

# 【激活引导解码：通过双向隐状态干预缓解大型视觉语言模型中的幻觉现象】

—— 【Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention】

## 1 相关资源

pdf: <https://aclanthology.org/2025.acl-long.634/>

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: ACL 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: Jingran Su, Jingfan Chen, Hongxin, Yuntao Chen, Qing Li , Zhaoxiang Zhang

单位:

1. The Hong Kong Polytechnic University
2. New Laboratory of Pattern Recognition, CASIA
3. State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
4. Hong Kong Institute of Science Innovation, CASIA
5. University of Chinese Academy of Sciences 6Shanghai Artificial Intelligence Laboratory

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。

本文的工作旨在探索一种新的视角来缓解 LVLMs 中的幻觉问题，通过研究模型生成过程中的中间激活 (hidden states)，提出一种无需重新训练的解决方案，以提高模型的可靠性和实用性。

### (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】**

当前的研究已经取得了一定进展，主要通过增强数据质量、设计训练目标、引入额外的视觉模块或在模型输出过程中进行干预等方式来缓解幻觉问题。例如，一些工作通过对比学习、对抗样本和数据增强来提高训练数据的多样性，还有一些通过强化学习技术来抑制模型的幻觉行为。

然而，这些方法要么需要大量的额外数据，要么涉及昂贵的训练过程，对于实际部署中的模型来说，这些方法可能难以快速适应新的场景。此外，现有的训练后解决方案可能在实际部署中面临挑战，因为模型需要在最小的计算开销下快速适应新场景。

本文指出，尽管这些方法提供了有价值的见解，但它们大多基于特定的假设，例如图像区域的注意力损失，而本文的目标是通过直接操纵模型的中间激活来更根本地解决这一问题。

### (3) 本文解决的问题

**【一句话概括本文解决的核心问题】**

本文解决的核心问题是通过直接干预 LVLMs 的中间激活来有效缓解幻觉现象，同时保持模型在一般视觉理解任务上的性能。

### (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. 通过系统性实证研究揭示了 LVLMs 中间激活空间中独特的幻觉模式，为理解其内部机制提供了新见解
2. 提出 ASD：一种通过针对性干预中间激活来减少幻觉的无训练新方法
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】**

**表征收集框架**

为系统研究基础模型的幻觉模式，我们开发了可扩展的隐藏状态-幻觉标签配对收集框架。该方法专注于对象幻觉——当模型生成输入图像中不存在的对象引用时，这种可量化定义的多模态幻觉形式。数据收集流程如下：

采用 MSCOCO 数据集作为主要数据源，因其丰富的分割标注与多样化视觉内容。对数据集中的每

幅图像，使用提示词”请详细描述该图像”查询基础模型以生成详细描述。生成描述反映模型对输入图像的内在感知，可能包含偏离实际视觉信息的幻觉内容。

$\mathcal{O} = \{o_1, o_2, \dots, o_{80}\}$  代表 MSCOCO 数据集中预定义的 80 个对象类别。对每个对象类别  $o$ ，收集同义词集  $C(o)$  以确保全面对象提取。每幅图像  $v_i$  关联基于 MSCOCO 标注的真实对象集  $G(v_i) \subseteq \mathcal{O}$ 。对每个生成描述  $y_i$ ，使用自然语言工具库将其分割为独立句子  $\{s_{i,1}, s_{i,2}, \dots, s_{i,j}\}$ ，其中每个  $s_{i,j}$  是表示单词的词元子序列：

$$s_{i,j} = (y_1^{i,j}, y_2^{i,j}, \dots, y_p^{i,j}), \text{ with } \bigcup_j s_{i,j} = y_i. \quad (1)$$

随后通过以下方式识别句子  $s_{i,j}$  中提及的所有对象  $O(s_{i,j})$ ：

$$O(s_{i,j}) = \{o \in \mathcal{O} \mid \text{substr}(o, s_{i,j}), \text{ or } \exists c \in C(o), \text{substr}(c, s_{i,j})\}, \quad (2)$$

`substr(x,y): x 是 y 的子串`

作者根据句子  $s_{i,j}$  是否包含任何不存在的对象，为标记  $y_p^{i,j} \in s_{i,j}$  定义幻觉标签  $L(y_p^{i,j})$ 。数学表达式如下：

$$L(y_p^{i,j}) = \begin{cases} 1 & \text{if } O(s_{i,j}) \setminus G(I_i) \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

设  $\mathbf{Z}(y)$  表示标记  $y$  在所有层的隐藏状态，最终构建的激活-标签配对数据集形式为：

$$\bigcup_i (\mathbf{Z}(y_p^{i,j}), L(y_p^{i,j})).$$

### 隐藏状态的线性探测

为探究幻觉发生时隐藏状态的模式特征，作者对 LLaVA1.5-7B 全架构进行线性探测。具体而言，从 MSCOCO 训练集随机抽取 500 张图像，采用前述方法提取所有 32 个 Transformer 层的隐藏状态表征。初始收集得到不平衡数据集，含 42,160 个非幻觉样本和 12,113 个幻觉样本。通过从每类随机抽取 11,000 个实例构建平衡数据集，最终获得 22,000 个样本。保留 2,000 个样本作为测试集，剩余 20,000 用于训练。我们使用不同量级的训练数据开展系列线性探测实验，独立训练 32 个层的线性分类器，追踪幻觉相关信息在模型各层的编码轨迹。

图片展示了不同训练集规模下各模型层的准确率与 F1 分数。分析发现：首先，训练数据量对分类器判别能力影响显著，约需 20k 个样本才能建立可靠模式，表明幻觉特征虽具一致性，但需足量数据才能准确刻画。其次，中后层隐藏状态表现出更优的幻觉检测表征能力，说明幻觉相关特征在模型层级间渐进累积。最值得注意的是，探测结果表明幻觉信息在隐藏状态空间中保持高度完整且线性可分，仅用 20k 训练标记即可在中层实现 82.49% 的探测准确率。这种显著的线性可分性有力证明：幻觉内容在模型隐藏状态空间中以独特、一致的模式显现，暗示针对隐藏状态的定向干预可有效缓解幻觉行为。

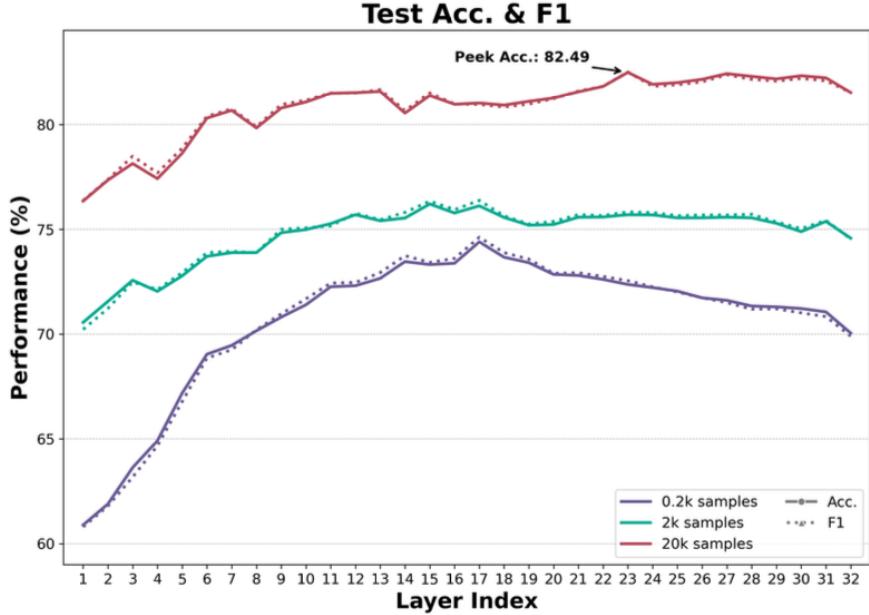


图 1: Test accuracy and F1 scores for hallucination versus non-hallucination classification across different layers of LLaVA-1.5-7B with varying training sample sizes (0.2k, 2k, and 20k).

### 激活导向解码

基于“幻觉模式在隐藏状态中具有独特编码且线性可分”的实证发现，作者提出激活导向解码策略——通过直接干预模型隐藏激活来抑制幻觉的新颖解码方法。

根据收集的配对数据  $\bigcup_i \{(\mathbf{Z}(y_p^{i,j}), L(y_p^{i,j}))\}$ ，作者计算捕获隐藏状态空间中“幻觉  $\rightarrow$  非幻觉”方向的导向向量。对每个第 1 层，计算非幻觉与幻觉标记的激活均值差：

$$\mathbf{v}^l = \frac{1}{P} \sum_{L(y)=1} \mathbf{z}_l(y) - \frac{1}{N} \sum_{L(y)=0} \mathbf{z}_l(y), \quad (4)$$

其中  $P$  和  $N$  分别表示真实标记与幻觉标记的数量。

利用提取导向向量的最直接方法是干预隐藏状态：

$$\mathbf{z}_l^{\text{steered}} = \mathbf{z}_l + \lambda \mathbf{v}_l, \quad (5)$$

其中  $\lambda$  调控导向强度。虽然该方法随  $\lambda$  增大能有效减少幻觉，但可能扭曲隐藏状态编码的语义信息。

为在保持生成质量的同时实现更稳定的幻觉抑制，作者提出激活引导解码方法。设  $\pi^+$  和  $\pi^-$  分别表示模型在正向（即  $\lambda > 0$ ）和负向（即  $\lambda < 0$ ）引导下的状态，通过式 (2) 施加方向相反的相同引导向量。最终通过以下方式获得下一词元预测的对数概率：

$$\text{logit}_{ASD} = (1 + \alpha) \cdot \text{logit}_{\pi^+} - \alpha \cdot \text{logit}_{\pi^-}, \quad (6)$$

其中  $\alpha$  为对比权重系数。该对比机制之所以有效，是因为差分运算放大了引导对输出对数概率的影响，同时允许使用较小的引导强度以更好地保持隐藏状态的语义完整性。相较于直接引导，这一特性使该方法更具鲁棒性，且不易干扰模型的正常生成过程。

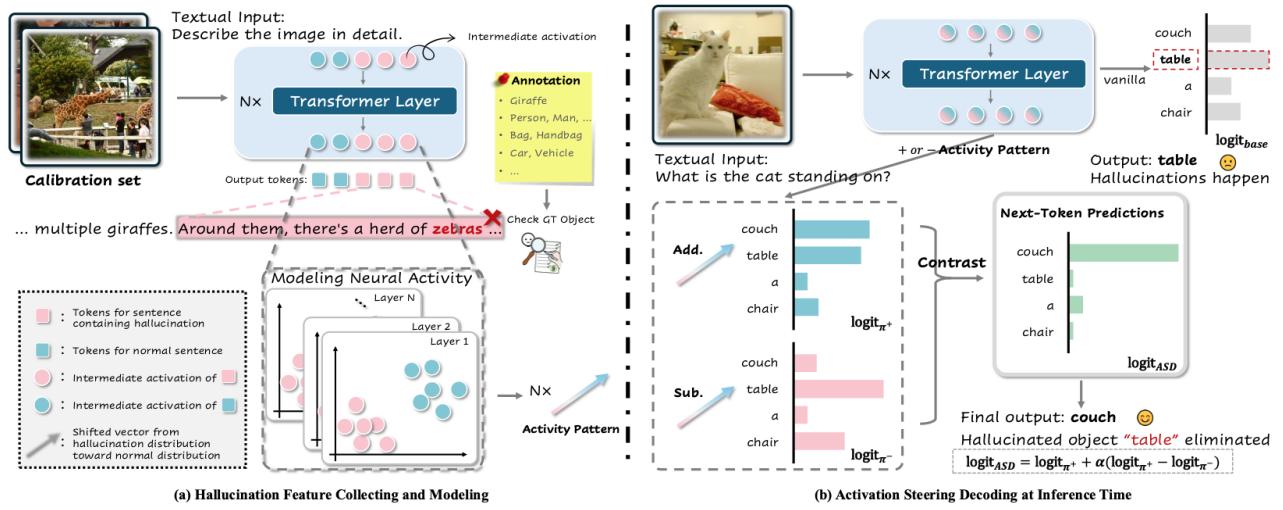


图 2: Overview of our proposed method.

## 实验细节

作者在两个基础模型上开展实验：LLaVA1.5-7B 和 Qwen-VL-Chat。针对每个模型，从 MSCOCO 训练集中随机抽取 1,000 张图像用于公式 (1) 的导向向量提取。对  $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$  两个参数进行网格搜索，范围分别为  $\pi^+$  和  $\pi^-$ 。作为对比，采用优化超参数实现了 VCD，并使用推荐参数实现了 VDD=None。

### (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？】**

本文的局限性在于，当前的方法主要针对对象级别的幻觉，因为引导向量是使用 COCO 对象注释提取的。这种对对象类别的关注限制了该方法处理其他类型幻觉的能力，例如属性错误（如错误的颜色或大小）、关系不准确（如错误的空间关系）或涉及抽象概念和动作的幻觉。

### (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】**

首先，其对 LVLMs 中间激活空间中幻觉模式的系统研究为理解模型内部机制提供了新的视角，这种对模型内部行为的深入分析可以为其他研究人员提供启发，帮助他们更好地理解模型的工作原理。其次，激活引导解码 (ASD) 方法的提出为处理 LVLMs 中的幻觉问题提供了一种新的技术途径，这种方法的训练自由特性使其具有很高的实用价值，尤其是在需要快速适应新场景而无需重新训练模型的情况下。

### (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……】**

本文展示了如何通过对比解码机制来增强模型的输出，这种方法可以应用于其他需要精确控制模型生成内容的场景。

## 5 评阅人

姓名:

时间: