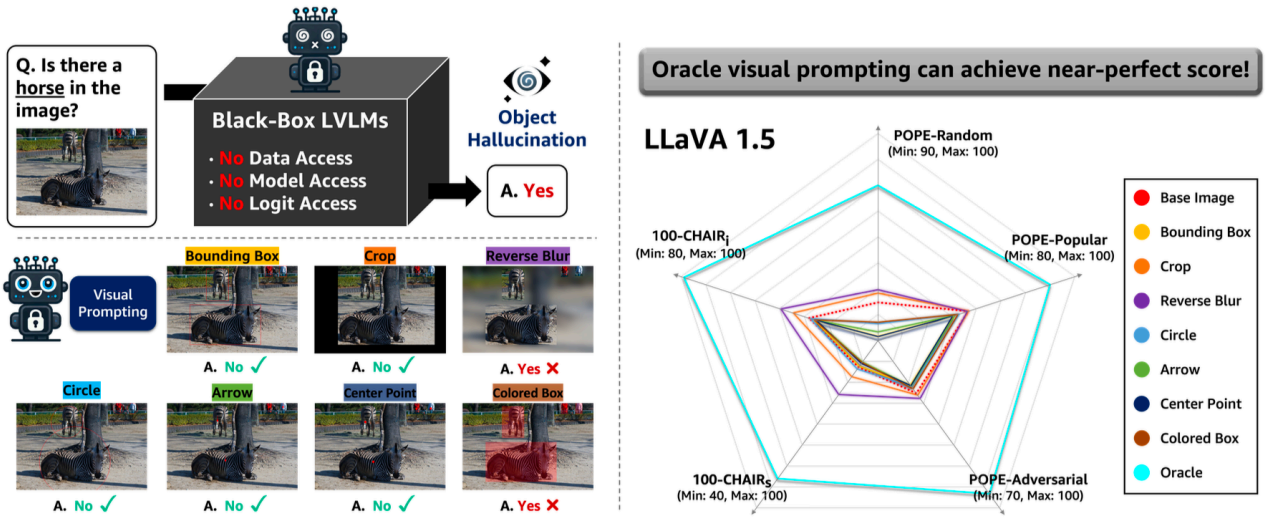

图 1: Motivation.

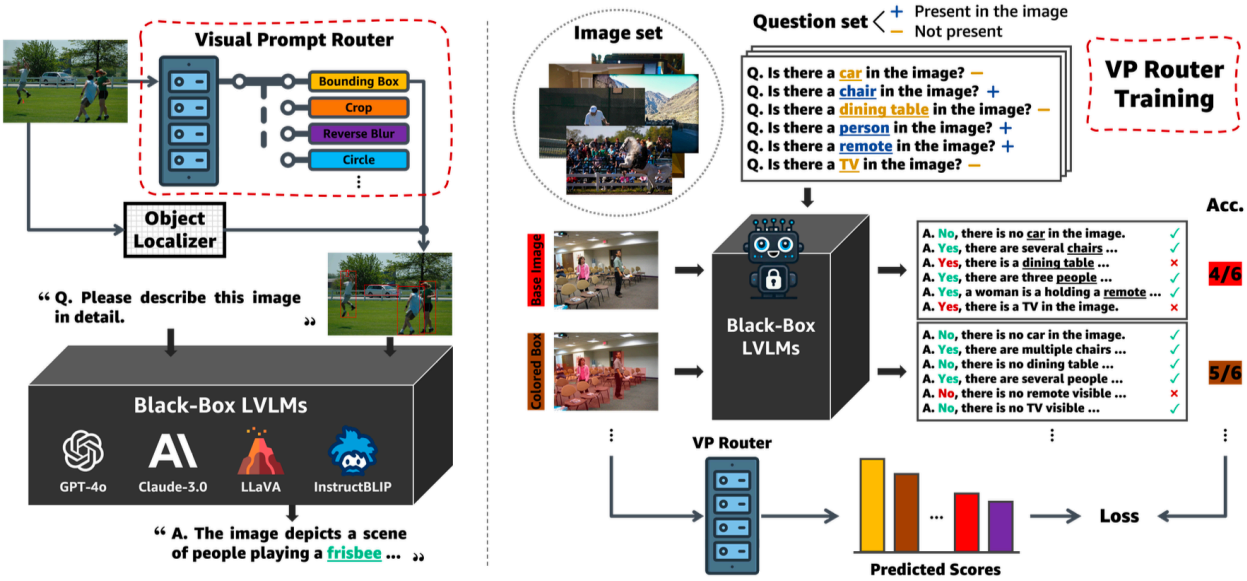

图 2: BBVPE.

实现细节：目标定位器：SAM2；视觉提示路由：采用冻结 CLIP 视觉编码器搭配可训练 MLP；大视觉语言模型：使用两种开源模型（LLaVA-1.5, InstructBLIP）和两种商业模型（GPT-4o, Claude-3.0-Sonnet）。

(6) 可改进的地方

【本文工作的局限性是什么？你觉得可以从哪些方面改进工作?】

当前方法主要针对自然图像，尚未扩展至文档 VQA、科学 VQA 或数学 VQA 等场景的抽象合成图像。这些合成图像通常具有不同的视觉特征，现有方案可能无法直接适用。

目前采用基于 Segment Anything 模型的边界框提示，转向细粒度的掩码式视觉提示可能进一步提升性能。

当前路由模型仅考虑图像特征，未融入问题上下文。初步实验表明，引入问题语境可进一步提升效果，这为未来探索问题感知的视觉提示研究指明了方向。

物体定位至关重要。

(7) 可借鉴的地方

【你觉得本文哪些方面可以借鉴？比如思路、方法、技术等】

本文的思路和方法在多个方面具有借鉴价值。首先，黑箱优化的思路为处理无法访问内部信息的模型提供了一种可行的解决方案，这种方法可以应用于其他类似的黑箱模型优化问题。其次，视觉提示工程的概念为减少幻觉提供了一种新的视角，这种方法可以扩展到其他视觉语言任务中，如图像描述生成、视觉问答等。

(8) 其他收获

【你有什么其他收获吗？比如了解了哪些团队和大牛在某领域做得很好，某类问题通常用什么技术解决，某些技术之间存在什么样的关联，某些会议和期刊在某领域很知名……】

我了解到视觉提示工程和黑箱优化是解决视觉语言模型问题的新兴技术，这些技术在减少幻觉方面具有很大的潜力。同时，我也意识到数据集的选择和模型的泛化能力在实际应用中的重要性。

5 评阅人

姓名:
时间: