# 【通过注意力透镜解读、检测和缓解物体幻觉】

## ——【Devils in Middle Layers of Large Vision-Language Models: Interpreting, Detecting and Mitigating Object Hallucinations via Attention Lens】

## 1 相关资源

pdf: https://arxiv.org/pdf/2411.16724
ppt:
短视频:
数据集:
源码: https://github.com/ZhangqiJiang07/middle_ layers_indicating_hallucinations
网站:
【除了网站，其他资源尽量下载】

## 2 论文属性

论文来源: CVPR 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: Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, Xu Yang
单位:

1. National University of Defense Technology
2. Southeast University
3. Key Laboratory of New Generation Artificial Intelligence Technology Its Interdisciplinary Applications (Southeast University), Ministry of Education
4. Nanyang Technological University

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。

本文旨在通过分析 LVLMs 处理视觉信息的方式，揭示幻觉产生的机制，并提出有效的检测和减轻幻觉的方法。

(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】

目前，对于 LVLMs 中幻觉问题的研究主要集中在语言方面，例如通过视觉指令微调、外部专家模型集成和对比解码策略等方法来减轻幻觉。然而，这些方法大多未能深入探究幻觉产生的根本原因，尤其是视觉信息处理方面的问题。一些研究指出，LVLMs 倾向于优先使用内部文本知识而非外部视觉信息，导致幻觉的产生。但这些研究大多忽略了视觉部分的处理问题，而直观上，视觉信息的不当处理可能是幻觉产生的主要原因之一。因此，当前的研究虽然在一定程度上缓解了幻觉问题，但对于幻觉产生的内在机制的理解仍然有限。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

本文的核心问题是探究 LVLMs 如何处理视觉信息，以及这种处理方式如何导致幻觉的产生，并提出一种在推理阶段调整视觉信息处理过程的方法，以减轻幻觉现象。

(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. 本文深入探究了 LVLM（大规模视觉语言模型）如何处理图像标记中的视觉信息及其对物体幻觉生成的影响。
2. 提出一种简易方法：在推理过程中整合多头注意力信息来调整中间层的视觉信息处理
3. 全面实验验证表明，该方法在多种场景下显著减少幻觉的同时，保持了模型在标准任务上的性能

(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】

**理论基础**

视觉注意力比率。对于第 $k$ 个标记 $y_k$，定义其在第 $\ell$ 层第 $h$ 个头的视觉注意力比率（VAR）：

$$\mathrm{VAR}^{(\ell,h)}(y_k) \triangleq \sum_{i=1}^{n} A_k^{(\ell,h)}(a_k, i), \tag{1}$$

其中 $A_k^{(\ell,h)}(a_k, i)$ 表示新生成标记 $y_k$ 对图像标记 $v_i$ 分配的注意力权重。VAR 量化了该标记与视觉信息的交互程度：VAR 值越高，表明图像标记在 $y_k$ 生成过程中的贡献越大。

Logit 透镜。运用该技术探究模型如何通过文本来解读推理过程中的视觉隐藏状态 $\mathbf{v}_i^{\ell}$。将 softmax

前的线性投影器记作 $\mathbf{W}_{\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}| \times d}$，它作用于 $\mathbf{y}_{k-1}^{\ell}$ 以预测词汇表 $\mathcal{V}$ 中下一标记的概率。随后，logit 透镜通过线性投影器 $\mathbf{W}_{\mathcal{V}}$ 将图像标记的隐藏状态 $\mathbf{v}_i^{\ell}$ 转换为词汇表上的预测概率分布：

$$\mathbf{p}(\mathcal{V} \mid \mathbf{v}_i^{\ell}) = \text{softmax}\left(\mathbf{W}_{\mathcal{V}} \cdot \mathbf{v}_i^{\ell}\right) \in \mathbb{R}^{|\mathcal{V}|}, \tag{2}$$

其中 $p_j(\mathcal{V} \mid \mathbf{v}_i^{\ell})$ 对应词汇表中第 $j$ 个文本标记。为解释处理后的图像标记，将概率最高的文本标记视为模型对第 $\ell$ 层隐藏状态 $\mathbf{v}_i^{\ell}$ 的解读。

为探究生成真实与幻觉物体标记的内部规律，作者使用贪婪搜索模型在”请详细描述该图像”提示下生成选定图像的描述。随后以真实标注为参照识别真实与幻觉物体标记。对于多标记物体，仅考虑首标记。最终获得 1,842 个幻觉标记和 4,397 个真实标记。

**发现 1：中层对视觉信息交互至关重要**

由图一，可视化显示 5-26 层持续对图像标记保持高注意力权重，表明视觉信息交互主要发生在中层。

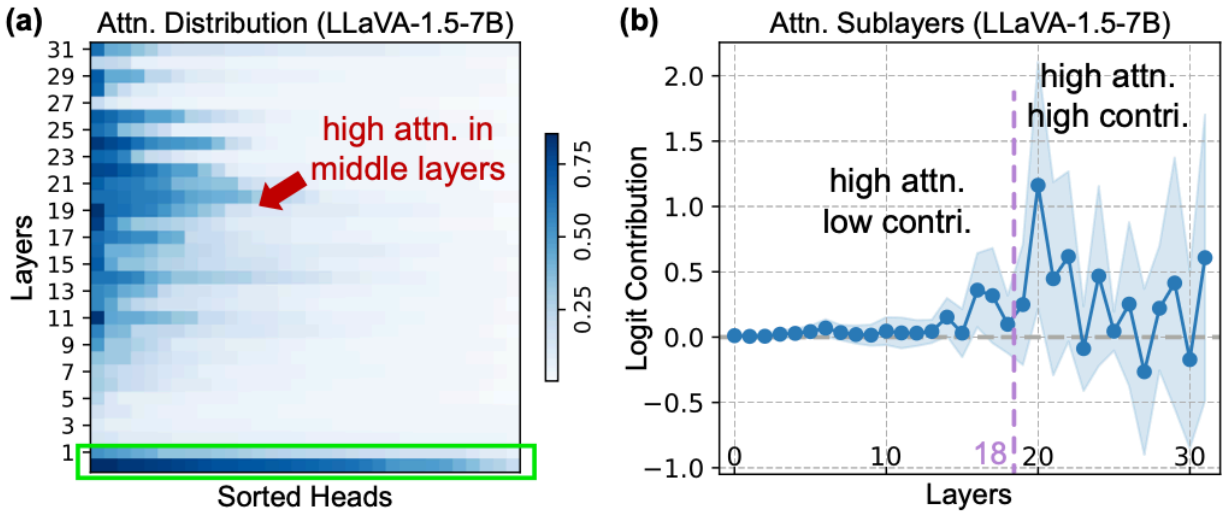


图 1: Overview of our proposed method.

图二揭示了中间层存在两个具有不同视觉信息处理机制的阶段：

阶段 1：视觉信息富集。第 5-18 层显示检索到的文本标记与对应图像块关联性较弱，表明模型尚未具备视觉信息解析能力。这种错位现象可解释图中观测到的预测贡献度偏低现象。值得注意的是，这些层仍保持较高的 VAR 分数，说明物体标记正通过自注意力机制持续积累视觉信息。后续将证明此类注意力是识别幻觉现象的关键指标。

阶段 2：语义精炼。第 19-26 层显示检索文本标记与图像块在语义上高度一致，表明模型已能解析并利用图像标记编码的视觉信息。鉴于该阶段持续的高 VAR 值，作者将其定义为语义精炼阶段——模型主动与图像标记的语义信息交互，进而推理物体标记的预测结果。

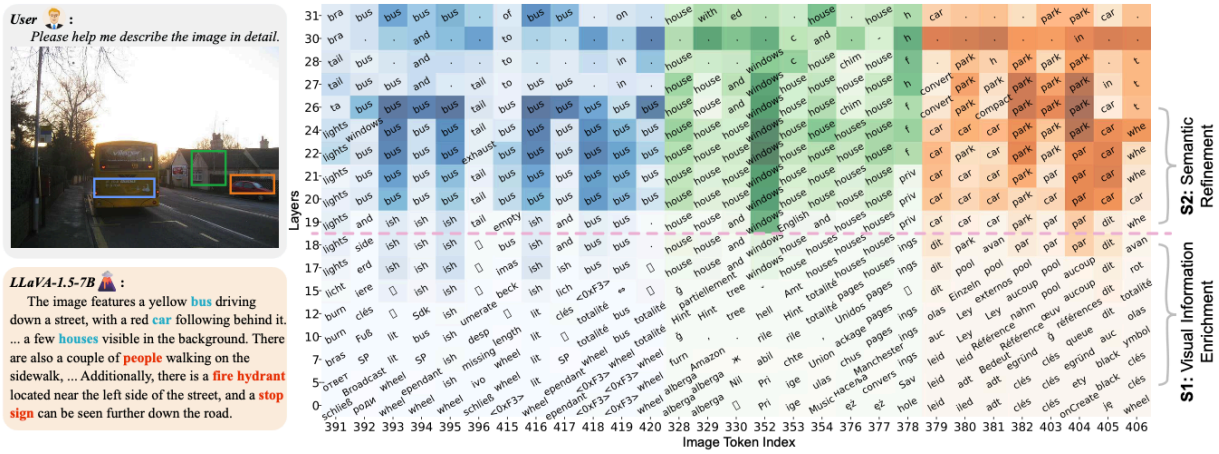


图 2: (a) Distribution of visual attention ratio for real ob- ject tokens across heads and layers in LLaVA-1.5-7B, sorted row- wise by attention ratios.(b) The logit contribution of attention sublayers to real token prediction.

**发现 2：中间层视觉注意力失活预示幻觉现象**

观测发现，在视觉信息富集阶段（5-18 层），幻觉标记相较于真实标记表现出注意力失活模式。我们推测该模式削弱了物体标记与图像标记在此阶段的交互，限制视觉信息传播并可能诱发幻觉。为量化该现象，作者引入” 视觉注意力总和比率 (SVAR)” 指标，通过对所有注意力头的 VAR 分数取平均并在选定层求和计算。

结果发现在视觉信息富集阶段，模型对真实物体标记分配的图像注意力权重显著高于幻觉标记。

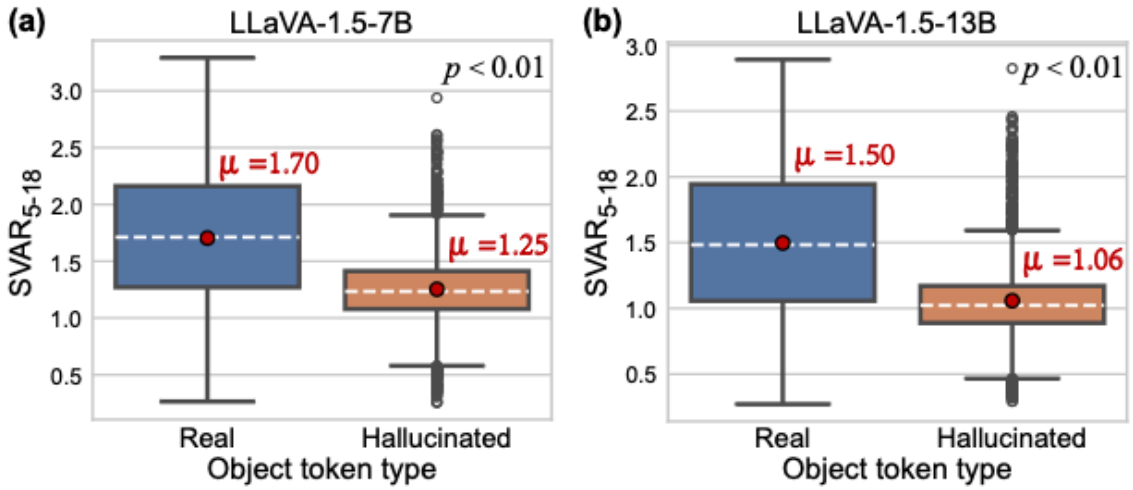


图 3: SVAR5-18 score distributions across object token types for the 7B (a) and 13B (b) versions of LLaVA-1.5.

**应用：物体幻觉检测**

基于发现二，作者采用 $SVAR_{5-18}$ 指标检测物体幻觉。通过案例数据集评估 $SVAR_{5-18}$ 反映物体幻

觉的有效性，其中正样本为真实物体标记，负样本为幻觉标记。作为对比基线，我们使用文献 [?] 提出的内部置信度指标，该指标通过 logit 透镜映射图像隐藏状态 $\{\mathbf{v}_i^\ell : i \in [n], \ell \in [L]\}$ 中物体标记的最大概率进行检测。相比内置信度，简单采用 $\text{SVAR}_{5-18}$ 指标使 AUROC 提升 8.82%，AP 提升 3.53%，验证了该指标的实用性。这表明中间层能为物体幻觉提供关键指示信息。

在 LLaVA-1.5-7B 中测试四个层范围。可见视觉信息富集层（5-18）表现最优，验证了发现二的有效性。其他层虽也编码幻觉相关信息，但相比 5-18 层的显式模式，其他范围的隐式模式需要额外计算成本进行特征学习。

**发现三：注意力头与多物体交互引发物体幻觉**

鉴于 MHSA 机制通过多注意力头聚合信息，作者分析真实/幻觉物体标记（' 公交车'vs' 人群'）的图像热力图，研究视觉信息富集阶段的注意力头行为。结果发现每个注意力头聚焦不同局部细节，但真实标记的注意力分布始终与物体空间范围一致；而幻觉标记的注意力头在此阶段会与图像中不一致物体交互。推测这种不稳定行为将多物体混合信息编码至物体标记，可能导致幻觉。

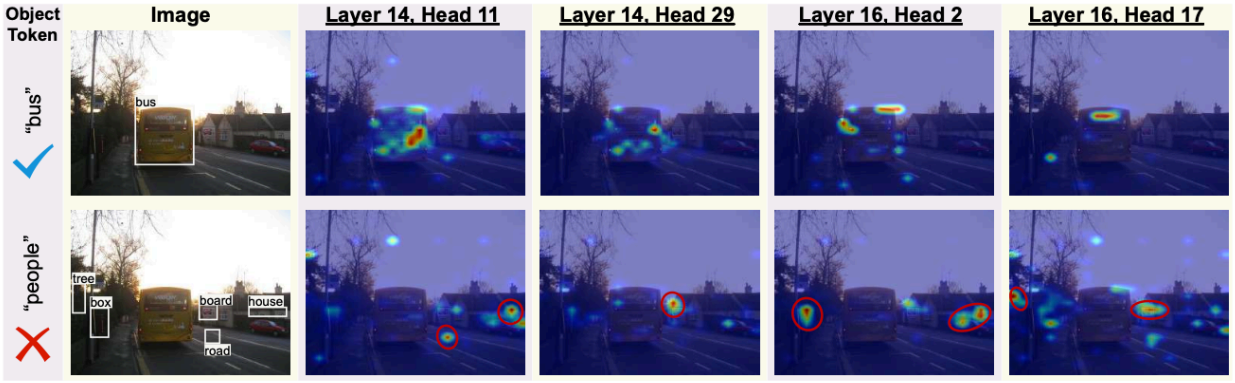


图 4: Visualization of attention maps over image for ‘bus’ (real) and ‘people’ (hallucinated) object tokens in heads ( ,h).

**物体幻觉缓解–注意力头引导的干预**

基于之前的重要发现——推理过程中视觉信息处理不当会导致物体幻觉，作者旨在修正这一过程以缓解幻觉现象。具体而言，作者在推理时对中间层各注意力头的图像标记（即 $\left\{\mathbf{A}_k^{(\ell,h)}(a_k, i)\right\}_{i=1}^{n}$）的注意力权重进行干预调整。为修改时间步 $k$ 的注意力权重，提取 softmax 前的注意力分数矩阵 $\mathbf{S}_k^{(\ell,h)}$。

$$\mathbf{S}_k^{(\ell,h)} = \left(\frac{\mathbf{Q}_{\ell,h}\mathbf{K}_{\ell,h}^\top}{\sqrt{d_k}}\right)_k, \tag{3}$$

其中 $\mathbf{Q}_{\ell,h}$ 和 $\mathbf{K}_{\ell,h} \in \mathbb{R}^{a_k \times d_k}$ 分别表示维度为 $d_k$ 的查询矩阵与键矩阵

根据发现 2 中的注意力模式差异，我们在视觉信息富集层通过添加正值来放大原始注意力分数，从而增强视觉信息交互。进一步结合发现 3 中观察到的注意力头行为，这些正值通过计算所有注意力头绝对分数的平均值得到，使得不同注意力头一致关注的区域获得更大增强。通过整合多注意力头信息，我们能找到更忠实且与物体相关的注意力转移方向以减少幻觉。形式上，对于第 $\ell \in [\ell_s, \ell_e]$ 层所有注意力头 $h \in [H]$ 中的每个图像标记 $i \in [n]$，我们通过以下方式调整视觉注意力分数：

$$\mathbf{S}_k^{(\ell,h)}(a_k, i) = \mathbf{S}_k^{(\ell,h)}(a_k, i) + \alpha \frac{1}{H} \sum_{h=1}^{H} \left|\mathbf{S}_k^{(\ell,h)}(a_k, i)\right| \tag{4}$$

其中 $[\ell_s, \ell_e]$ 表示视觉信息富集范围，参数 $\alpha$ 作为平衡因子控制干预强度

(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?】

本文的方法依赖于对中间层注意力模式的调整，而这种调整可能对不同类型的 LVLMs 效果不同。未来的研究可以进一步探索更通用的注意力调整策略，使其适用于更广泛的模型架构。

(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】

本文在分析 LVLMs 幻觉问题时采用的注意力机制分析方法具有很强的借鉴意义。通过将注意力模式与幻觉现象联系起来，为理解模型行为提供了一个新的视角。这种方法可以应用于其他类型的模型和任务，帮助研究人员更好地理解模型的内部工作机制。

(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……】

我了解到注意力机制在模型解释和优化中的重要作用，以及如何通过分析注意力模式来揭示模型的内部工作机制。

## 5 评阅人

姓名:

时间: